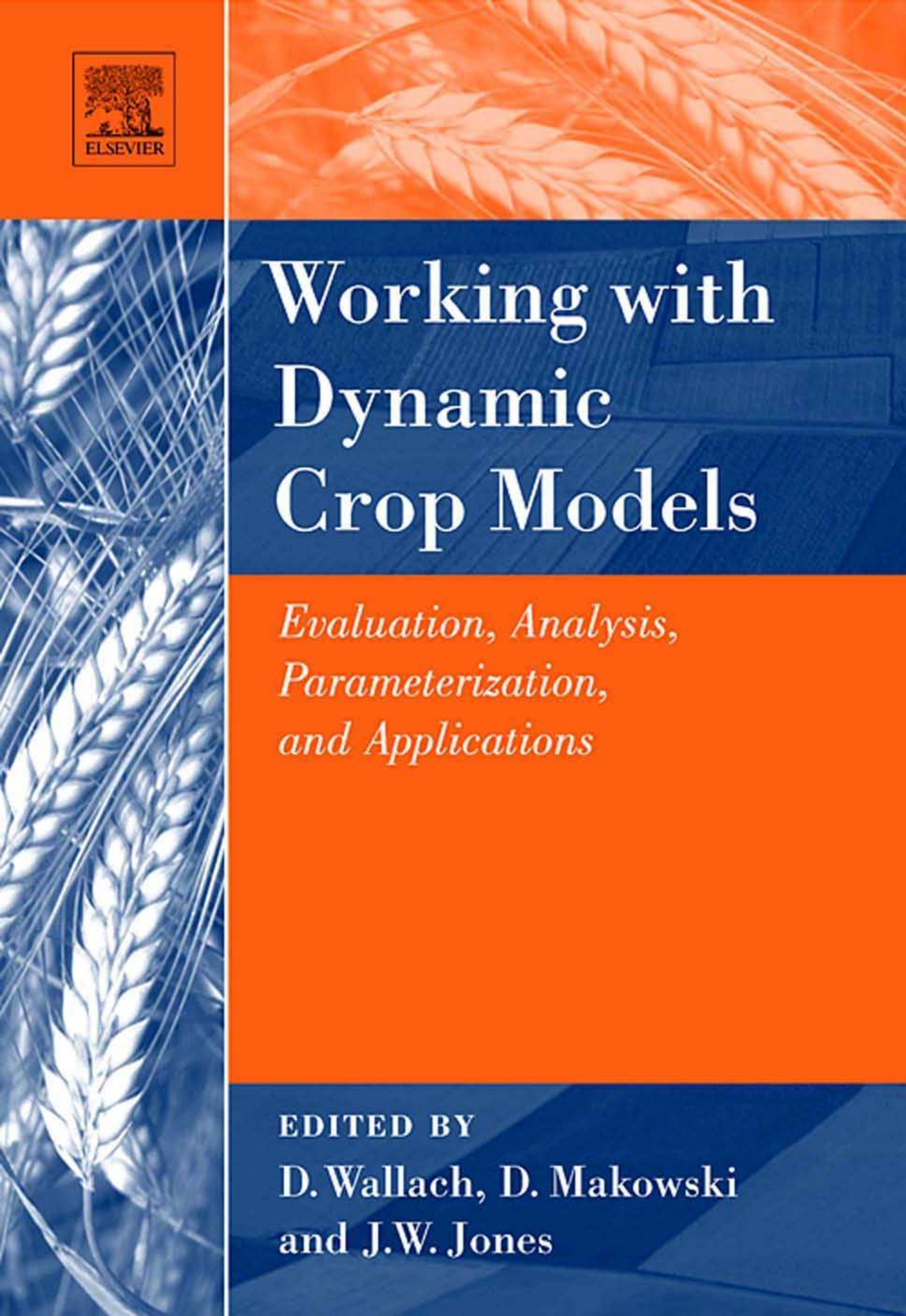


Working with Dynamic Crop Models

Evaluation, Analysis, Parameterization, and Applications

EDITED BY
D. Wallach, D. Makowski
and J.W. Jones

ELSEVIER

Working with Dynamic Crop Models

Evaluation, Analysis, Parameterization, and Applications

Working with Dynamic Crop Models

Evaluation, Analysis, Parameterization, and Applications

Edited by

Daniel Wallach
INRA, Castanet Tolosan, France

David Makowski
INRA, Thiverval-Grignon, France

James W. Jones
University of Florida, Gainesville, USA

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

ELSEVIER
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

First edition 2006

Copyright © 2006 Elsevier B.V. All rights reserved

No part of this publication may be reproduced, stored in a retrieval system
or transmitted in any form or by any means electronic, mechanical, photocopying,
recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at http://elsevier.com/locate/permissions, and selecting *Obtaining permission to use Elsevier material*

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN-13: 978-0-444-52135-4
ISBN-10: 0-444-52135-6

For information on all Elsevier publications
visit our website at books.elsevier.com

Printed and bound in The Netherlands

06 07 08 09 10   10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER   BOOK AID International   Sabre Foundation

Contents

Contributors vii
Preface xi
Overview xiii

I Methods 1

1 The two forms of crop models 3
D. Wallach
2 Evaluating crop models 11
D. Wallach
3 Uncertainty and sensitivity analysis for crop models 55
H. Monod, C. Naud and D. Makowski
4 Parameter estimation for crop models 101
D. Makowski, J. Hillier, D. Wallach, B. Andrieu and M.-H. Jeuffroy
5 Data assimilation with crop models 151
D. Makowski, M. Guérif, J.W. Jones and W. Graham
6 Representing and optimizing management decisions with crop models 173
J.E. Bergez, F. Garcia and D. Wallach
7 Using crop models for multiple fields 209
D. Leenhardt, D. Wallach, P. Le Moigne, M. Guérif, A. Bruand and M.A. Casterad

II Applications 249

8 Introduction to Section II 251
J.W. Jones, D. Makowski and D. Wallach
9 Fundamental concepts of crop models illustrated by a comparative approach 257
N. Brisson, J. Wery and K. Boote
10 Crop models with genotype parameters 281
M.-H. Jeuffroy, A. Barbottin, J.W. Jones and J. Lecoeur
11 Model-assisted genetic improvement of crops 309
C.D. Messina, K.J. Boote, C. Löffler, J.W. Jones and C.E. Vallejos
12 Parameterization and evaluation of a corn crop model 337
D. Wallach
13 Evaluation of a model for kiwifruit 349
F. Lescourret and D. Wallach

14 Sensitivity and uncertainty analysis of a static denitrification model 359
B. Gabrielle
15 Sensitivity analysis of PASTIS, a model of nitrogen transport and transformation in the soil 367
P. Garnier
16 Sensitivity analysis of GeneSys, a model for studying the effect of cropping systems on gene flow 377
N. Colbach and N. Molinari
17 Data assimilation and parameter estimation for precision agriculture using the crop model STICS 391
M. Guérif, V. Houlès, D. Makowski and C. Lauvernet
18 Application of Extended and Ensemble Kalman Filters to soil carbon estimation 399
J.W. Jones and W.D. Graham
19 Analyzing and improving corn irrigation strategies with MODERATO, a combination of a corn crop model and a decision model 409
J.-E. Bergez, J.-M. Deumier and B. Lacroix
20 Managing wheat for ethanol production: a multiple criteria approach 419
C. Loyce, J.P. Rellier and J.M. Meynard

Appendix. Statistical notions 429

Answers to Exercises 437

Index 441

Contributors

Bruno Andrieu INRA, Unité mixte de recherche Environnement et grandes cultures, 78850 Thiverval-Grignon, France

Aude Barbottin UMR INRA SAD APT-INA-PG, BP 01, 78 850 Thiverval-Grignon, France

Jacques-Eric Bergez INRA, UMR INRA/ENSAT ARCHE, BP 52627, 31326 Castanet Tolosan Cedex, France

Ken Boote Agricultural and Biological Engineering Department, University of Florida, PO Box 110570, Gainesville, FL 32611, USA

Nadine Brisson INRA, Unité CSE, Domaine Saint-paul, 84914 Avignon Cedex 9, France

Ary Bruand Institut des Sciences de la Terre d'Orléans (ISTO), UMR 6113 CNRS UO, Université d'Orléans, Géosciences BP 6759-45067 Orléans Cedex 2, France

Auxiliadora Casterad Centro de Investigación y Tecnología Agroalimentaria de Aragón (CITA), Unidad de Suelos y Riegos, Apartado 727, 50080 Zaragoza, Spain

Nathalie Colbach Unité mixte de recherche Biologie et gestion des adventices, 17, Rue Sully BP 86510, 21065 Dijon Cedex, France

Jean-Marc Deumier ARVALIS - Institut du végétal, 6, chemin de la cote vieille, 31450 Baziege, France

Benoît Gabrielle INRA, Unité mixte de recherche Environnement et grandes cultures, 78850 Thiverval-Grignon, France

Frédéric Garcia INRA, Unité de Biométrie et intelligence artificielle, BP 27, 31326 Castanet Tolosan Cedex, France

Patricia Garnier INRA, Unité d'Agronomie Laon-Reims-Mons, Rue Fernand Christ, 02007 Laon Cedex, France

Wendy Graham Department of Agricultural and Biological Engineering, PO Box 110570, Rogers Hall, University of Florida, Gainesville, FL 32611-0570, USA

Martine Guérif INRA, Unité CSE, Domaine Saint-Paul, 84914 Avignon Cedex 9, France

Jonathan Hillier University College London, Department of Geography, 26 Bedford Way, London, WC1H OAP, UK

Vianney Houlès Laboratoire GEOTROP du CIRAD/AMIS/Agronomie, Maison de la Télédétection, 500, rue J.F. Breton, 34 093 Montpellier Cedex 5, France

Marie-Hélène Jeuffroy INRA, UMR Agronomie INRA-INA-PG, B. P. 01, 78850 Thiverval-Grignon, France

James W. Jones Agricultural and Biological Engineering Department, University of Florida, PO Box 110570, Gainesville, FL 32611, USA

Bernard Lacroix ARVALIS-Institut du végétal, 6 chemin de la côte vieille 31450 Baziège, France

Claire Lauvernet UMR INRA/Univ. Avignon et Pays de Vaucluse: Climat, Sol et Environnement, Domaine Saint-Paul, Site Agroparc, 84914 Avignon Cedex 9, France

Jérémie Lecoeur UMR LEPSE, Agro M-INRA, 2, place Viala, 34060 Montpellier Cedex 01, France

Delphine Leenhardt INRA, UMR INRA/ENSAT ARCHE, BP 52627, 31326 Castanet Tolosan Cedex, France

Patrick Le Moigne Météo-France CNRM/GMME/MC2, 42, av. G. Coriolis, 31057 Toulouse Cedex, France

Françoise Lescourret INRA, Unité Plantes et systèmes de culture horticoles, Domaine Saint-Paul - Site Agroparc, 84914 Avignon Cedex 9, France

C. Löffler DuPont Agriculture & Nutrition, 7200 NW 62nd Ave. P.O. Box 184, Johnston, IA 50131-0184, USA

Chantal Loyce INA P-G Département AGER, UMR Agronomie, Bâtiment EGER, 78850 Thiverval-Grignon, France

David Makowski INRA, UMR Agronomie INRA-INA-PG, B. P. 01, 78850 Thiverval-Grignon, France

Carlos D. Messina DuPont Agriculture & Nutrition, 7200 NW 62nd Ave. P.O. Box 184, Johnston, IA 50131-0184, USA

Jean-Marc Meynard INRA, Département sciences pour l'action et le développement, 78850 Thiverval-Grignon, France

Nicolas Molinari IURC, 641 Av. Gaston Giraud, Montpellier, France

Hervé Monod INRA, Unité de recherche Mathématiques et Informatique Appliquées, Domaine de Vilvert, 78352 Jouy-en-Josas Cedex, France

Cédric Naud INRA, UMR Agronomie INRA-INA-PG, B. P. 01, 78850 Thiverval-Grignon, France

C. Eduardo Vallejos 1143 Fifield Hall, Program in Plant Molecular and Cellular Biology and Horticultural Sciences, University of Florida, Gainesville, FL 32611-0690, USA

Daniel Wallach I.N.R.A. Toulouse, *UMR1248 ARCHE*, BP 52627, 31326 Castanet Tolosan Cedex, France

Jacques Wery UMR INRA/ENSAM/CIRAD: Fonctionnement et Conduite des Systèmes de Cultures Tropicaux et Méditerranéens, CIRAD Avenue Agropolis TA 80/01, 34398 Montpellier Cedex 5, France

Acknowledgments

A number of scientists were willing to review various chapters and make suggestions concerning both content and form. We gratefully acknowledge our indebtedness to Dr. Senthold Asseng (CSIRO, Wembley, WA, Australia), Dr. Jean-Noel Aubertot (INRA, Grignon, France), Dr. Frédéric Baret (INRA, Avignon, France), Dr. Patrick Durand (INRA, Rennes, France), Dr. Carine Hue (INTA, Toulouse, France), Dr. Eric Malézieux (CIRAD, Montpellier, France), Dr. Walter Rossing (WUR, Wageningen, The Netherlands), and Dr. Jean-Philippe Steyer (INRA, Narbonne, France).

Preface

This book focuses on the methods for working with crop models, in particular, on mathematical and statistical methods. Most crop modelers are painfully aware of the need for such methods. Parameter estimation, model evaluation, and sensitivity analysis are called for in essentially every modeling project. Other methods treated in this book, related to the use of in-season measurements for improving model predictions, to optimization of management decisions, to the use of models on a large spatial scale or to the use of models to aid in genetic improvement of crops, are also important but only in certain cases.

In crop modeling as in all fields, it is a challenge to keep up with progress, and this is particularly difficult when it comes to mathematical and statistical methods, which are developed outside the framework of crop models. The purpose of this book is to make these methods easily available to crop modelers. We felt that there is a gap in the literature in this respect. Many books treat the way to describe a crop system in terms of equations, but none seems to provide an in-depth presentation of a large range of methods for working with crop models.

This book is intended for use in a graduate level course in crop modeling (hence the exercises), and for researchers who wish to use crop models. It should be useful to biologists, agronomists, and plant physiologists who are comfortable with describing and quantifying the soil–plant–atmosphere system, but are not familiar with rigorous methods for dealing with complex dynamic models. Others who may benefit are students and researchers with more mathematical and statistical backgrounds who are interested in the applications of applied mathematics to crop models. The emphasis throughout is on crop models, but in fact, much of the material applies more generally to dynamic models of complex systems.

While preparing the contents of this book, we had three main goals. First, the book should reflect the latest knowledge about the different topics covered. Second, the material should be adapted to and applicable to complex dynamic models, in particular, crop models. This is achieved by discussing each method in the specific context of crop models, by using simple crop models to provide the illustrative examples in the text and by furnishing case studies involving crop models. Finally, the material should be accessible to someone who has had basic courses in statistics and linear algebra. To this end, we have tried to explain each method simply, but without sacrificing detail or accuracy. To help the reader, an appendix reviews the statistical notions that are used in the text.

The origins of this book go back to the year 2000, when a group of French researchers began to prepare an intensive week-long school for modelers. The book began as a syllabus for that course. The years since then have gone into testing the material in other courses, expanding the coverage and refining the contents.

Statistician G.E.P. Box once wrote, “All models are wrong, but some are useful”. Our hope is that this book, by improving access to important methods, will contribute to increasing the usefulness of crop models.

D. Wallach
D. Makowski
J. W. Jones

Overview

Herein is a brief overview of the contents of the book.

I. Methods

1. The two forms of crop models. This chapter is concerned with the mathematical form of crop models. Crop models consist of a set of dynamic equations (form 1), which one integrates to get predictions of responses *versus* inputs (form 2). The uses of the two forms are quite different.

2. Evaluation. This chapter first presents and discusses different measures of the distance between model predictions and observed values. It then discusses the notion of prediction error and insists on the difference between how well the model reproduces past data and predicts future values. There is also a discussion on how to evaluate a model when it is used to propose crop management decisions.

3. Uncertainty and sensitivity. Such analyses are aimed at describing how variations in input factors (variables or parameters) affect the output variables. The chapter begins by reviewing the uses of such analyses. The rest of the chapter discusses different sensitivity or uncertainty indices and how they are calculated, in particular, in the case where multiple input factors vary.

4. Parameter estimation. There is a very large statistical literature about parameter estimation, but most of it cannot be directly applied to crop models. The specific problems of crop models include the large number of parameters compared to the amount of field data and the complex structure of that data (several variables, at various dates). On the other hand, there is often outside knowledge about many of the parameter values (from controlled environment studies, similar crops, etc.). The chapter begins with a basic introduction to the principles and methods of parameter estimation. Then the specific case of complex crop models is considered. A number of approaches to parameter estimation that have been or could be used are described and illustrated. Included here is the Bayesian approach to parameter estimation, which is particularly adapted to the efficient use of the outside information.

5. Data assimilation. In-season information about crop growth, for example from satellite photos, is becoming increasingly available. This information can be used to adjust a crop model to reflect the specific trajectory of the field in question. This chapter discusses and illustrates how this adaptation can be done. In particular, variants of the Kalman filter approach are explained and illustrated.

6. Representing and optimizing management decisions. Improving crop management is a major use of crop models. The first part of this chapter concerns how to express management decisions, and discusses in particular decision rules, which express decisions as functions of weather or state of the crop. The second part of the chapter presents and discusses algorithms for calculating optimal decisions. The problem is very complex because of the multiple decisions and the uncertainty in future climate, but efficient algorithms exist.

7. Using crop models for multiple fields. One is often faced with the problem of running a crop model for multiple fields, for example in order to predict regional yields or nitrogen leaching for each field in a watershed. This chapter discusses the specific problems posed by this use of crop models. A major problem is that in general one cannot obtain all the necessary input variables for every field. The chapter presents the different solutions that have been proposed for each type of input data.

II. Applications

8. Introduction to Section II

9. Fundamental concepts of crop models. This chapter discusses the way crop models represent a crop–soil system, with examples from five different crop models.

10. Crop models with genotype parameters. The existence of multiple varieties for each crop, and the fact that many new varieties are developed each year, is a problem specific to crop models. It is important that models be variety specific, but this raises the problem of how to identify and estimate the variety specific parameters. This chapter discusses the approaches that have been proposed.

11. Model assisted genetic improvement in crops. This chapter covers the very new field of the use of crop models in plant breeding. It explains the different ways in which crop models can contribute to selection and includes examples of such uses.

12–20. Case studies. These chapters illustrate a diversity of applications of crop models, and show how the methods presented in Section I can be useful.

Section I

Methods

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 1

The two forms of crop models

D. Wallach

1. **Introduction** 3
2. **A crop model is a dynamic system model** 3
3. **A crop model is a response model** 6
4. **Working with crop models. Which form?** 7
5. **Conclusions** 9
Exercises 10

1. Introduction

Crop models are mathematical models which describe the growth and development of a crop interacting with soil. They can be viewed in two different, complementary ways. First, a crop model can be seen as a system of differential or difference equations, which describe the dynamics of the crop–soil system. Second, the model can be thought of as a set of equations for responses of interest as functions of explanatory variables. We present and discuss these two viewpoints in this chapter. As we shall see, the different methods described in this book may call for one or the other of these viewpoints.

2. A crop model is a dynamic system model

The general form of a dynamic system model in discrete time is

$$\begin{aligned} U_1(t+\Delta t) &= U_1(t) + g_1\left[U(t), X(t); \theta\right] \\ &\vdots \\ U_S(t+\Delta t) &= U_S(t) + g_S\left[U(t), X(t); \theta\right] \end{aligned} \tag{1}$$

where t is time, Δt is some time increment, $U(t) = [U_1(t), \ldots, U_S(t)]^{\mathrm{T}}$ is the vector of state variables at time t, $X(t)$ is the vector of explanatory variables at time t, θ is the vector of parameters and g is some function. For crop models, Δt is often one day. The state variables $U(t)$ could include for example leaf area index (leaf area per unit soil area), biomass, root depth, soil water content in each of several soil layers, etc. The explanatory variables $X(t)$ typically include initial conditions (such as initial soil moisture), soil characteristics (such as maximum water holding capacity), climate variables (such as daily maximum and minimum temperature) and management variables (such as irrigation dates and amounts). Chapter 9 contains an overview of the processes generally described by crop models.

The model of Eq. (1) is dynamic in the sense that it describes how the state variables evolve over time. It describes a system in the sense that there are several state variables that interact.

To illustrate, we present a very simplified crop model with just 3 state variables, namely the temperature sum *TT*, plant biomass B and leaf area index *LAI*. The equations are:

$$TT(j+1) = TT(j) + \Delta TT(j)$$

$$B(j+1) = B(j) + \Delta B(j)$$

$$LAI(j+1) = LAI(j) + \Delta LAI(j)$$

with

$$\Delta TT(j) = \max\left[\frac{T\mathrm{MIN}(j) + T\mathrm{MAX}(j)}{2} - T_{\mathrm{base}},\ 0\right] \tag{2}$$

$$\begin{aligned} \Delta B(j) &= RUE(1 - \mathrm{e}^{-K \cdot LAI(j)})I(j) && TT(j) \le TT_{\mathrm{M}} \\ &= 0 && TT(j) > TT_{\mathrm{M}} \end{aligned} \tag{3}$$

$$\begin{aligned} \Delta LAI(j) &= \alpha \Delta TT(j) LAI(j) \max[LAI_{\max} - LAI(j), 0] && TT(j) \le TT_{\mathrm{L}} \\ &= 0 && TT(j) > TT_{\mathrm{L}} \end{aligned} \tag{4}$$

The index j is the day. The model has a time step Δt of one day. The explanatory variables are TMIN(j), TMAX(j) and $I(t)$ which are respectively minimum and maximum temperature and solar radiation on day j. The parameters are T_{base} (the baseline temperature for growth), RUE (radiation use efficiency), K (excitation coefficient, which determines the relation between leaf area index and intercepted radiation), α (the relative rate of leaf area index increase for small values of leaf area index), $LAI_{\max}$ (maximum leaf area index), TT_{M} (temperature sum for crop maturity) and TT_{L} (temperature sum at the end of leaf area increase).

2.1. The elements of a dynamic system model

2.1.1. State variables U(t)

The state variables play a central role in dynamic system models. The collection of state variables determines what is included in the system under study. A fundamental choice is involved here. For example, if it is decided to include soil mineral nitrogen within the system being studied, then soil mineral nitrogen will be a state variable and the model will include an equation to describe the evolution over time of this variable. If soil mineral nitrogen is not included as a state variable, it could still be included as an explanatory variable, i.e. its effect on plant growth and development could still be considered. However, in this case the values of soil mineral nitrogen over time would have to be supplied to the model; they would not be calculated within the model. The limits of the system being modeled are different in the two cases.

The choice of state variables is also fundamental for a second reason. It is assumed that the state variables at time t give a description of the system that is sufficient for calculating the future trajectory of the system. For example, if only root depth is included among the state variables and not variables describing root geometry, the implicit assumption is that the evolution of the system can be calculated on the basis of just root depth. Furthermore, past values of root depth are not needed. Whatever effect they had is assumed to be taken into account once one knows all the state variables at time t.

Given a dynamic model in the form of Eq. (1), it is quite easy to identify the state variables. A state variable is a variable that appears both on the left side of an equation, so that the value is calculated by the model, and on the right side, since the values of the state variables determine the future trajectory of the system.

2.1.2. Explanatory variables and parameters ($X(t)$, θ)

The explanatory variables likewise imply a basic decision about what is important in determining the dynamics of the system. In the chapter on model evaluation, we will discuss in detail how the choice of explanatory variables affects model predictive quality. Briefly, adding additional explanatory variables has two opposite effects. On the one hand, added explanatory variables permit one to explain more of the variability in the system, and thus offer the possibility of improved predictions. On the other hand, the additional explanatory variables normally require additional equations and parameters which need to be estimated, which leads to additional error and thus less accurate predictions.

Explanatory variables and parameters can be recognized by the fact that they appear only on the right-hand side of Eq. (1). They enter into the calculation of the system dynamics but are not themselves calculated. The difference between explanatory variables and parameters is that explanatory variables are measured or observed for each situation where the model is applied, or are based on measured or observed values. Thus for example maximum soil water holding capacity is measured for each field, or perhaps derived from soil texture, which would then be the measured value. Potentially at least, an explanatory variable can differ depending on the situation while a parameter is by definition constant across all situations of interest.

2.2. The random elements in the dynamic equations

We have written the dynamic equations as perfect equalities. In practice however they are only approximations. The actual time evolution of a state variable in a system as complex as a crop–soil system can depend on a very large number of factors. In a crop model this is generally reduced to a small number of factors; those considered to be the most important. The form of the equation is also in general chosen for simplicity and may not be exact. Thus the equations of a crop model should actually be expressed as

$$U_i(t+\Delta t) = U_i(t) + g_i\,[U(t), X(t);\theta] + \eta_i(t), \quad i = 1,\ldots,S \tag{5}$$

where the error $\eta_i(t)$ is a random variable. This is a stochastic dynamic equation.

Another major source of uncertainty in the dynamic equations comes from the explanatory variables and in particular climate. When crop models are used for prediction, future climate is unknown and this adds a further source of uncertainty about the time evolution of the system.

3. A crop model is a response model

We can integrate the differential equations or difference equations of the dynamic system model. Often we talk of "running" the model when the equations are embedded in a computer program and integration is done numerically on the computer. For the difference equations, one simply starts with the initial values at $t = 0$ of the state variables, uses the dynamic equations to update each state variable to time $t = \Delta t$, uses the dynamic equations again to get the state variable values at $t = 2\Delta t$, etc. up to whatever ending time one has chosen.

The result of integration is to eliminate intermediate values of the state variables. The state variables at any time T are then just functions of the explanatory variables for all times from $t = 0$ to $t = T - \Delta t$ i.e. after integration the state variables can be written in the form

$$U_i(T) = f_{i,T}\,[X(0), X(\Delta t), X(2\Delta t), X(3\Delta t), \ldots, X(T-\Delta t);\theta], \quad i = 1,\ldots,S \tag{6}$$

In general, there are a limited number of model results that are of primary interest. We will refer to these as the model response variables. They may be state variables at particular times or functions of state variables. The response variables may include: variables that are directly related to the performance of the system such as yield, total nitrogen uptake or total nitrogen leached beyond the root zone; variables that can be used to compare with observed values, for example leaf area index and biomass at measurement dates; variables that help understand the dynamics of the system, for example daily water stress.

We note a response variable Y. According to Eq. (6) the equation for a response variable can be written in the form

$$Y = f(X;\theta) \tag{7}$$

where X stands for the vector of explanatory variables for all times from $t = 0$ to whatever final time is needed and θ is the same parameter vector as in Eq. (1). When we want to emphasize that the model is only an approximation, we will write $\hat{Y}$ in place of Y.

3.1. The random elements in the response equations

Since the dynamic equations are only approximate, the response equations are also only approximate. Including error, the equation for a response variable can be written

$$Y = f(X;\theta) + \varepsilon \tag{8}$$

where ε is a random variable. For the moment we ignore the uncertainty in X. Since the response equations derive directly from the dynamic equations, ε is the result of the propagation of the errors in Eq. (5). However, it is not obligatory to first define the errors in the dynamic equations and then derive the errors in the response equations. An alternative is to directly make assumptions about the distribution of ε. In this case, Eq. (8) is treated as a standard regression equation. If there are several response variables to be treated, then one is dealing with a (generally non-linear) multivariate regression model.

The error arises from the fact that the explanatory variables do not explain all the variability in the response variables, and from possible errors in the equations. In addition there may be uncertainties in the explanatory variables, in particular climate when the model is used for prediction.

4. Working with crop models. Which form?

In developing and working with crop models, both the dynamic equations and response equations are used, though for different objectives one will in general concentrate on one or the other.

During the initial development of a crop model one generally works with the dynamic equations. Several reasons have led to the use of dynamic crop models. First, we have a great deal of information about the processes underlying crop growth and development, and the dynamic equations allow us to use this information in studying the evolution of the overall crop–soil system. A second reason is that they allow us to break down the very complex crop–soil system into more manageable pieces and to model each of those pieces separately. It is possible to develop response models directly, without the intermediate step of dynamic equations. However, such models are in general limited to much simpler representations of a crop–soil system than is possible with dynamic crop models. The individual dynamic equations in crop models may also be quite simple, but their combination and interaction in the overall model results in complex response equations.

Historically, researchers have had two quite different attitudes towards crop models. On the one hand, a crop model can be considered a scientific hypothesis. Testing the hypothesis involves both forms of a crop model. The dynamic equations represent the hypothesis, which is tested by comparing the response equations with observations. The second attitude is that crop models are engineering tools. They are useful in relating outputs to inputs, but it is not necessary that the dynamic equations mimic exactly the way the system functions. The dynamic equations are simply a way of deriving useful input–output relationships. In this case, the response equations are of main interest and the evaluation of the model measures the quality of the input–output relationships. Evaluation is treated in Chapter 2.

Especially from an engineering perspective, the general behavior of the model responses as functions of the explanatory variables is of interest and importance. However, the response equations are in general not available as analytic expressions, but only after numerical integration of the dynamic equations. It is thus difficult to analyze the effect of input variables on response variables directly from the model equations. This has led to the use of sensitivity analysis, which is the study of how input factors (both explanatory variables and parameters) affect the outputs of a response model. This topic is treated in Chapter 3.

A major problem with crop models is obtaining the values of the parameters. The complexity of crop models means that there are in general, many (often a hundred or more) parameters. The amount of experimental data on the other hand is in general limited because experimentation on crop systems is necessarily lengthy and expensive in terms of land, equipment and manpower. If we consider just the response equations, then we have a regression problem involving simultaneously all the parameters in the model, and their estimation from the experimental data may be impossible or at least lead to large errors. However, the fact that a crop model has two forms often leads to additional information that can be used for parameter estimation. In particular, one often assumes that the dynamic equations have validity beyond the range of conditions described by the response model. This implies that one can do experiments on some processes under other conditions than those where the crop model will be used. For example, the temperature dependence of some processes may be studied in controlled temperature environments. The result is additional data, independent of the data on the overall system, that can be used to estimate parameters. The problem of parameter estimation for crop models is discussed in Chapter 4.

A specific problem related to crop models is that for each crop species there are in general many varieties, and plant breeders add new varieties each year. From a crop model perspective, this greatly exacerbates the problem of parameter estimation, since at least some of the model parameters vary from variety to variety. A possible solution is to use both the dynamic and response forms of a crop model. Some varietal parameters can be obtained from studies on the individual processes, others can be estimated from the response equations. This approach is discussed in Chapter 10. One can also treat this problem at a more fundamental level, by seeking to relate the model parameters more closely to genetic information (see Chapter 11).

A very promising approach to improving crop models is data assimilation, where one injects in-season data into the model and adjusts the values of the state variables or the parameters to that data. Assimilation is based on Eq. (5). It is necessary to have an estimate of error in the dynamic equations, in order to determine the respective weights to give to the data and to the model when combining those two sources of information. Data assimilation is treated in Chapter 5.

Testing different possible crop management strategies is a major use of crop models. One aspect of this use is mathematical optimization of management strategies. Chapter 6 presents two different approaches to optimization. Optimization by simulation is based on the response form of crop models. Here, management strategies are parameterized. Optimization consists of calculating the values of the management parameters that maximize an objective function, which in general depends on a small number of model response variables such as yield or grain protein content. The second approach treats optimization as a control problem. Here, the dynamic equations are used to calculate the transition probabilities from one time step to the next, as a function of explanatory variables including management decisions.

5. Conclusions

The fact that crop models exist in two forms, as dynamic equations and as response equations, is both a complication and an advantage. One complication is that in general this leads to quite complex models. A second is that model error must be treated at two levels, that of the dynamic equations and that of the overall system response.

The advantage is that the model can be developed and analyzed at two levels. One can study the individual processes and the overall system, and results from both can be integrated into the model. This allows us to profit from knowledge of how the system functions in order to better understand and manage crop–soil systems. The connection between processes and the overall system can also be used to test and improve our knowledge of the processes.

Exercises

1. Write equations for the response variables $B(2)$ and $LAI(2)$ using Eqs. (3) and (4). The resulting expressions should depend on explanatory variables, including the initial values of the state variables, but not on values of the state variables at other times.
2. On what explanatory variables does $B(2)$ depend? Compare with the explanatory variables in the dynamic equation for biomass. Explain the difference.
3. Let $\eta_{\mathrm{B}}(j)$ and $\eta_{\mathrm{L}}(j)$ represent respectively the errors in the dynamic equations for biomass and leaf area index at time j. Write the dynamic equations for biomass and leaf area index as stochastic equations using this notation.
4. Write the equations for the response variables $B(2)$ and $LAI(2)$ including the error terms $\eta_{\mathrm{B}}(j)$ and $\eta_{\mathrm{L}}(j)$ from the dynamic equations. The resulting expressions show how the errors in the dynamic equations propagate through the system model.
5. Let $\varepsilon_{\mathrm{B}}(2)$ and $\varepsilon_{\mathrm{L}}(2)$ represent the errors in the equations for the response variables $B(2)$ and $LAI(2)$. Write the equations for $B(2)$ and $LAI(2)$ as a multivariate regression model using this notation.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 2

Evaluating crop models

D. Wallach

1. **Introduction** **11**
2. **Comparing a model with data** **13**
3. **Evaluating the predictive quality of a model** **26**
4. **Evaluating a model used for decision support** **44**
5. **Using a model to test a scientific hypothesis** **47**

References **48**
Exercises **51**

1. Introduction

1.1. Definition

The dictionary definition of evaluation is to “ascertain the value of,” and that is the meaning that we use here. The goal of evaluation is to determine the value of a crop model, with respect to the proposed use of the model. The results of an evaluation study can include graphs comparing observed and predicted values, numerical measures of quality or qualitative conclusions about the quality of a model.

In the literature, one often encounters the term “validation” rather than “evaluation.” A rather common definition is that validation concerns determining whether a model is adequate for its intended purpose or not. This emphasizes the important fact that a model should be judged with reference to an objective. On the other hand, this definition seems to indicate that the result of a validation exercise is “yes” (the model is valid) or “no” (not valid). In practice, it is rarely the case that one makes, or even wishes to make, such a categorical decision. Rather one seeks a diversity of indications about how well the model represents crop responses. We therefore prefer the term “evaluation.”

General discussions and reviews of evaluation for ecological or crop models can be found in Swartzman and Kaluzny (1987), Loehle (1987), Mayer and Butler (1993), Rykiel (1996) and Mitchell and Sheehy (1997).

1.2. The importance of evaluation

Model evaluation is important for several reasons. Firstly, the simple fact of deciding to evaluate a model obliges one to answer some basic questions, including what is the objective of the model, what is the range of conditions where the model will be used, what level of quality will be acceptable.

Secondly, model improvement is impossible without evaluation. In the absence of a measure of model quality, how can one decide whether improvement is called for, and how can one know if a modified model is an improvement? As we shall see, evaluation can provide not only an overall indication of quality but can also quantify the errors resulting from different causes. Then, further efforts can be focused on reducing the major errors.

Finally, evaluation is important for potential users of a model. The user needs information about the quality of the model in order to decide how much credence to give to model results.

1.3. The role of evaluation in a modeling project

Evaluation should not be envisioned as an activity that is undertaken just once, at the end of a modeling project. It is rather an activity that, in its different forms, accompanies the project throughout its lifetime.

Evaluation should begin at the beginning of a modeling project. At that time it is necessary to identify the goals of the project, and consequently the criteria for model evaluation. It is at the beginning of the project that the range of conditions to be simulated is specified, that the output variables of interest are identified and that the acceptable level of error is defined.

A second evaluation step involves the model equations, which will be compared with results in the literature or with expert opinion. In general, the model is then embodied in a computer program and there is the essential step of evaluation of the computer program. Testing a computer program, to ensure that it performs the intended calculations, is an important field with its own literature and we will not consider it here.

Once the computer program exists, one often continues with sensitivity analysis (Chapter 3) and parameter estimation (Chapter 4). Both of these activities include elements of evaluation. Sensitivity analysis allows one to evaluate whether model response to input factors is reasonable, and also to see if the most important input factors are sufficiently well known. Parameter estimation normally includes indications of the quality of the parameter estimators, which is an important aspect of model quality as we shall see.

Once the parameter values are fixed, one can proceed to evaluate the model results. This is the subject of this chapter. In general modeling is an iterative exercise. If the model and/or the data are modified, then a new round of evaluation is required.

1.4. In this chapter

A first approach to the evaluation of model results is to compare model results with data. Various approaches and criteria are discussed in Section 2. The real objective for a model is often prediction. A criterion of predictive quality and its analysis are presented in Section 3. Another possible objective is to use the model as an aid to decision-making. In Section 4, we discuss how one can evaluate a model specifically with respect to this objective. The Sections 2–4 treat the model as an engineering model, which relates inputs to outputs. Evaluation here concerns how well the model reproduces input–output relationships, and is not concerned with the realism of the processes included in the model. In Section 5, we adopt a different point of view. Here, the model is used to test a hypothesis about how the system under study functions. We wish to test the hypothesis that the processes as described by the model are identical to the way the real world functions. Here, it is logical to use the term validation rather than evaluation.

2. Comparing a model with data

An essential part of model evaluation, probably the first aspect that comes to mind, is comparison of model predictions with observed data. This can help to identify problems with the model and give ideas for improvement. However, one must be careful in drawing general conclusions from the comparison of data and predictions. If the observed situations are not representative of the situations of interest, then comparison with past data may be different from the agreement with future measurements. Also, if the observed data have been used in model development, the degree of fit to that data is in general better than the agreement with future measurements. We will return to these problems when we consider predictive quality.

We first discuss graphical comparisons between measured and calculated values. Graphs are extremely useful for providing a quick visual summary of data and of the comparison between model and data. We then discuss numerical comparisons. Among the many different measures of agreement that have been proposed, we present those that seem to be most widely used or that offer some particular quality. We divide the measures into 4 groups: simple measures of the difference between observed and predicted values; measures which are normalized for easier interpretation; measures that can be decomposed into separate contributions, and which thus give additional information about the sources of error; measures based on a threshold of model quality. Note that for many of the measures of agreement between model and measurements, there is no standardized vocabulary. In order to avoid ambiguity, it is important to give the equation that is used in the calculations.

There is no single best method of comparison between a model and data (Table 2). Different comparisons highlight different features of the data and of model behavior. Therefore one should use a number of methods described below. The main difficulty is in obtaining the data and then in obtaining the corresponding model predictions. Once data and predictions are available, producing graphs or calculating measures of agreement is in general quite simple, which is another argument in favor of exploring several types of comparison. Software is also available to aid in evaluating model performance (Fila et al., 2003).

To illustrate the methods of this section, we use the model and data set of Example 1. This is not a dynamic crop model, but rather a static linear model between some output and 5 input variables. It is legitimate to use this simple model because in fact the comparisons that we illustrate are not specific to crop models but rather apply generally to any model.

2.1. Graphical representations of model error

2.1.1. Graph of model predictions versus observed values

Probably the most widespread graphical presentation of the agreement between measured and calculated values for crop models is a plot as in Figure 1. For each measurement, the

Example 1

Suppose that our model for predicting response Y for individual i is

$$\hat{Y}_i = f(X_i;\theta) = \hat{\theta}^{(0)} + \hat{\theta}^{(1)}x_i^{(1)} + \hat{\theta}^{(2)}x_i^{(2)} + \hat{\theta}^{(3)}x_i^{(3)} + \hat{\theta}^{(4)}x_i^{(4)} + \hat{\theta}^{(5)}x_i^{(5)} \tag{1}$$

The explanatory variables are $X = (x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)})^{\mathrm{T}}$. The parameter vector is

$$\hat{\theta} = (\hat{\theta}^{(0)},\hat{\theta}^{(1)},\hat{\theta}^{(2)},\hat{\theta}^{(3)},\hat{\theta}^{(4)},\hat{\theta}^{(5)})^{\mathrm{T}} = (1.9, 7.8, 2.5, -0.2, 0.1, 0.7)^{\mathrm{T}} \tag{2}$$

The hat notation is used to indicate that the parameter values are estimates. We do not need to bother here with the origin of these estimates. The data set for evaluating the model is given in Table 1.

Table 1. Measured values (Y_i), 5 explanatory variables, calculated values ($\hat{Y}_i$) and model errors (D_i) for 8 situations.

i	Y_i	$x_i^{(1)}$	$x_i^{(2)}$	$x_i^{(3)}$	$x_i^{(4)}$	$x_i^{(5)}$	$\hat{Y}_i$	$D_i = Y_i - \hat{Y}_i$
1	−9.39	−1.63	0.80	0.44	−0.45	−0.47	−9.31	−0.08
2	−3.23	−0.95	1.07	0.50	0.53	−0.33	−3.10	−0.13
3	0.37	−0.25	0.20	0.51	0.82	0.45	0.72	−0.35
4	7.10	0.38	1.02	−0.22	0.82	−1.93	6.18	0.92
5	5.82	0.15	1.14	1.07	1.63	−0.55	5.45	0.38
6	−11.21	−1.20	−1.79	0.35	−0.26	0.26	−11.84	0.63
7	−5.81	−0.97	−0.15	0.11	1.13	0.10	−5.90	0.09
8	2.82	0.39	−1.20	2.01	−0.79	−1.44	0.47	2.35

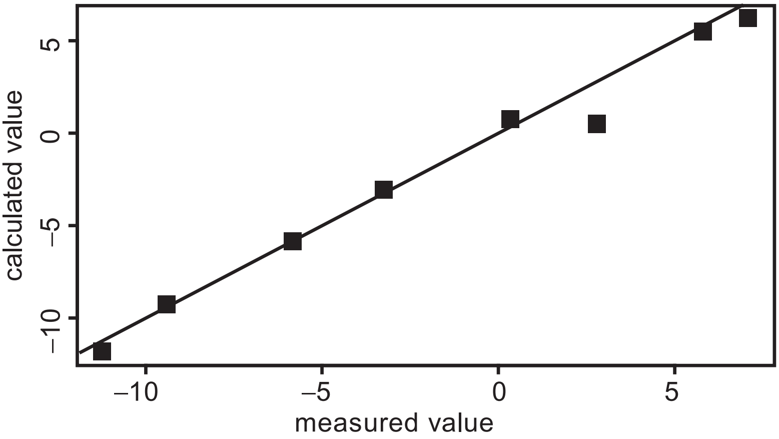


Figure 1. Calculated *versus* measured values, using the model and data of Table 1.

x value is the measured value and the *y* value is the corresponding calculated value. It is also usual to draw the 1:1 line on such a graph. If there is no model error the calculated values and measured values are identical and then each point will be exactly on the 1:1 line.

The advantage of this type of graph is that one can see at a glance how well model calculations and measurements agree. However, a word of caution is required. There may be a tendency to underestimate the level of error. The eye tends to evaluate the shortest distance from the point to the 1:1 line, but model error is given by the vertical or equivalently the horizontal distance from the point to the line.

In some cases, one shows the regression line between calculated and measured values in addition to, or instead of, the 1:1 line. Calculating the regression can be of interest, as we shall see. However, to include the regression line on a graph of this type seems more misleading than useful. If the points fall close to the regression line one can have the impression that the model is quite good, but in fact that is not a measure of model error.

A graph of measured *versus* calculated values can be used not only for output variables that have a single value for each situation (for example, yield), but also for output variables that are functions of time (for example, LAI). There may then be several points for each situation (for example, several measurements of LAI at different dates).

As with any graph, it can be useful to distinguish (for example by using different symbols) different groups of points. For example, if there are several measurements of LAI per field, the use of different symbols for each field will make it easier to see if there is a field effect on error. Another example would be where different levels of fertilizer are studied. The use of different symbols for different levels will allow one to see if model error tends to be different depending on fertilizer level.

2.1.2. Measured and calculated values versus time or other variable

In the specific case of an output variable which is a function of time, it is fairly common practice to graph measured and calculated values *versus* time, for each situation separately. In general, the model produces calculated values every day while the measurements are much sparser. In this case, the graph takes the form of a regular dotted line (one dot each day) for the calculated values, which can be compared with the occasional points for the measurements. For example, Robertson et al. (2002) present graphs of this type for comparing measured and calculated values of aboveground biomass and observed and calculated values of leaf biomass. One can present more than one situation in each graph (Robertson et al., 2002 present 2 situations in a graph), but such graphs quickly become cluttered as more situations are represented.

Mayer and Butler (1993) discuss the difficulty of visual evaluation of model error based on graphs like this when the output variable fluctuates with time. This is often the case, for example for soil moisture. They present an artificial example where the "measurements" are generated by a random mechanism, with no relation to the model. Nonetheless, a rapid visual examination of the data seems to indicate that the model is "reasonable." The reason is that we tend to focus on the smallest distance between the measured points and the calculated curve. However, error is in fact given by the vertical distance between the measured value and the calculated curve. In the artificial example, there is always some part of the calculated curve fairly close to the "measured" values because of the fluctuations in the calculated values. This is not to say that graphs of measured and calculated values *versus* time are not useful. The conclusion is rather that they must be analyzed carefully. The fact that such graphs may be difficult to interpret is an additional reason for using several different types of graphs for assessing model agreement with data.

The above discussion concerns graphs with time on the abscissa. It may also be of interest to graph measured and calculated values *versus* some other variable. For example, Pang et al. (1997) present total absorbed nitrogen as a function of applied nitrogen. The general objective is to see how model error varies with the dependent variable.

2.1.3. Graphing the residues

The classical method of examining model error in statistics is to plot model error (the difference between measured and calculated values) on the y axis, against measured values on the x axis. This type of graph is seen seldom for crop models, which is very unfortunate. The data of Table 1 are plotted in this way in Figure 2. The advantage of this type of graph, compared to Figure 1, is that the model errors appear directly. They are thus easier to evaluate and to compare. For example, consider the point for individual 8 with the response $Y = 2.82$. One can see from Figure 1 that this point has the largest error, but the size of this error relative to the others stands out more clearly in Figure 2.

Residue graphs are very important for bringing attention to systematic patterns in the errors. For example, a residue graph might show that the errors in yield are exceptionally large for very small observed yield values. This might suggest analyzing in detail the way the model handles extreme stresses.

It is also of interest to plot model error *versus* explanatory variables such as total water input (rainfall plus irrigation), date of sowing, total applied nitrogen, etc. If the

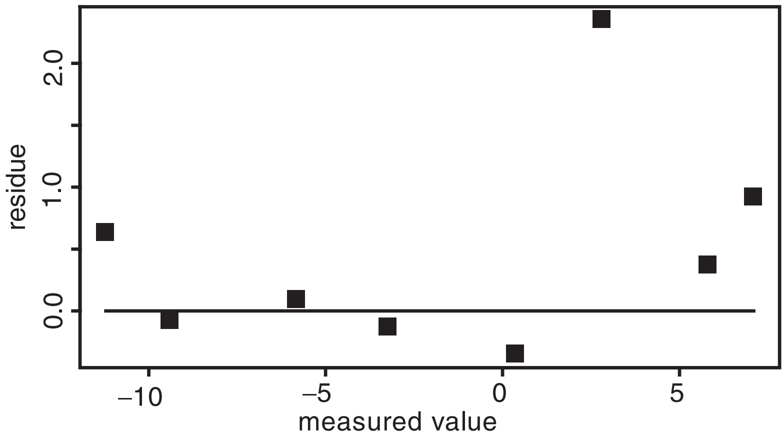


Figure 2. Residues for model and data of Table 1.

model is correctly specified, there should be no trends in the residues. If the residues do show some systematic trend, then there is an effect of the explanatory variable which has not been taken into account in the model. Chapter 12 presents examples of this type of analysis.

The residue graph will, furthermore, give indications about the variability of model error, which is important for parameter estimation. The simplest assumption is that model errors have zero mean and constant variance. Residue graphs allow one to examine visually whether such an assumption is reasonable. Specifically, the residues should then be centered on zero and have roughly the same spread for different values of the variable against which they are plotted.

2.2. Simple measures of agreement between measured and calculated values

The basic quantity for measuring the agreement between model and observations is the difference between the two, noted

$$D_i = Y_i - \hat{Y} \tag{3}$$

where Y_i is the measured value for situation *i* and $\hat{Y}_i$ is the corresponding value calculated by the model. The output variable Y can be any model output, for example yield, LAI 30 days after emergence, days to flowering, etc. It is the differences D_i that are plotted in a graph of residues.

A very simple way to summarize the D_i values for several situations is to calculate their average, also known as model bias.

$$Bias = \frac{1}{N}\sum_{i=1}^{N} D_i \tag{4}$$

where N is the total number of situations. The *bias* measures the average difference between measured and calculated values. If on the average the model under-predicts, the *bias* is positive, and conversely if the model over-predicts on the average, the *bias* is negative. The interpretation is thus very simple, which makes this measure useful as a guide for model improvement. For example, if yield is systematically under-predicted (positive *bias*), one might start by examining whether final biomass or harvest index or both are under-predicted. If one has information on yield components, one can examine individually seed number and weight per seed to see which is under-predicted.

Bias alone, however, is not sufficient as a summary of model errors. A *bias* value near zero may be the consequence of very small model errors in all situations, or alternatively of large errors that approximately cancel each other between under- and over-prediction. The interpretation of a sizeable positive or negative *bias* is also subject to some ambiguity. A positive *bias* can arise because the model systematically under-predicts, or because the model both under- and over-predicts but with a preponderance of under-prediction. Negative *bias* has an analogous ambiguity.

For the data of Example 1, $bias = 0.48$. An examination of the data shows that the model rather systematically under-predicts, which is also apparent from Figure 2.

There are two classical measures of agreement that eliminate the problem of compensation between under- and over-prediction. The first and most widely used is the mean squared error, defined as

$$MSE = (1/N)\sum_{i=1}^{N}(D_i)^2 \tag{5}$$

Often it is convenient to work with the square root of *MSE*, called the root mean squared error;

$$RMSE = \sqrt{MSE} \tag{6}$$

The advantage is that *RMSE* has the same units as Y and thus is easier to understand.

Because *MSE* is an average of squared differences, large differences are heavily weighted. It is worthwhile to verify if *MSE* is not essentially due to one or two large differences. If this is the case, it might be more astute to examine those specific cases (problem with the data? exceptional circumstances such as extreme stress?) rather than the overall model.

For the data of example 1, $MSE = 0.88$ and $RMSE = 0.94$. The largest error $D_8 = 2.35$ contributes 78% of the total value of *MSE*. This is a case where one might start by examining the situation with large error.

The second measure which avoids compensation between under- and over-prediction is the mean absolute error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|D_i|$$

The units of *MAE* are the same as for Y. Furthermore, there is no over-weighting of large differences here. Thus *MAE* has advantages over *MSE* or *RMSE*, if the objective is simply to examine overall model error. On the other hand, the important advantage of *MSE* is that it can be decomposed into separate contributions, which is useful in identifying the sources of error.

For the data of Example 1, $MAE = 0.62$. The largest error, D_8, only contributes 48% of the total here.

A variant of the above measures is obtained by dividing *RMSE* by the average of the observed values. The relative root mean squared error is then

$$RRMSE = \frac{RMSE}{\bar{Y}}$$
$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad (7)$$

where $\bar{Y}$ is the average of the Y_i values. Robertson et al. (2002) for example calculate $RRMSE = 23\%$ for their data on peanut yield, using the Agricultural Production Systems Simulator (APSIM) model. An advantage of *RRMSE* is that it seems more meaningful than *RMSE* for comparing errors based on different data sets. Another advantage is that *RRMSE* is independent of the units used to measure *Y*. *RRMSE* will have the same value whether yield is measured in kg/ha or t/ha.

Mayer and Butler (1993) propose a relative mean absolute error,

$$RMAE = \frac{1}{N}\sum_{i=1}^{N} \frac{|Y_i - \hat{Y}_i|}{|Y_i|} \qquad (8)$$

Note that here one divides each difference by the corresponding observed value.

For the data of Example 1, the relative root mean squared error is $RRMSE = -56\%$. The value is negative because $\bar{Y}$ is negative, and large because $\bar{Y}$ is small. The relative mean absolute error is $RMAE = 0.26$.

In all of the above formulas, each observation enters just once, with no difference in weighting for different observations. This may not always be appropriate. For example, in the case of spatial data, one might want to weight each point by the area it represents (Willmott et al., 1985).

2.3. Normalized measures

Here, we consider distance measures which have an upper and/or lower bound. Such measures are easily interpreted and can be particularly convenient for comparing completely different cases (different data, different models).

Probably the most widely used measure of this type is the modeling efficiency, defined as

$$EF = 1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} \qquad (9)$$

Garnier et al. (2001) provide an example of the use of this measure. If the model is perfect, then $Y_i = \hat{Y}_i$ for each situation i and $EF = 1$. If one uses the average of observed values as the predictor for every case, so that $\hat{Y}_i = \bar{Y}$ for all i, then $EF = 0$. Thus, a model that gives $EF = 0$ has the same degree of agreement with the data as using the average to predict for every situation. A crop model with EF close to 0 would not normally be considered a good model. There is in general no lower limit to EF. A model can be a worse predictor than the average of observed values ($EF < 0$).

A second normalized measure that is sometimes used is the correlation coefficient between measured and calculated values defined by

$$r = \frac{\hat{\sigma}_{Y\hat{Y}}}{\hat{\sigma}_Y^2 \hat{\sigma}_{\hat{Y}}^2} \tag{10}$$

where $\hat{\sigma}_Y^2$, $\hat{\sigma}_{\hat{Y}}^2$ and $\hat{\sigma}_{Y\hat{Y}}$ are sample estimates of the variance of Y, the variance of $\hat{Y}$ and the covariance of Y and $\hat{Y}$ respectively.

$$\hat{\sigma}_Y^2 = \frac{1}{N}\sum_{i=1}^{N}[(Y_i - \bar{Y})^2]$$

$$\hat{\sigma}_{\hat{Y}}^2 = \frac{1}{N}\sum_{i=1}^{N}[(\hat{Y}_i - \bar{\hat{Y}})^2]$$

$$\hat{\sigma}_{Y\hat{Y}} = \frac{1}{N}\sum_{i=1}^{N}[(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})]$$

$$\bar{\hat{Y}} = \frac{1}{N}\sum_{i=1}^{N}\hat{Y}_i$$

The range of r is $-1 \leq r \leq 1$. A value of $r = 1$ indicates that there exists a perfect linear relationship between $\hat{Y}_i$ and Y_i. Note, however, that this does not necessarily imply that the model is perfect. Suppose for example that $\hat{Y}_i = 0.1\ Y_i$ for all i. Then $r = 1$, but in fact the model systematically predicts values that are smaller by a factor of 10. Thus r by itself is not a good measure of how well a model agrees with measurements. Addiscott and Whitmore (1987) suggest that one should use both bias and r, in order to have measures that concern two different aspects of model quality. A good model would have both small bias and an r value close to 1. The idea of using more than one measure, in order to bring out different aspects of model agreement, is important. In the next section we will go into this in more detail.

Willmott (1981) propose an agreement index defined as

$$index = 1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} \left(|\hat{Y}_i - \bar{Y}| + |Y_i - \bar{Y}|\right)^2} \tag{11}$$

The numerator is the mean squared error *MSE*. The denominator is related to the variability in the measured and in the calculated values. If the model is perfect, then $Y_i = \hat{Y}_i$ and $index = 1$. If the model predictions are identical in all cases and equal to the average of the observed values, i.e. $\hat{Y}_i = \bar{Y}$, then $index = 0$. These limiting values are the same as for *EF*, but for other cases, the two criteria will have different values.

For the model and data of Example 1, $EF = 0.98$ and $index = 0.99$. Here the values of *EF* and of *index* are very similar, and both seem to indicate that the model is much better than just using the average of the observations as predictor. This can also be seen from Figure 1, which shows that much of the variability in the observations is tracked by the model predictions. However, we have also examined the residuals and so we know that in at least one case the residual is actually quite large.

2.4. Measures for identifying different types of error

Certain measures of agreement between measured and calculated values can be decomposed into different contributions to the overall error. The effort for model improvement can then be concentrated on the dominant source of error.

Kobayashi and Salam (2000) show that *MSE* can be decomposed as

$$MSE = (Bias)^2 + SDSD + LCS \tag{12}$$

with

$$SDSD = (\sigma_Y - \sigma_{\hat{Y}})^2$$

$$LCS = 2\sigma_Y \sigma_{\hat{Y}}(1 - r)$$

r is the correlation coefficient defined in Eq. (10), σ_Y^2 and $\sigma_{\hat{Y}}^2$ are the variances of the measured and calculated variables respectively and their square roots, σ_Y and $\sigma_{\hat{Y}}$, are the corresponding standard deviations.

The first term in Eq. (12) is the bias squared. The cause of model bias is in many cases relatively easy to identify and perhaps to correct. The second term in the decomposition is related to the difference between the standard deviation of the measurements and the standard deviation of the calculated values. Once again, the causes of the difference can sometimes be identified. For example, if the model predicts that yield for different

Table 2. Measures of agreement between a model and measured data.

Name	Equation
Bias	$Bias = \frac{1}{N}\sum_{i=1}^{N} D_i$
Mean squared error	$MSE = (1/N)\sum_{i=1}^{N}(D_i)^2$
Root mean squared error	$RMSE = \sqrt{MSE}$
Mean absolute error	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert D_i\rvert$
Relative root mean squared error	$RRMSE = \frac{RMSE}{\bar{Y}}$
Relative mean absolute error	$RMAE = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert Y_i - \hat{Y}_i\rvert}{\lvert Y_i\rvert}$
Modeling efficiency	$EF = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$
Correlation coefficient	$r = \frac{\sum_{i=1}^{N}[(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})]}{\sqrt{\sum_{i=1}^{N}[(Y_i - \bar{Y})^2]\sum_{i=1}^{N}[(\hat{Y}_i - \bar{\hat{Y}})^2]}}$
Agreement index	$index = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(\lvert\hat{Y}_i - \bar{Y}\rvert + \lvert Y_i - \bar{Y}\rvert)^2}$
Concordance correlation coefficient	$\rho_c = \frac{2\sigma_{Y\hat{Y}}}{\sigma_Y^2 + \sigma_{\hat{Y}}^2 + (\mu_{\hat{Y}} - \mu_Y)^2}$
Total deviation index	$TDI(p)$ = minimal value of d such that $\lvert D_i\rvert \le \lvert d\rvert$ for at least $p\%$ of the observed situations.
Coverage probability	$CP(d)$ = the smallest value of p such that, for a percentage p of the observed situations, $\lvert D_i\rvert \le \lvert d\rvert$

situations varies only slightly whereas the measurements show a larger variation, one might look at the effect of stress. Is the difference due to the fact that the calculated values are not sufficiently sensitive to water stress, for example? The last term in the decomposition is related to the correlation between observed and predicted values. This term depends in detail on how well the model mimics the observed variation of Y from situation to situation. As such, it may often be the result of many small errors rather than a single major error, and thus be relatively difficult to analyze and correct.

For the data of Example 1, Eq. (12) gives MSE (0.88) $= Bias^2(0.23) + SDSD$ (0.06) $+$ LCS (0.59). The major source of error is the LCS term, whose origin is often difficult to ascertain. However, the squared bias term represents about one quarter of the total mean squared error, so model improvement could begin by searching for the origin of the bias.

Gauch et al. (2003) suggest that it would be advantageous to have a decomposition with terms explicitly related to the regression of Y on $\hat{Y}$. The decomposition they propose is

$$MSE = (Bias)^2 + NU + LC \tag{13}$$

$$NU = (1 - b_{Y\hat{Y}})^2 \sigma^2_{\hat{Y}}$$

$$LC = (1 - r^2)\sigma^2_Y$$

$$b_{Y\hat{Y}} = \frac{\sigma^2_{Y\hat{Y}}}{\sigma^2_{\hat{Y}}}$$

The term $b_{Y\hat{Y}}$ is the slope of the regression of Y on $\hat{Y}$. The decomposition of Eq. (13) is quite similar to that of Eq. (12). The first term is again the squared bias, and the last depends in detail on how variation in Y and $\hat{Y}$ are correlated. The second term, NU, depends on how close the slope of the regression of Y on $\hat{Y}$ is to 1.

For the data of Example 1, Eq. (13) gives MSE (0.88) $= Bias^2(0.23) + NU$ (0.04) $+$ LC (0.62). The numerical results, and conclusions, are similar to those for the decomposition of Eq. (12).

Willmott (1981) proposed a decomposition of MSE based on the linear regression of $\hat{Y}$ as a function of Y. The result is a regression equation $\tilde{\hat{Y}}_i = a + b_{\hat{Y}Y} Y_i$ where $\tilde{\hat{Y}}_i$ is the value of $\hat{Y}_i$ calculated from the regression model. The regression parameters are given by the standard formulas for linear regression,

$$b_{\hat{Y}Y} = \frac{\sigma_{Y\hat{Y}}}{\sigma^2_Y}$$

$$a = \bar{\hat{Y}} - b_{\hat{Y}Y}\bar{Y}$$

The decomposition is then

$$MSE = MSE_{\text{s}} + MSE_{\text{u}} \tag{14}$$

with $MSE_{\text{s}} = (1/N)\sum(\tilde{\hat{Y}}_i - Y_i)^2$ and $MSE_{\text{u}} = (1/N)\sum(\hat{Y}_i - \tilde{\hat{Y}}_i)^2$

The term MSE_s is called the systematic part of MSE and MSE_u the unsystematic part.

To understand the first term, suppose that $a = 0$ and $b = 1$ so that the regression line is the 1:1 line. Then $\hat{\hat{Y}}_i = Y_i$ and $MSE_s = 0$, i.e. the systematic contribution is zero in this case. In general, MSE_s is a measure of how far the regression line deviates from the 1:1 line. The MSE_u term on the other hand measures the variability of $\hat{Y}_i$ around the regression line. The systematic term is the contribution that is likely to be relatively easy to analyze and perhaps correct. An example of the use of this decomposition is given in Ben Nouna et al. (2000).

For the data of Example 1, $a = -0.55$, $b = 0.95$, $MSE_s = 0.31$, $MSE_u = 0.57$ and $MSE = 0.88$.

Some authors propose using a statistical test of the hypothesis H_0: $a = 0$, $b = 1$. If H_0 is true then the regression line is the 1:1 line. However, the above decomposition shows clearly that this test is related to only part of the model error. If the variability of the predicted values around the regression line is large then MSE_u and therefore MSE will be large, regardless of the values of a and b. Thus, the hypothesis should not be regarded as testing overall model quality. For this and other reasons, several authors (for example Mitchell, 1997) have criticized this test as a basis for judging the quality of a model.

A third decomposition of the error is proposed by Lin et al. (2002). This is not based on mean squared error but rather on a measure of error which they call the "concordance correlation coefficient," ρ_c, defined by

$$\rho_c = \frac{2\sigma_{Y\hat{Y}}}{\sigma_Y^2 + \sigma_{\hat{Y}}^2 + (\mu_{\hat{Y}} - \mu_Y)^2} \tag{15}$$

where $\sigma_{Y\hat{Y}}$, σ_Y^2 and $\sigma_{\hat{Y}}^2$ have been defined above and μ_Y and $\mu_{\hat{Y}}$ are respectively the averages of the Y and $\hat{Y}$ values. If $Y_i = \hat{Y}_i$ for all i, then $\mu_Y = \mu_{\hat{Y}}$, $\sigma_{Y\hat{Y}} = \sigma_Y^2 = \sigma_{\hat{Y}}^2$ and so $\rho_c = 1$. At the other extreme, if the observed and predicted values are completely uncorrelated ($r = 0$) then $\sigma_{Y\hat{Y}}$ is zero and so $\rho_c = 0$. Lin et al. (2002) propose this measure for comparing two populations, but it can also be used here for comparing two samples (measured and calculated values).

ρ_c can be decomposed as $\rho_c = r\chi_a$. The first term is the correlation coefficient. The second is

$$\chi_a = \frac{2}{(\sigma_{\hat{Y}}/\sigma_Y) + (\sigma_Y/\sigma_{\hat{Y}}) + (\mu_{\hat{Y}} - \mu_Y)^2/(\sigma_{\hat{Y}}\sigma_Y)}$$

It is this term that represents the systematic part of the error. More precisely, χ_a is equal to 1 if and only if the mean of the measured values is equal to the mean of the calculated values and also the variance of the measured values is equal to the variance of the calculated values.

For the data of Example 1, $\rho_c = 0.9888$ and the decomposition into two factors gives $\chi_a = 0.9963$ and $r = 0.9925$. In this particular case, ρ_c is very close to the best value of 1.0, which limits the usefulness of the decomposition.

2.5. Measures based on a threshold of model quality

In some cases, one might accept a few large differences between measured and calculated values as long as the agreement is good for the majority of situations. The two measures presented in this section are adapted to this viewpoint.

In the first measure, one fixes the percentage p of situations which should show acceptable agreement. The measure is the total deviation index;

$TDI(p)$ = minimal value of d such that $|D_i| \leq |d|$ for at least $p\%$ of the observed situations.

The second measure in this group fixes a maximum error $|d|$ and measures the percentage of situations with errors smaller than that of threshold. This is called the coverage probability, defined as

$CP(d)$ = the largest value of p such that, for a percentage p of the observed situations, $|D_i| \leq |d|$

For the data of Example 1, if we set $p = 80\%$, then $TDI\ (p = 80) = 0.93$, i.e. at least 80% of the observed situations (in fact 7 situations out of 8 or 87.5%) have $|D_i|$ values that are less than or equal to 0.93.

Fixing the threshold of error at $|d| = 0.5$, we have $CP\ (d = 0.5) = 62.5\%$, i.e. 62.5% of the model errors (5 out of 8) are smaller than or equal to 0.5 in absolute value.

2.6. Treating complex output variables

The measures presented above apply directly to outputs like yield or time to flowering, which have a single value for each situation. For more complex types of output, it is not always clear how to apply the above formulas.

Consider first output variables that are functions of time, such as LAI or soil moisture or root depth. One possibility is simply to apply the above measures to all of the observations, ignoring the fact that the observations are structured by situation, with several observations corresponding to the same situation. However, the values at different times may be very different, with small values shortly after emergence and much larger values later in the season. This would often be the case for LAI and root depth, for example. When this occurs, the results can be quite different depending on whether we consider the errors themselves or relative errors (errors divided by the observed values). It will then be important to choose the more meaningful measure of agreement. A second note of caution concerns the distribution of observations among situations. If a few situations have most of the observations, then the measures of agreement may essentially concern those few situations. If this seems to be a problem, it might be worthwhile to first calculate the measure of agreement for each situation and then average over situations. In this way, each situation has the same weight.

A different difficulty occurs if the output is not a single variable, but rather a distribution of values for each situation. An example would be the fraction of fruits in different size classes (see Chapter 13). One possibility here is to convert to a single output variable. For example, from the distribution of fruit sizes one could calculate an average fruit size. Then one can apply the measures that are appropriate to a single output variable per situation. In other cases, the major interest is in the distribution itself, and so one wants a measure of model agreement that specifically measures how well the calculated and observed distributions agree. A useful measure in this case would be the statistic used in the Kolmogorov–Smirnov test for comparing two distributions (Sokal and Rolf, 1981). Let H be the variable that is being subdivided into classes, for example fruit size. Let $F_H(h)$ be the fraction of observed items with $H \leq h$, and $\hat{F}_H(h)$ be the fraction of items with $H \leq h$ according to the model. Let $H_{\max}$ be the maximum value of $|F_H(h) - \hat{F}_H(h)|$. $H_{\max}$ is the value of our measure of agreement. If there are several situations, one might use the value of $H_{\max}$ averaged over situations.

The final special case that we mention is that of stochastic models. In this case, each run of the model gives a different value of the output variables. The problem then is to compare the distribution of values calculated by the model with the single observed value for each situation. Waller et al. (2003) suggest that in this case the question is not whether the model is consistent with the measurements, but rather whether the measurements fall within the observed variability of the model. They suggest a Monte Carlo approach to evaluate the probability of a value as large as or larger than the measured value, assuming that the measured value is drawn from the distribution of calculated values.

3. Evaluating the predictive quality of a model

3.1. Introduction

In the preceding section, we presented measures that summarize the agreement between the model and past measurements. In general, however, our real interest is not in how well the model reproduces data that has already been measured, but rather in how well it can predict new results. The assumption, often implicit, underlying the use of past measurements is that the agreement of the model with those data can inform us about how the model will perform in the future. However, that assumption is not always founded. In this section, we consider in detail the definition, analysis and estimation of prediction error.

3.2. A criterion of prediction quality

The standard criterion of prediction quality in statistics is the mean squared error of prediction or *MSEP*. For a model with fixed parameter vector $\hat{\theta}$, *MSEP* is defined as

$$MSEP(\hat{\theta}) = E\{[Y - f(X; \hat{\theta})] | \hat{\theta}\}^2 \quad (16)$$

This is the squared difference between the observations Y and the corresponding values calculated with the model, averaged over situations of interest. The notation $|\hat{\theta}$ means that the parameter vector estimator is treated as fixed, so the expectation is not over possible values of the parameters. The notation $MSEP(\hat{\theta})$ emphasizes that the mean squared error of prediction is specific to the parameter vector that is used in the model.

The definition of $MSEP(\hat{\theta})$ is superficially very similar to that for MSE in Eq. (5). Both involve the squared difference between true and calculated values. However, MSE concerns just the situations that have actually been measured while $MSEP(\hat{\theta})$ concerns all possible situations of interest. This implies that MSE can be a poor estimator of $MSEP(\hat{\theta})$ for two reasons. First of all, if the measured situations are not representative of the full range of situations of interest, then MSE may be very different than $MSEP(\hat{\theta})$. Secondly, if MSE involves data that was used for model development, then model error for those data will not be representative of model error for other situations of interest. We will go into both of these subjects in more detail.

The units of $MSEP(\hat{\theta})$ are the units of Y squared. Often one uses the root mean squared error of prediction in order to deal with a quantity that has the same units as Y. The definition is

$$RMSEP(\hat{\theta}) = \sqrt{MSEP(\hat{\theta})}$$

3.2.1. A criterion for the prediction of a time-dependent variable

The definition in Eq. (16) assumes that there is a single model output of interest, like yield or days to flowering. Suppose, however, that one is interested in predicting a model output that varies with time such as leaf area index. A criterion often used in this case is the integrated mean squared error of prediction, defined as

$$IMSEP(\hat{\theta}) = E\left\{\int [Y(t) - f(t, X; \hat{\theta})]^2 \mathrm{dt}\right\}$$

where we have shown explicitly the time dependence of Y and of the model predictions. For crop models with a time step of 1 day, the integral would be replaced by a sum over days. Wallach et al. (1990) studied a very similar criterion for evaluating models of nitrogen uptake over time by the root systems of young peach trees.

3.2.2. Prediction for what range of conditions?

A prerequisite to evaluating the predictive quality of a model is to specify the situations for which predictions will be made. We will speak of the “target distribution” to refer to the distribution of situations of interest.

Often the target distribution is defined implicitly, by describing the physical types of situations that are of interest. For example, the target distribution might be fields with irrigated corn in southwestern France, with standard management practices. This defines a joint distribution of soil characteristics, weather, initial conditions, management practices, weed, disease and pest levels, etc.

The definition of a target distribution is very important. It is closely related to the notion that a model should be evaluated in relation to the projected use of the model. If for example, the model is intended for use with corn that is irrigated to obtain near-potential yields, the model should be evaluated for such situations. The target distribution will then not include situations with extreme water stress. The value of $MSEP(\hat{\theta})$ for the same model might be quite different depending on the target distribution.

It is useful to distinguish two aspects of the target distribution. First of all, the target distribution concerns all the explanatory variables that appear in the model. This will often include initial conditions, soil characteristics, daily climate and management variables such as sowing date and density, irrigation dates and amounts, etc. The target distribution also concerns all variables not in the model that affect the output variables in question. This might include pest damage, disease incidence, initial phosphorous level, the spatial variability of soil characteristics, etc.

We need not only define the target distribution, we also need to sample from it or generate values from it. To estimate $MSEP(\hat{\theta})$, we need a sample from the target distribution. For crop models two difficulties often arise. First, often one does not draw situations independently from the target distribution. Very often one has measurements from several fields in the same year, and/or measurements in the same field over several years. In the first case, the year and thus the climate are not chosen independently for each field. In the second case, fields are not chosen independently. The structure of the data set may be quite complex and difficult to take into account in the estimation of $MSEP(\hat{\theta})$. The second difficulty is that in general the number of situations sampled is fairly small, while the diversity of situations in the target distribution may be very large. As a result, even if we have a truly random sample, whole sections of the target distribution may be missing from it. For example, the target distribution may include average spring temperatures that cover a wide range, while the available sample only includes years with warm spring weather. Another example would be where the target distribution includes a range of soil depths, while the shallow soils are not represented in the available data. In such cases, one must be aware of the limitations of the sample, and be very wary of drawing conclusions about situations that are far removed from those sampled.

We will also want to generate samples of model explanatory variables representative of the target distribution. A simple assumption is that initial conditions, weather, soil and management decisions are independent. Then we can generate samples from each independently. A more complex but often more realistic assumption is that the management decisions depend on the other variables, through decision rules (Chapter 6). Then one would first generate the other variables and from them deduce the management decisions.

Example 2

The Y values in Table 1 were generated using the relationship

$$\begin{aligned} &Y = \theta^{(0)} + \theta^{(1)}x^{(1)} + \theta^{(2)}x^{(2)} + \theta^{(3)}x^{(3)} + \theta^{(4)}x^{(4)} + \theta^{(5)}x^{(5)} + \varepsilon \\ &\theta = (\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}, \theta^{(5)})^{\mathrm{T}} = (2, 8, 2, 0.05, 0.01, 0.002)^{\mathrm{T}}. \end{aligned} \tag{17}$$

where ε has a normal distribution $\varepsilon \sim N(0, 0.04)$. The 8 values of ε required for Table 1 were drawn independently from the distribution of ε. We will use this true relationship between

X and Y to evaluate $MSEP(\hat{\theta})$ for various models. Of course this is only possible in an artificial example like this one. In practice the true relationship between X and Y is unknown, and so $MSEP(\hat{\theta})$ cannot be calculated exactly but only estimated.

Note what is meant by a "true" relation between Y and X. Equation (17) does not allow us to calculate Y exactly, given X. The extent to which Y is not completely determined by X is represented by the random variable ε. The relationship is "true" in the sense that we have given the true distribution of ε.

We will also need the target distribution for the explanatory variables X. We suppose that the components of X are independent. Then the joint distribution is the product of the distributions for each component of X, so that $f_X(X) = f_{X_1}(x_1)f_{X_2}(x_2)f_{X_3}(x_3)f_{X_4}(x_4)f_{X_5}(x_5)$. We also assume that $f_{X_i}(x_i) \sim N(0, 1)$ for $i = 1, \dots, 5$. Finally, we assume that X and ε are independent.

We can now calculate $MSEP(\hat{\theta})$ for the model given in Example 1. Plugging Eqs. (17) and (1) into Eq. (16) gives,

$$\begin{aligned} MSEP(\hat{\theta}) = E\{[(\theta^{(0)} - \hat{\theta}^{(0)}) &+ (\theta^{(1)} - \hat{\theta}^{(1)}) \times x^{(1)} + (\theta^{(2)} - \hat{\theta}^{(2)}) \times x^{(2)} \\ &+ (\theta^{(3)} - \hat{\theta}^{(3)}) \times x^{(3)} + (\theta^{(4)} - \hat{\theta}^{(4)}) \times x^{(4)} \\ &+ (\theta^{(5)} - \hat{\theta}^{(5)}) \times x^{(5)} + \varepsilon] | \hat{\theta}\}^2 \end{aligned} \tag{18}$$

The expectation over the target distribution here involves an expectation over X, the explanatory variables in the model, and over ε, whose variability results from the variability in conditions not represented by the explanatory variables of the model. Taking the expectation gives

$$\begin{aligned} MSEP(\hat{\theta}) &= (\theta^{(0)} - \hat{\theta}^{(0)})^2 + (\theta^{(1)} - \hat{\theta}^{(1)})^2 1 + (\theta^{(2)} - \hat{\theta}^{(2)})^2 1 + (\theta^{(3)} - \hat{\theta}^{(3)})^2 1 \\ &\quad + (\theta^{(4)} - \hat{\theta}^{(4)})^2 1 + (\theta^{(5)} - \hat{\theta}^{(5)})^2 1 + \mathrm{var}(\varepsilon) \\ &= 0.90 \end{aligned}$$

We have used the fact that all the random variables are independent and have expectation 0 and that the components of X have variance 1. Thus $E(x^{(i)}x^{(j)}) = E(x^{(i)})E(x^{(j)}) = 0$ for all $i \neq j$, $E(x^{(i)}\varepsilon) = E(x^{(i)})E(\varepsilon) = 0$ for all i and $E(x^{(i)}x^{(i)}) = 1$ for all i.

We can illustrate the importance of the target distribution using this example. To do so, we evaluate $MSEP(\hat{\theta})$ for a second target population. The distribution of X for this new target population is the same as for the first, except that now the variance of $x^{(5)}$ is 4 instead of 1. For example, if $x^{(5)}$ represents (soil depth in cm – 100 cm)/10, then the soil depth for both target populations is centered at 100 cm, but in the original distribution 95% of the soils have depths between 80 and 120 cm (expected value ± 2 standard deviations), whereas in the new distribution soil depth is more variable, 95% of the soils having depths between 60 and 140 cm. For this new target distribution, $MSEP(\hat{\theta}) = 2.2$ compared to 0.90 for the original target distribution. The origin of the difference is the term $(\theta^{(5)} - \hat{\theta}^{(5)})^2 E(x^{(5)})^2$, which is now equal to $(0.002 - 0.7)^2 \times 4$ instead of $(0.002 - 0.7)^2 \times 1$. That is, in changing the target population we have changed the way model errors (in this case the error in estimating $\theta^{(5)}$) contribute to the error of prediction.

3.3. $MSEP(\hat{\theta})$ and the choice of model complexity

We can develop the mean squared error of prediction as

$$\begin{aligned} MSEP(\hat{\theta}) &= E\big\{[Y - E(Y|X) + E(Y|X) - f(X;\theta)]^2\big\} \\ &= E_X\big\{E_Y\{[Y - E_Y(Y|X) + E_Y(Y|X) - f(X;\hat{\theta})]|X\}^2\big\} \\ &= E_X\big\{E_Y\{[Y - E_Y(Y|X)]|X\}^2\big\} + E_X\{[E_Y(Y|X) - f(X;\hat{\theta})]|X\}^2\big\} \\ &= \Lambda + \Delta \end{aligned} \tag{19}$$

where

$$\Lambda = E_X\{E_Y\{[Y - E_Y(Y|X)]^2|X\}\} = E_X\,[\mathrm{var}(Y|X)] = \text{population variance} \tag{20}$$

$$\Delta = E_X\{[E_Y(Y|X) - f(X;\hat{\theta})]^2\} = \textit{squared bias} \tag{21}$$

(Bunke and Droge, 1984; Wallach and Goffinet, 1987).

The second line of Eq. (19) follows because one can take an expectation over X and Y by first fixing X and taking the expectation over Y, and then taking the expectation over X. (In the notation here, we indicate specifically which variables are concerned by the expectation. Thus, E_Y is an expectation over Y.) In the third line we develop the square. The cross term is null because it involves $E_X\{E_Y\{[Y - E_Y(Y|X)]|X\}\} = E_X\{E_Y(Y|X) - E_Y(Y|X)\} = 0$.

The two components of $MSEP(\hat{\theta})$ are noted Λ (lambda) and Δ (delta). The population variance, Λ, depends on how much Y varies for fixed values of the explanatory variables in the model. When X is fixed Y still varies, within the target population, because not all the variables that affect Y are included in the model. That variability is then averaged over X. Note that Λ does not involve $f(X;\hat{\theta})$, i.e. the exact equations of the model are irrelevant here. It is only the choice of the explanatory variables that is important. If the explanatory variables in the model do not explain most of the variability in Y, then the remaining variability in Y for fixed X is large and Λ is large. Consider for example, a model which does not include initial soil mineral nitrogen. If Y (for example yield) for the target population is strongly affected by initial soil nitrogen, then Λ will be large. We see that the choice of explanatory variables is a major decision as far as prediction accuracy is concerned. That choice sets a minimum value for mean squared prediction error. Even if the model is the best possible, the mean squared error of prediction cannot be less than the population variance Λ.

The squared bias term, Δ, does depend on the form of the model. Once the choice of explanatory variables in the model is made, then the best model (minimum value of $MSEP(\hat{\theta})$) is the model that predicts a value equal to $E_Y(Y|X)$ at each value of X. The bias measures the distance between this best prediction and the model prediction, averaged over the target distribution of X values. The bias may be due to errors in the form of the model or to errors in the parameter values. Figure 3 illustrates the two contributions to the mean squared error of prediction.

The above decomposition of $MSEP(\hat{\theta})$ into two terms can help to understand the consequences of choosing different levels of detail for a model. Adding more detail in

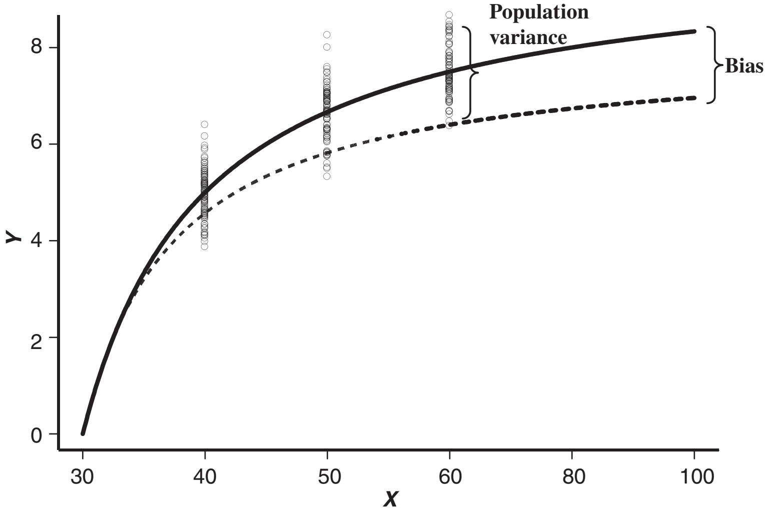


Figure 3. Response Y as a function of a single explanatory variable x. For 3 specific x values, the variability of Y is shown. The solid line is $E(Y|x)$ and the dashed line a hypothetical model.

general involves including additional explanatory variables. This has two opposing consequences. On the one hand, adding additional explanatory variables will reduce (or at worst leave unchanged) the unexplained variability in Y once X is fixed, i.e. Λ will decrease or at worst remain unchanged. On the other hand, there will in general be additional equations and parameters to estimate in conjunction with the additional explanatory variables. This will in general lead to an increase in the squared bias term Δ.

Suppose that one has a preliminary model and wants to decide whether or not to add additional explanatory variables. There is a better chance that the additional explanatory variables will reduce $MSEP(\hat{\theta})$ if

(1) they play an important role in determining the variability in Y in the target population, so that adding them to the model reduces Λ by a substantial amount.
(2) the associated equations and parameters can be well estimated from the available data, so that the additional detail does not cause a substantial increase in Δ.

In Example 2, the true relation of Y to X is given. We can then calculate the two contributions to $MSEP(\hat{\theta})$. The variance of Y for fixed X is $\mathrm{var}(\varepsilon)$, and this is the same for all X. Thus $\Lambda = E_X\,[\mathrm{var}(\varepsilon)] = \mathrm{var}(\varepsilon) = 0.04$. The squared bias term is

$$\begin{aligned} MSEP(\hat{\theta}) &= (\theta^{(0)} - \hat{\theta}^{(0)})^2 + (\theta^{(1)} - \hat{\theta}^{(1)})^2 1 + (\theta^{(2)} - \hat{\theta}^{(2)})^2 1 + (\theta^{(3)} - \hat{\theta}^{(3)})^2 1 \\ &\quad + (\theta^{(4)} - \hat{\theta}^{(4)}) 1 + (\theta^{(5)} - \hat{\theta}^{(5)})^2 1 \\ &= 0.86 \end{aligned}$$

In this example, almost all the error arises from the squared bias term Δ. The variability in Y that is not explained by the explanatory variables makes only a very small contribution to prediction error. Adding additional explanatory variables in this case could at best only reduce Λ by 0.04 units, which would only very marginally improve model predictions.

In general, bias can arise from errors in the model equations and/or from errors in the parameter values. In Example 2, the form of the model is correct (in the model Y is a linear function of X as in the true relation), so the bias arises solely from the error in the parameter vector.

Example 3

The purpose of this example is to show how Λ, Δ and their sum $MSEP(\hat{\theta})$ evolve as additional explanatory variables are added to a model. We assume that the true relation between Y and X is given by Eq. (17). We consider the following sequence of models:

$$f_1(X;\hat{\theta}) = \hat{\theta}^{(0)} + \hat{\theta}^{(1)}x^{(1)}$$

$$f_2(X;\hat{\theta}) = \hat{\theta}^{(0)} + \hat{\theta}^{(1)}x^{(1)} + \hat{\theta}^{(2)}x^{(2)}$$

$$f_3(X;\hat{\theta}) = \hat{\theta}^{(0)} + \hat{\theta}^{(1)}x^{(1)} + \hat{\theta}^{(2)}x^{(2)} + \hat{\theta}^{(3)}x^{(3)}$$

$$f_4(X;\hat{\theta}) = \hat{\theta}^{(0)} + \hat{\theta}^{(1)}x^{(1)} + \hat{\theta}^{(2)}x^{(2)} + \hat{\theta}^{(3)}x^{(3)} + \hat{\theta}^{(4)}x^{(4)}$$

$$f_5(X;\hat{\theta}) = \hat{\theta}^{(0)} + \hat{\theta}^{(1)}x^{(1)} + \hat{\theta}^{(2)}x^{(2)} + \hat{\theta}^{(3)}x^{(3)} + \hat{\theta}^{(4)}x^{(4)} + \hat{\theta}^{(5)}x^{(5)}$$

The parameter values for each model, estimated using ordinary least squares and the data in Table 1, are shown in Table 3.

The values of Λ, Δ and $MSEP(\hat{\theta})$ for each model are also given in Table 3. The calculations are easily done for this artificial example. We illustrate for the model $f_1(X;\hat{\theta})$. Here $X = (x^{(1)}, x^{(1)})^{\mathrm{T}}$. We write

$$f_1(X;\hat{\theta}) = \hat{\theta}^{(0)} + \hat{\theta}^{(1)}x^{(1)} + \varepsilon_1$$

Comparing with Eq. (17) shows that

$$\varepsilon_1 = \theta^{(2)}x^{(2)} + \theta^{(3)}x^{(3)} + \theta^{(4)}x^{(4)} + \theta^{(5)}x^{(5)} + \varepsilon.$$

It is then easily seen that $\varepsilon_1 \sim N\left[0, \sigma_1^2\right]$ with

$$\sigma_1^2 = \theta^{(2)2}\mathrm{var}(x^{(2)}) + \theta^{(3)2}\mathrm{var}(x^{(3)}) + \theta^{(4)2}\mathrm{var}(x^{(4)}) + \theta^{(5)2}\mathrm{var}(x^{(5)}) + \mathrm{var}(\varepsilon) = 4.04.$$

The population variance term is then $\Lambda = E_X\ \mathrm{var}(Y|X) = \mathrm{var}(\varepsilon_1) = 4.04$. The squared bias term is

$$\Delta = E_X\{E_Y(Y|X) - f(X;\hat{\theta})^2\} = E_X\{[\theta^{(0)} + \theta^{(1)}x^{(1)} - (\hat{\theta}^{(0)} + \hat{\theta}^{(1)}x^{(1)})]^2\}$$

$$= (\theta^{(0)} - \hat{\theta}^{(0)})^2 + (\theta^{(1)} - \hat{\theta}^{(1)})^2 E(x^{(1)})^2 = 0.36$$

Finally, $MSEP(\hat{\theta}) = \Lambda + \Delta = 4.04 + 0.36 = 4.40$. Analogous calculations apply to the other models.

Table 3 shows how Λ, Δ and $MSEP(\hat{\theta})$ vary as we add additional explanatory variables to the model. As more explanatory variables are added to the model, the amount of unexplained variability (represented by the term Λ) must decrease or at worst remain constant. This is indeed the behavior of Λ in Table 3. The decrease in Λ with each new explanatory variable depends on the importance of that variable in explaining the variability in Y. Adding $x^{(2)}$ to the model decreases Λ substantially, but further explanatory variables have little importance. The term Δ has a more complex behavior, decreasing at first then increasing. The result for $MSEP(\hat{\theta})$ is that it decreases to a minimum for the model $f_3(X; \hat{\theta})$, then increases as further explanatory variables are added. This is typical behavior, for complex models like a crop model as well as for simple linear models like our example. Eventually, as more explanatory variables are added to a model, the accumulation of errors in the model equations and in the parameter estimates outweighs decreases in Λ and so $MSEP(\hat{\theta})$ begins to increase. In the example here, $f_3(X; \hat{\theta})$ is the best model for prediction.

Table 3. A sequence of increasingly complex models adjusted to the data in Table 1.

Model	Parameters in the model Least squares parameter values	Λ	Δ	$MSEP(\hat{\theta})$	MSE
$f_1(X;\theta)$	$\theta^{(0)}, \theta^{(1)}$ 2.535, 8.275	4.04	0.36	4.40	4.61
$f_2(X;\theta)$	$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}$ 2.121, 8.005, 2.065	0.04	0.02	0.06	0.01
$f_3(X;\theta)$	$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}$ 2.046, 7.971, 2.085, 0.091	0.04	0.01	0.05	0.01
$f_4(X;\theta)$	$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}$ 1.906, 7.906, 2.036, 0.169, 0.156	0.04	0.05	0.09	0.004
$f_5(X;\theta)$	$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}, \theta^{(5)}$ 1.641, 7.735, 1.967, 0.237, 0.230, −0.174	0.04	0.35	0.39	0.0003

3.4. Estimating $MSEP(\hat{\theta})$

How then does one estimate the value of the mean squared error of prediction in real-life situations, where the true relation between Y and X is unknown?

This is the topic of the following sections. The approach is quite different depending on whether the data available for estimating $MSEP(\hat{\theta})$ have been used to guide model development (in particular for parameter estimation) or not.

3.4.1. Model and data are independent

The simplest situation arises when we have a random sample of data from the target population, and the model has been developed independently of that data. Then an unbiased estimator of $MSEP(\hat{\theta})$ is simply

$$\hat{M}SEP(\hat{\theta}) = MSE = \frac{1}{N}\sum_{i=1}^{N}[Y_i - f(X_i;\hat{\theta})]^2 \tag{22}$$

where N is the number of observations. Each squared error in the sum is an unbiased estimator of $MSEP(\hat{\theta})$ and the best overall estimator is simply the average of the squared errors. It is also important to have an idea of how much this estimator would vary if a different data set had been chosen from the target distribution. Since the estimator is a mean of independent terms, the estimated variance of the estimator is

$$\hat{v}ar[\hat{M}SEP(\hat{\theta})] = \frac{1}{N(N-1)}\sum_{i=1}^{N}\{[Y_i - f(X_i;\hat{\theta})]^2 - \hat{M}SEP(\hat{\theta})\}^2 \tag{23}$$

We have already discussed the difficulty in practice of obtaining a random sample from the target distribution. We may however be able to obtain a hierarchical random sample, where the first level involves random sampling from the target distribution but then soils or climates are repeated. In this case Eq. (22) is still an unbiased estimator of $MSEP(\hat{\theta})$, but Eq. (23) can no longer be used to estimate the variance of this estimator.

For the model of Eq. (1), using the data in Table 1, we calculated $MSEP(\hat{\theta}) = 0.90$ and $MSE = 0.88 = \hat{M}SEP(\hat{\theta})$. The estimated and true values of $MSEP(\hat{\theta})$ are very close, but this is somewhat of a coincidence since the estimated standard deviation of $\hat{M}SEP(\hat{\theta})$, calculated using Eq. (23), is relatively large (0.67).

3.4.2. *Data are used for model development*

A very common situation is that where we have a single data set, and want to use it both to estimate certain parameter values for the model and to estimate $MSEP(\hat{\theta})$. In this case *MSE* in general underestimates $MSEP(\hat{\theta})$. The reason is easy to understand. First, one specifically fits the model to the data, then one calculates how well the model fits that same data. Clearly, that fit will in general be better than the fit of the model to other situations chosen at random from the target distribution.

When many parameters are estimated relative to the amount of data, the difference between *MSE* and $MSEP(\hat{\theta})$ can be very important. Furthermore, *MSE* and $MSEP(\hat{\theta})$ will have qualitatively different behavior as model complexity increases. *MSE* can never increase as additional explanatory variables are added to a model, assuming that the associated parameters are adjusted to the data. Thus model choice based on *MSE* will always lead to choosing the most complex model. $MSEP(\hat{\theta})$ on the other hand leads to choosing a model with some intermediate level of complexity.

Table 3 shows that *MSE* and $MSEP(\hat{\theta})$ are comparable for the model with just 2 parameters adjusted to the data, but for more complex models *MSE* seriously underestimates $MSEP(\hat{\theta})$. For the model $f_4(X;\hat{\theta})$ the ratio $MSEP(\hat{\theta})/MSE$ is 25 and for model $f_5(X;\hat{\theta})$ it is 1400! If the criterion for choosing a model were minimum *MSE*, one would choose $f_5(X;\hat{\theta})$, the most complex model. The model with the smallest mean squared error of prediction on the other hand is $f_3(X;\hat{\theta})$.

3.4.3. Cross-validation

How then can one estimate $MSEP(\hat{\theta})$ when one also needs to use the data for parameter estimation? For the moment, we continue to assume that the data are a random sample from the target distribution. A simple solution is to split the data into two parts, say half the data in data set 1 and the other half in data set 2. Then data set 1 is used to estimate the parameters and data set 2 is used to estimate $MSEP(\hat{\theta})$. Since data set 2 was not used for parameter estimation, *MSE* calculated using data set 2 is an unbiased estimator of $MSEP(\hat{\theta})$. However, this approach has major drawbacks. First, we now use only half of our data for estimating the parameter values and also for estimating $MSEP(\hat{\theta})$. The estimates will therefore be less precise. A second drawback is the arbitrariness in this procedure. Why split into two equal halves rather than using some other proportion? On what basis are observations put into one data set rather than the other?

The method of cross-validation is based on this same principle of data splitting, but avoids to a large degree the above drawbacks. The cross-validation estimator of $MSEP(\hat{\theta})$ is

$$\hat{M}SEP_{\mathrm{CV}}(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}[Y_i - f(X_i;\hat{\theta}_{-i})]^2 \tag{24}$$

The notation $\hat{\theta}_{-i}$ indicates that the parameter values are estimated using all the data in the data set except Y_i.

Concretely, one begins by estimating the parameter values using all the data except Y_1. The result is the estimated parameter vector $\hat{\theta}_{-1}$. Since Y_1 was not used to estimate $\hat{\theta}_{-1}$, the squared error $[Y_1 - f(X_1;\hat{\theta}_{-1})]^2$ is an unbiased estimator of $MSEP(\hat{\theta})$. This gives the first term in the sum in Eq. (24). Then the procedure is repeated, this time basing parameter estimation on all the data except Y_2 and calculating $[Y_2 - f(X_2;\hat{\theta}_{-2})]^2$. The calculations continue, adjusting the parameters to all the data except Y_3, then except Y_4, and so on. Each adjusted parameter vector gives rise to a term in the sum of Eq. (24). The final estimator of $MSEP(\hat{\theta})$ is the average of the N unbiased estimators $[Y_i - f(X_i;\hat{\theta}_{-i})]^2$ for $i = 1, \ldots, N$.

At the end of the procedure, we have N different estimates of the parameter vector $\hat{\theta}$. Which should we use? None, in fact. The best estimator is the one based on all the data, and that is the estimator to use in practice.

In this approach, all of the data are used to estimate the parameters as well as to estimate $MSEP(\hat{\theta})$. Furthermore, all the data are used in the same way, which eliminates the problem of arbitrarily assigning a data value to some particular subset. A disadvantage of the method is the calculation time. It is now necessary to estimate the model parameters not once but $N+1$ times. Also, we are in fact estimating $MSEP(\hat{\theta}_{-1})$, $MSEP(\hat{\theta}_{-2})$, etc. So there is an additional assumption that those quantities are good estimators of $MSEP(\hat{\theta})$. Nevertheless, this estimator of prediction error is largely used in the statistical literature (Harrell, 2001) and has been used for crop models by Jones and Carberry (1994), Colson et al. (1995) and others.

Table 4 illustrates the evaluation of $\hat{M}SEP_{\mathrm{CV}}(\hat{\theta})$ (Eq. (24)) for the model $f_5(X;\hat{\theta})$ of Example 3 using the data of Table 1. Each line in Table 4 corresponds to estimating the parameters from a sample missing a different data point. The last column shows the squared error for the data point not used for model estimation. The last line of the table gives the average of these squared errors, which is the cross-validation estimate $\hat{M}SEP_{\mathrm{CV}}(\hat{\theta}) = 0.313$. The true value is $MSEP(\hat{\theta}) = 0.390$ (Table 3).

Table 4. Calculation of cross-validation estimate of $MSEP(\hat{\theta})$ for model$f_5(X;\hat{\theta})$.

i	Data used for parameter adjustment	Y_i	$f_5(X_i;\hat{\theta}_{-i})$	$[Y_i - f_5(X_i;\hat{\theta}_{-i})]^2$
1	$Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, Y_8$	−9.39	9.44	1.06
2	$Y_1, Y_3, Y_4, Y_5, Y_6, Y_7, Y_8$	−3.23	11.96	0.97
3	$Y_1, Y_2, Y_4, Y_5, Y_6, Y_7, Y_8$	0.37	−2.45	0.12
4	$Y_1, Y_2, Y_3, Y_5, Y_6, Y_7, Y_8$	7.10	1.01	0.00
5	$Y_1, Y_2, Y_3, Y_4, Y_6, Y_7, Y_8$	5.82	−8.33	0.03
6	$Y_1, Y_2, Y_3, Y_4, Y_5, Y_7, Y_8$	−11.21	9.14	0.15
7	$Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_8$	−5.81	−8.48	0.06
8	$Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7$	2.82	−4.04	0.10
	$\hat{M}SEP_{\mathrm{CV}}(\hat{\theta})$			0.31

Parameter estimation for crop models is sometimes based on a trial and error procedure. Such an approach has many disadvantages, but a further disadvantage is that cross-validation becomes essentially impossible. Cross-validation requires that parameter estimation be repeated several times. The implicit assumption is that the method of estimation does not vary, only the data vary. For this to be true, one needs a reproducible algorithm for parameter estimation.

The above approach is referred to as leave-one-out cross-validation, because a single data point is left out of the sample at each step. This may not be appropriate for sampling schemes other than random sampling. If for example there are several data points from each site, it would be necessary to leave out all the data points from each site in turn. An example of cross-validation for a crop model leaving out more than one data point at a time is given in Wallach et al. (2001).

3.4.4. Bootstrap estimation

The bootstrap, like cross-validation, is a data re-sampling approach, i.e. the same data are used several times to provide an estimator of the quantity of interest, here the mean squared error of prediction.

There are many variants of the bootstrap approach. Here we present just one relatively simple version (Efron, 1983). We do not directly estimate $MSEP(\hat{\theta})$ but rather the difference

$$op = MSEP(\hat{\theta}) - MSE.$$

The notation "*op*" comes from "optimistic" and was chosen to emphasize the fact that when the parameters are estimated from the data, MSE is in general smaller (more optimistic about the quality of the model) than the mean squared error of prediction. The true mean squared error of prediction will in general be larger. Thus the idea here is to calculate *MSE* and then to augment it by an estimator of *op*, in order to obtain an estimator of $MSEP(\hat{\theta})$.

The bootstrap approach is equivalent to supposing that the original data set, with N observations, constitutes the full target distribution. We will refer to this as the bootstrap target population. Thus the expectation over the bootstrap target population is equivalent to an average over the N observations. We create B bootstrap samples, each with N elements, by drawing N points with replacement from the bootstrap target population. Since we are sampling with replacement, a bootstrap sample may have some of the data points represented several times, while other data points are absent. Using the data in Table 1, 3 bootstrap samples could be (Y_6, Y_7, Y_3, Y_7, Y_6, Y_7, Y_6, Y_8), (Y_3, Y_1, Y_7, Y_8, Y_5, Y_4, Y_7, Y_2) and (Y_1, Y_7, Y_8, Y_7, Y_5, Y_2, Y_4, Y_7).

The parameters are adjusted to each bootstrap sample, giving estimated parameter vector $\hat{\theta}_b$ for the bth bootstrap sample. The mean squared error of prediction of the model based on bootstrap sample b for the bootstrap target population is

$$MSEP_b(\hat{\theta}_b) = \frac{1}{N}\sum_{i=1}^{N}[Y_i - f(X_i;\hat{\theta}_b)]^2$$

We can also calculate *MSE* for sample b as

$$MSE_b = \frac{1}{N}\sum_{i=1}^{N}[Y_{bi} - f(X_{bi};\hat{\theta}_b)]^2$$

where Y_{bi} and X_{bi} refer to the ith data point of bootstrap sample b. The value of op for sample b is then $op_b = MSEP_b(\hat{\theta}_b) - MSE_b$ and the final bootstrap estimator of op is

$$\hat{op} = \frac{1}{B}\sum_{b=1}^{B} op_b$$

Finally, the bootstrap estimator of $MSEP(\hat{\theta})$ is

$$MS\hat{E}P_{\text{bootstrap}}(\hat{\theta}) = MSE + \hat{o}p$$

where MSE is calculated using the original sample.

It is usually recommended to have several tens or hundreds of bootstrap samples. Since the parameter vector must be adjusted to each sample, the overall calculation time can be quite long. Furthermore, there can be numerical difficulties in adjusting the parameters for certain bootstrap samples, especially if the original sample is quite small so that it is likely to have some bootstrap samples with only a few distinct data points. On the positive side, in a simulation study Efron (1983) found that this bootstrap method gave better predictions of $MSEP(\hat{\theta})$ than did cross-validation.

Wallach and Goffinet (1989) use the above bootstrap approach to estimate the difference in $MSEP(\hat{\theta})$ between two models. Wallach and Goffinet (1987) used a variant of the above bootstrap approach, adapted to the specific case where the data have a hierarchical structure. Their study concerned a static model for predicting the maintenance requirements of sheep.

3.5. *Effect of errors in Y or in X on MSEP($\hat{\theta}$)*

3.5.1. *Measurement error in Y and MSEP($\hat{\theta}$)*

In the above discussion, we have assumed that Y is measured without error. If that is not the case, there are two different mean squared errors of prediction that are of interest. The first, noted $MSEP^{\text{obs}}(\hat{\theta})$, refers to the difference between calculated values and observed values. The second, $MSEP(\hat{\theta})$, refers to the difference between calculated values and the true response values. We show here how these two quantities are related.

We suppose that

$$Y^{\text{obs}} = Y + \eta$$

where η, the measurement error, is a random variable independent of X and Y with $E(\eta) = 0$ and $\text{var}(\eta) = \sigma_\eta^2$. Then (Wallach and Goffinet, 1987)

$$\begin{aligned} MSEP^{\text{obs}}(\hat{\theta}) &= E\{[Y^{\text{obs}} - f(X;\hat{\theta})]^2\} \\ &= E\{[Y^{\text{obs}} - Y + Y - f(X;\hat{\theta})]^2\} \qquad (25) \\ &= \sigma_\eta^2 + MSEP(\hat{\theta}) \end{aligned}$$

If we estimate the mean squared error of prediction using observed Y values with measurement error, it means we are estimating $MSEP^{\text{obs}}(\hat{\theta})$. We can obtain an estimate of $MSEP(\hat{\theta})$ by subtracting an estimate of the measurement error, i.e.

$$M\hat{S}EP(\hat{\theta}) = M\hat{S}EP^{\text{obs}}(\hat{\theta}) - \hat{\sigma}_{\eta}^{2}$$

3.5.2. *Measurement error in X and MSEP($\hat{\theta}$)*

We have so far assumed that the explanatory variables are measured without error. However, this is often not the case. For example, soil characteristics such as moisture content at field capacity and at wilting point are difficult to determine and may have appreciable errors. If there is no weather station in the field in question, there may be errors in the weather data. Initial soil moisture and initial nitrogen may also have errors, in particular if they are estimated rather than measured.

Let U be the subset of explanatory variables that has measurement error. Suppose $U^{\text{obs}} = U + \tau$ where U^{obs} is the observed value, U is the true value and τ is the error. We suppose that $E(\tau) = 0$ and we note $\text{var}(\tau) = \Sigma_{\tau}$. If U (and therefore τ) is of dimension n, then Σ_{τ} is an $n \times n$ matrix. Wallach and Génard (1998) showed that, approximately, the effect of measurement errors in the explanatory variables is to increase the mean squared error of prediction by adding on the term

$$\Gamma_U = E\left\{\left[\frac{\partial f(X;\hat{\theta})}{\partial U}\right]^{\mathrm{T}} \Sigma_{\tau} \left[\left[\frac{\partial f(X;\hat{\theta})}{\partial U}\right]\right]\right\} \tag{26}$$

The expectation is over the target distribution. The partial derivative with respect to the vector U is a column vector.

While Eq. (26) may look forbidding, it is actually quite easy to understand and to evaluate. The partial derivatives measure how the model output changes when each explanatory variable in U changes. It is logical that the importance of error in an explanatory variable depends on the sensitivity of the output to that of explanatory variable. The other factor in Γ_U is the variance–covariance matrix Σ_{τ}. Consider for simplicity the case where this matrix is diagonal. The diagonal terms are just the variances of the elements of τ. Then the larger the errors, the larger the variances and the larger the effect on Γ_U.

The expression for Γ_U involves the model, the errors in the explanatory variables and the target distribution, but does not require any measured outputs. This term can then be estimated in the absence of data on the response variable. For example, one could calculate how important certain errors would be before actually doing measurements.

We illustrate the calculation of Γ_U using the model of Example 1. Suppose that just the two components $x^{(1)}$ and $x^{(2)}$ are measured with error. Specifically, suppose that

$$U^{\text{obs}} = \begin{pmatrix} x^{(1)\text{obs}} \\ x^{(2)\text{obs}} \end{pmatrix} = \begin{pmatrix} x^{(1)} + \tau_1 \\ x^{(2)} + \tau_2 \end{pmatrix}, \quad E\begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\text{var}\begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.02 \end{pmatrix}$$

The fact that Σ_τ is diagonal implies that the errors in the explanatory variables $x^{(1)}$ and $x^{(2)}$ are independent. There is no assumption about whether the explanatory variables themselves are independent or not.

The partial derivative vector will normally be calculated numerically. For our simple example however it can be calculated analytically as,

$$\frac{\partial f(X;\hat\theta)}{\partial U}=\begin{pmatrix}\dfrac{\partial(\hat\theta^{(0)}+\hat\theta^{(1)}x_i^{(1)}+\hat\theta^{(2)}x_i^{(2)}+\hat\theta^{(3)}x_i^{(3)}+\hat\theta^{(4)}x_i^{(4)}+\hat\theta^{(5)}x_i^{(5)})}{\partial x_i^{(1)}}\\ \dfrac{\partial(\hat\theta^{(0)}+\hat\theta^{(1)}x_i^{(1)}+\hat\theta^{(2)}x_i^{(2)}+\hat\theta^{(3)}x_i^{(3)}+\hat\theta^{(4)}x_i^{(4)}+\hat\theta^{(5)}x_i^{(5)})}{\partial x_i^{(2)}}\end{pmatrix}=\begin{pmatrix}\hat\theta^{(1)}\\ \hat\theta^{(2)}\end{pmatrix}$$

The partial derivatives do not depend on X, so we can simply ignore the expectation over X in Γ_U.

Substituting into Eq. (26) gives

$$\Gamma_U=\begin{pmatrix}\hat\theta^{(1)} & \hat\theta^{(2)}\end{pmatrix}\begin{pmatrix}0.05 & 0\\ 0 & 0.02\end{pmatrix}\begin{pmatrix}\hat\theta^{(1)}\\ \hat\theta^{(2)}\end{pmatrix}=3.2$$

In the absence of measurement error in X we had $MSEP(\hat\theta)=0.90$. With measurement error $MSEP(\hat\theta)=0.90+3.2=4.1$. In this particular example, measurement error in the explanatory variables is the major contribution to the mean squared error of prediction.

3.6. Parameter uncertainty and mean squared error of prediction

In general, there is a degree of uncertainty about the parameter values in a model. For crop models, this may include uncertainties in the parameters adjusted to data using regression techniques and also uncertainties in parameter values taken from the literature. In the previous sections, we were not required to consider this uncertainty because we treated the model with fixed parameter values. The fact that parameter estimation could have given different parameter values was not relevant.

It can also be of interest, however, to consider $E[MSEP(\hat\theta)]$, where the expectation is over the distribution of $\hat\theta$. We first derive an expression for this expectation, then discuss its usefulness. The treatment is similar but not identical to that presented in Wallach and Goffinet (1987) and in Wallach and Génard (1998). We have

$$\begin{aligned}E[MSEP(\hat\theta)]&=E_{\hat\theta}E_X\{E_Y\{[Y-f(X;\hat\theta)]|X\}^2\}\\&=E_{\hat\theta}E_X\{E_Y\{[Y-E_{\hat\theta}[f(X;\hat\theta)]+E_{\hat\theta}[f(X;\hat\theta)]-f(X;\hat\theta)]|X\}^2\}\\&=E_X\{E_Y\{[Y-E_{\hat\theta}[f(X;\hat\theta)]]|X\}^2+E_{\hat\theta}E_X\{E_{\hat\theta}[f(X;\hat\theta)]\\&\qquad-f(X;\hat\theta)]|X\}^2\}\\&=\Lambda+\Delta_{E_{\hat\theta}[f(X;\hat\theta)]}+E_X\{\text{var}[f(X;\hat\theta)|X\}]\}\end{aligned}\qquad(27)$$

This can be compared to the decomposition of $MSEP(\hat{\theta})$ in Eq. (19). The first term is the same population variance term as in Eq. (20). The second term is the squared model bias of Eq. (21), but now for the model averaged over the distribution of $\hat{\theta}$. The last term is new. It is the variance of the model due to the variability of the parameter values, averaged over the target distribution of explanatory variables. If there is no uncertainty in the parameter values, then $\mathrm{var}[f(X;\hat{\theta})|X] = 0$. Since the variance cannot be negative, any uncertainty necessarily increases $E[MSEP(\hat{\theta})]$.

One might argue that in fact we are only interested in a single value of $\hat{\theta}$, the one that we intend to use, and so the expectation over the distribution of $\hat{\theta}$ does not concern us. However, the uncertainty in $\hat{\theta}$ reflects the fact that we do not know the parameter values exactly, and it is just a matter of chance that we have obtained a particular value and not another. For example, adding new data will change $\hat{\theta}$. To have confidence in the model, we would not want $E[MSEP(\hat{\theta})]$ to be large, even if the estimated value of $MSEP(\hat{\theta})$ for our particular parameters is acceptable.

To estimate $E[MSEP(\hat{\theta})]$ one could use

$$\hat{E}[MSEP(\hat{\theta})] = M\hat{S}EP(\hat{\theta}) + \hat{E}_X\{\mathrm{var}[f(X;\hat{\theta})|X]\}$$

The last term on the right is an estimator of $E_X\{\mathrm{var}[f(X;\hat{\theta})|X]\}$. Note that it involves the model, the distribution of the parameter vector and the target distribution, but not measured response variables. This term can then be estimated for scenarios that have not been observed or for which there is very little data. For example, one might want to use the model to predict crop performance in a new environment, or under changed climatic conditions, or with modified management practices. All these cases correspond to using the model for new target distributions. Extrapolating a model to conditions outside those where it has been adjusted and tested is always a perilous exercise, and any information about the validity of the extrapolation is important. Without data we cannot estimate $MSEP(\hat{\theta})$ but we can estimate the contribution of the term $E_X\{\mathrm{var}[f(X;\hat{\theta})|X\}]\}$ to $E[MSEP(\hat{\theta})]$. If this contribution is small we have at least eliminated from consideration one often major source of error, though population variance and expected squared bias are still unknown. If the contribution is large, then we are forewarned that the model will be a poor predictor for the new target distribution. We can also be assured that reducing the variance of the parameter estimators will be worthwhile.

Example 4

We will calculate here $E[MSEP(\hat{\theta})]$ and its components (Eq. (27)) for the models of Example 3. We present the explicit calculations only for the model $f_1(X;\hat{\theta})$. The other models can be treated analogously.

The first term in Eq. (27), the population variance, is the same as in Table 3. For $f_1(X;\hat{\theta})$, $\Lambda = \mathrm{var}(\varepsilon_1) = 4.04$. The second term is

$$E_X\{\{[E(Y|X)] - E_{\hat{\theta}}[f_1(X;\hat{\theta})]\}^2\} = E_X\{[\theta^{(0)} + \theta^{(1)}x^{(1)} - E_{\hat{\theta}}(\theta^{(0)} + \theta^{(1)}x^{(1)})]\}^2 = 0$$

We have used the fact that the least square parameter estimators are unbiased estimators of the true parameters, so that $\mathrm{E}_{\hat{\theta}}(\theta^{(0)}) = \theta^{(0)}$ and $\mathrm{E}_{\hat{\theta}}(\theta^{(1)}) = \theta^{(1)}$. The last term of Eq. (27) is

$$\begin{aligned} E_X\{\mathrm{var}[f(X;\hat{\theta})|X]\} &= E_X\{E_{\hat{\theta}}[\hat{\theta}^{(0)} + \hat{\theta}^{(1)}x^{(1)} - (\theta^{(0)} + \theta^{(1)}x^{(1)})]^2\} \\ &= E_{\hat{\theta}}[(\hat{\theta}^{(0)} - \theta^{(0)})^2 + (\hat{\theta}^{(1)} - \theta^{(1)})^2] = \mathrm{var}(\hat{\theta}^{(0)}) + \mathrm{var}(\hat{\theta}^{(1)}) \end{aligned}$$

In going from the first line to the second we have used the fact that $E(x^{(1)2}) = 1$ and $E(x^{(1)}) = 0$. To evaluate the above quantity we use estimates of the variances provided by the least squares fitting program, accepting the fact that the estimates can be poor with so few data.

The calculated values of $E_{\hat{\theta}}[MSEP(\hat{\theta})]$ and the separate contributions are presented in Table 5. These values can be compared to the values for $MSEP(\hat{\theta})$ for the same models in Table 3. The contribution of population variance is the same in both cases, by definition. This term depends only on the choice of explanatory variables in the model and not on the form of the model. Thus it does not change when we consider an average over parameter values rather than the model with fixed parameter values. The differences between the two tables are in the other contributions to the mean squared error of prediction. In Table 3, the Δ term arises from the differences between the estimated and true parameter values. In Table 5, that term is replaced by an average over parameter estimates, and so its contribution is null because the parameter estimators are unbiased. However, there is now a contribution from the variance of the parameter estimators, i.e. in place of the errors in a specific set of parameter values we now have a term that represents an average error. The values of $E_{\hat{\theta}}[MSEP(\hat{\theta})]$ in Table 5 are not identical to the values of $MSEP(\hat{\theta})$ in Table 3. Nevertheless, in both cases a model of intermediate complexity (3 explanatory variables in the case of $MSEP(\hat{\theta})$, 2 in the case of $E_{\hat{\theta}}$ $[MSEP(\hat{\theta})]$ minimizes the mean squared error of prediction.

Table 5. Contributions to $E_{\hat{\theta}}[MSEP(\hat{\theta})]$ for the models of Example 3.

Adjusted parameters	Λ	$\Delta_{E_{\hat{\theta}}[f(X;\hat{\theta})]}$	$E_X\{\mathrm{var}[f(X;\hat{\theta})|X]\}$	$E_{\hat{\theta}}[MSEP(\hat{\theta})]$
$\theta^{(0)}, \theta^{(1)}$	4.04	0	2.56	6.60
$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}$	0.04	0	0.01	0.05
$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}$	0.04	0	0.03	0.07
$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}$	0.04	0	0.03	0.07
$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}, \theta^{(5)}$	0.04	0	0.02	0.06

3.6.1. A particular model, the average

An extremely simple model is what we will call the "average" model. This model uses the average of past Y measurements for all predictions. This model is of interest despite its simplicity because it can serve as a standard against which to measure other models. Any model which does not predict better than the simple average of past observations should be seriously questioned. We now show that it is straightforward to estimate $E_{\hat{\theta}}[MSEP(\hat{\theta})]$ for this model.

We write

$$Y = \mu + \varepsilon$$

with $E(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2$. Thus μ is the expectation of Y and ε is the random variability around the expectation. There are no assumptions involved here. One can always express a random variable as an expectation plus a random variability of expectation zero. The assumption that we do make is that we have a random sample, so that all the observations Y_i are independent and have the same distribution as Y. Our model is then

$$f(X; \hat{\theta}) = f(\hat{\theta}) = \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

This model has one parameter, $\hat{\mu}$, and no explanatory variables.

The population variance for this model is

$$\Lambda = E_Y\{[Y - E_Y(Y)]^2\} = E_Y\{[\mu + \varepsilon - \mu]^2\} = \sigma^2$$

The second term in the decomposition of $E_{\hat{\theta}}[MSEP(\hat{\theta})]$ in Eq. (27) is

$$E_X\{\{E_Y(Y) - E_{\hat{\theta}}[f(\hat{\theta})]\}^2\} = (\mu - \mu)^2 = 0$$

The final term in Eq. (27) is just the variance of the average of the observed Y values, and so

$$\mathrm{var}[f(\hat{\theta})] = \frac{\sigma^2}{N}$$

Overall then

$$E_{\hat{\theta}}[MSEP(\hat{\theta})] = \left(1 + \frac{1}{N}\right)\sigma^2 \tag{28}$$

This is an extreme case of a prediction model. It has no explanatory variables and so the population variance is maximal. A crop model should be able to do better, by introducing important explanatory variables that reduce population variance by more than the inevitable increase in error related to parameter estimation or bias.

An estimator of $E[MSEP(\hat{\theta})]$ is easily obtained, by replacing σ^2 in Eq. (28) by its usual estimator

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$$

where $\bar{Y}$ is the average of the Y_i values.

For the data in Table 1, the average response is 0.80 and so $f(\hat{\theta}) = 0.80$ for the "average" model. For this model, $E[MSEP(\hat{\theta})] = 71.0$ and the estimated value is $\hat{E}_{\hat{\theta}}[MSEP(\hat{\theta})] = \hat{\sigma}^2 = 70.5$. All of the models with explanatory variables do much better (Table 3).

4. Evaluating a model used for decision support

Prediction quality is often a major objective for a crop model, but it is not the only possible objective. Another common goal is to compare different management decisions. In this section, we discuss how a model could be evaluated with respect to this specific objective.

First, we illustrate that prediction quality and quality of model-based decisions can be quite different. Figure 4 shows, for one particular situation, predictions of profit (value of yield minus cost of nitrogen) *versus* applied nitrogen using two different models, labeled "A" and "B." We assume that the goal is to maximize profit. Model "A" predicts that profit is maximized at fertilizer level 200, which would therefore be the recommended dose according to this model. The true profit for this dose is 786, which is in fact the highest profit attainable. According to model "B" the optimal dose is 140, with corresponding real profit of only 730, i.e. basing the fertilizer decision on model "A" leads to substantially higher profit than using model "B." On the other hand, for every value of applied nitrogen, model "B" predicts profit better than does model "A." The criterion of predictive quality would lead us to choose the model which gives poorer management recommendations.

4.1. A criterion of decision quality

Let *d* be the vector of management variables to be optimized. This could be for example amount of nitrogen. It could also be a vector of several decisions, for example sowing

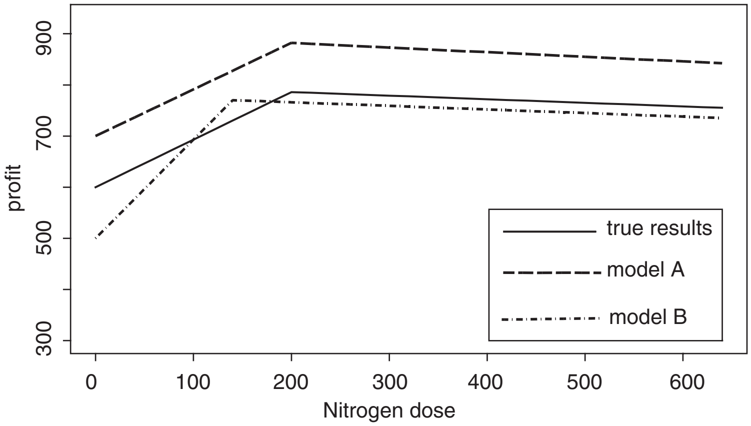


Figure 4. Profit *versus* amount of applied nitrogen.

date, sowing density and amount of fertilizer. Let *J(d)* be the objective function that specifies what "optimized" means exactly. *J(d)* could for example be profit, or it could be a combination of economic and environmental indicators.

Suppose that according to our model, the optimal decision for the situation with explanatory variable vector X is $d_{\mathrm{M}}(X)$. This is the value of d that maximizes *J(d)* according to the model. As indicated, the model-calculated optimal decision can depend on the situation through the explanatory variable X. (The calculation of optimal decisions is treated in Chapter 6.) Our criterion of model quality when the goal is management recommendations is then

$$C = E\left\{J\left[d_{\mathrm{M}}(X)\right]\right\} \tag{29}$$

The expectation in Eq. (29) is over the target distribution. The criterion C is the true expected value of the objective function that would be achieved, if the recommended decisions of the model were implemented everywhere in the target distribution.

The criterion C was implicitly used in discussing Figure 4. There the target distribution reduces to just a single situation. For model "A,"

$$C = E\left\{J\left[d_{\mathrm{A}}(X)\right]\right\} = J\,[200] = 786$$

For model "B"

$$C = E\left\{J\left[d_{\mathrm{B}}(X)\right]\right\} = J\,[140] = 730$$

The criterion C is thus larger for model "A," which would be the model of choice.

4.2. Estimating C

We suppose that the data available provide values of the objective function J for several values of the decision vector in each of several situations. For example, the data might be yield (from which we could calculate profit) for a series of nitrogen doses in each of several fields. We consider the case where the situations represented in the data are a random sample from the target population. We furthermore assume that the data have not been used for parameter development.

Suppose that the model-based decision $d_{\mathrm{M}}(X)$ is among the decisions tested for each situation in the data. Then a natural estimator of the criterion C is

$$\hat{C} = \frac{1}{N}\sum_{i=1}^{N} J_i\left[d_{\mathrm{M}}(X_i)\right] \tag{30}$$

where the index i refers to the situation and N is the number of situations in the data.

Often the observations available will not correspond to values of $d_{\mathrm{M}}(X)$. For example, the model recommended nitrogen dose for a particular situation might be 137 kg/ha, whereas the observations refer to doses of 0, 50, 100, 150 and 200 kg/ha N. Then one cannot directly use Eq. (30).

One possible approach would be to interpolate between observed values to obtain an estimate of $J_i[d_{\mathrm{M}}(X)]$ for each situation. Antoniadou and Wallach (2000) proposed a variant where one first combines all the data and then interpolates. This method was applied by Makowski et al. (2001) and by Makowski and Wallach (2001, 2002) to evaluate static models with respect to the quality of their recommendations for a single management variable, namely the total amount of nitrogen to be applied to wheat. According to this method, one first translates the observed doses for each situation. A dose d_i for situation i is translated to $d_i^{\mathrm{t}} = d_i - d_{\mathrm{M}}(X_{\mathrm{i}})$. For the model-based optimal dose $d_i = d_{\mathrm{M}}(X_i)$ and so the translated dose is $d_i^{\mathrm{t}} = 0$. On this new scale all situations have the same model-based optimal dose. The data provide values of $J(d_i^{\mathrm{t}})$. A nonparametric regression based on all the data is used to obtain $\hat{J}(0)$, the estimate of $J(0)$. The estimated value of C is then simply equal to $\hat{J}(0)$. This method has important advantages over interpolating for each situation individually. It allows one to use all the data including situations with only a single measurement. Situations with few data points for which interpolation is problematic do not degrade the estimator. Finally, the method automatically gives more weight to data close to the model-based optimal decisions, which should be more useful for estimating the $J_i[d_{\mathrm{M}}(X_i)]$ values.

In the case of multiple decisions, the above method will be difficult or impossible to apply. The problem is that the amount of data necessary to obtain a reasonable estimate of C increases rapidly with the dimension of the decision vector. Antoniadou and Wallach (2000) suggest that their approach may still be feasible in 2 dimensions, but beyond this the data requirements are probably prohibitive.

In the case of multiple decisions, a different approach must be adopted. One possibility would be to restrict the decision space to decisions that were actually applied. Suppose that the N_i decisions $d_{i,1}, \ldots, d_{i,N_i}$ were tested in situation i. The model is then used to evaluate each of these decisions. Suppose that the best decision among these, according to the model, is $d_{i,i*}$. Then the estimator of the criterion C becomes

$$\hat{C} = \frac{1}{N} \sum_{i=1}^{N} J_i\left(d_{i,i*}\right) \tag{31}$$

Table 6 shows artificial data from two situations. Three different decisions were tested in situation 1 and two different decisions in situation 2. The decisions can be vectors. $J_i(d_{i,j})$ and $\hat{J}_i(d_{i,j})$ represent respectively the true value of the objective function and the value of the objective function according to the model, for the jth decision in situation i. According to the model, the best decision for situation 1 is decision A, since that decision has the largest value of $\hat{J}_i(d_{i,j})$ among the decisions that were tested in situation 1. The true value of decision A is 79. The best decision for situation 2 according to the model is decision E with true value 93. The estimated value of the criterion, according to Eq. (31), is $\hat{C} = (1/2)(79 + 93) = 86$.

It is of interest to compare this value with the largest value that could be attained. From Table 6, the true best decisions are respectively B and E for situations 1 and 2. The maximum value of the criterion corresponds to these choices. It is $C_{\max} = (1/2)(113 + 93) = 103$.

Table 6. Model-estimated and true values of objective function for 3 different decisions in situation 1 and 2 different decisions in situation 2.

Situation i	Decision d	Model prediction of objective function $\hat{J}_i(d)$	Measured objective function $J_i(d)$
1	A	89	79
1	B	86	113
1	C	81	85
2	D	28	47
2	E	99	93

5. Using a model to test a scientific hypothesis

Suppose that the model represents a scientific hypothesis about how the real world functions and we wish to test whether this hypothesis is true or not. The basis of the test is simple. If the model is truly a correct representation of reality, and if it is applied within the range of conditions where it is applicable, then observations on the real world and model predictions should coincide perfectly (for the moment we ignore measurement error).

To formalize this idea, we define three logical statements;

A: We observe a system where the model is meant to apply.

B: The model has the same behavior as the true world for systems where the model is meant to apply.

C: The observed output of the real system and the outcomes calculated using the model are identical.

Each statement has one of two values, "true" or "false." The basis of our hypothesis test is the syllogism

IF A AND B, THEN C (32)

Equation (32) says that if both A and B are true then C must be true.

Our objective is to draw conclusions about the statement B concerning the relation between the model and reality. Logic allows us to conclude from Eq. (32) that

IF (NOT C) THEN (NOT A) OR (NOT B) (33)

If "*NOT C*" is true (i.e. if C is false) then either A or B or both must be false. If A is known to be true, B must is false, i.e. if model and observations do not coincide, and the model is applicable, then the model must be false.

Suppose however that the model and the data do agree. It is not correct to conclude from Eq. (32) that *IF C THEN A AND B*, i.e. if model and data agree we cannot conclude that B is true. In other words, it is possible to prove a model incorrect (using Eq. (33)), but it is not possible to prove that it is correct. One can invalidate hypotheses about the real world using a model, but one cannot unequivocally validate them.

If there is measurement error, categorical statements such as "C is true" or "C is false" will be replaced by probabilistic statements like "if C is true, the probability of observing a difference between measured and calculated values at least as large as that actually observed is *p*." If *p* is small (5% for example), one concludes that it is unlikely (but not impossible) that C is true. The model then is probably not identical to reality (assuming A is true). If on the other hand *p* is large, the evidence does not indicate that C is false. As before, however, we cannot conclude from this that the model is probably correct.

In fact, the formal testing of a crop model to determine whether it is "true" or "false" is usually of little interest. A crop model is by design a simplification of reality, so that it is not meant to be identical to the real world. There are few if any cases where we seriously entertain the hypothesis that the model outputs are strictly identical to real-world outputs. It may occur that in a case with measurement error we have a fairly large *p* value, so that we do not conclude that the model is incorrect. However, this will often simply be the result of poor data. Data with large measurement errors will tend to lead to large *p* values.

Does this mean that models have nothing to say about how a crop functions? They do, but not in the sense of proving a hypothesis. A more reasonable question is which description of crop functioning, among a small number of clearly expressed alternatives, is most compatible with observed data or has the smallest prediction error. An example is provided by Gent (1994) who compares three different hypotheses concerning the source of the carbon used for crop respiration. The approach is to develop mathematical models that embody each hypothesis and then to compare the outputs of the different models with observed data. The hypothesis that is preferred is the one that leads to the closest correspondence between observations and calculated values. The objective here is to obtain a better understanding of how the system operates. The methods however are the same as those we have described in the previous sections.

Other discussions of the role of crop models for understanding crop functioning can be found in Doucet and Sloep (1992) and Sinclair and Seligman (2000).

References

Addiscott, T.M., Whitmore, A.P., 1987. Computer simulation of changes in soil mineral nitrogen and crop nitrogen during autumn, winter and spring. Journal of Agricultural Science, Cambridge 109, 141–157.

Antoniadou, T., Wallach, D., 2000. Evaluating decision rules for nitrogen fertilizer. Biometrics 56, 420–426.

Ben Nouna, B., Katerji, N., Mastrolilli, M., 2000. Using the CERES-Maize model in a semi-arid Mediterranean environment. Evaluation of model performance. European Journal of Agronomy 13, 309–322.

Bunke, O., Droge, B., 1987. Estimators of the mean squared error of prediction in linear regression. Technometrics 26, 145–155.

Colson, J., Wallach, D., Bouniols, A., Denis, J.-B., Jones, J.W., 1995. Mean squared error of yield prediction by SOYGRO. Agronomy Journal 87, 397–402.

Doucet, P., Sloep, P.B., 1992. Mathematical Modeling in the Life Sciences. Ellis Horwood Limited, Chichester, England, pp. 489.

Efron, B., 1987. Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association 78, 316–331.

Fila, G., Bellocchi, G., Acutis, M., Donatelli, M., 2003. IRENE: a software to evaluate model performance. European Journal of Agronomy 18, 369–372.

Garnier, P., Néel, C., Mary, B., Lafolie, F., 2001. Evaluation of a nitrogen transport and transformation model in bare soil. European Journal of Soil Science 52, 253–268.

Gauch, H.G., Hwang, J.T.G., Fick, G.W., 2003. Model evaluation by comparison of model-based predictions and measured values. Agronomy Journal 95, 1442–1446.

Gent, M.P.N., 1987. Photosynthate reserves during grain filling in winter wheat. Agronomy Journal 86, 159–167.

Harrell, Jr. F.E., 2001. Regression Modeling Strategies. Springer, New York, pp. 568.

Irmak, A., Jones, J.W., Mavromatis, T., Welch, S.M., Boote, K.J., Wilkerson, G.G., 2000. Evaluating methods for simulating soybean cultivar responses using cross validation. Agronomy Journal 92, 1140–1149.

Jones, P.N., Carberry, P.S., 1987. A technique to develop and validate simulation models. Agricultural Systems 46, 427–442.

Kobayashi, K., Salam, M.U., 2000. Comparing simulated and measured values using mean squared deviation and its components. Agronomy Journal 92, 345–352.

Lin, L., Hedayat, A.S., Sinha, B., Yang, M., 2002. Statistical methods in assessing agreement: models, issues and tools. Journal of the American Statistical Association 97, 257–270.

Loehle, C. 1987. Errors of construction, evaluation, and inference: a classification of sources of error in ecological models. Ecological Modelling 36, 297–314.

Makowski, D., Wallach, D., 2001. How to improve model-based decision rules for nitrogen fertilizer. European Journal of Agronomy 15, 197–208.

Makowski, D., Wallach, D., 2002. It pays to base parameter estimation on a realistic description of model errors. Agronomie 22, 179–189.

Makowski, D., Wallach, D., Meynard, J.-M., 2001. Statistical methods for predicting responses to applied nitrogen and calculating optimal nitrogen rates. Agronomy Journal 93, 531–539.

Mayer, D.G., Butler D.G., 1993. Statistical validation. Ecological Modelling 68, 21–32.

Mitchell, P.L., 1997. Misuse of regression for empirical validation of models. Agricultural Systems 54, 313–326.

Mitchell, P.L., Sheehy, J.E., 1997. Comparison of predictions and observations to assess model performance: a method of empirical validation. Agricultural Systems 54, 437–451.

Pang, X.P., Letey, J., Wu, L., 1997. Yield and nitrogen uptake prediction by CERES-Maize model under semiarid conditions. Soil Science Society of America Journal 61, 254–256.

Power, M. 1993. The predictive validation of ecological and environmental models. Ecological Modelling 68, 33–50.

Robertson, M.J., Carberry, P.S., Huth, N.I., Turpin, J.E., Probert, M.E., Poulton, P.L., Bell, M., Wright, G.C., Yeates, S.J., Brinsmead, R.B., 2002. Simulation of growth and development of diverse legume species in APSIM. Australian Journal of Agricultural Research 53, 429–446.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecological Modelling 90, 229–244.

Sinclair, T.R., Seligman N., 2000. Criteria for publishing papers on crop modelling. Field Crops Research 68, 165–172.

Sokal, R.R., Rolf, F.J., 1981. Biometry. W.H. Freeman and Company, San Francisco, pp. 859.

Swartzman, G.L., Kaluzny, S.P., 1987. Simulation model evaluation. In: Swartzman, G.L., Kaluzny, S.P. (Eds), Ecological Simulation Primer. Macmillan, New York, pp. 209–251.

Wallach, D., Génard, M., 1998. Effect of uncertainty in input and parameter values on model prediction error. Ecological Modelling 105, 337–345.

Wallach, D., Goffinet, B., 1987. Mean squared error of prediction in models for studying ecological and agronomic systems. Biometrics 43, 561–573.

Wallach, D., Goffinet, B., 1989. Mean squared error of prediction as a criterion for evaluating and comparing system models. Ecological Modelling 44, 299–306.

Wallach, D., Loisel, P., Goffinet, B., Habib, R., 1990. Modeling the time dependence of nitrogen uptake in young trees. Agronomy Journal 82, 1135–1140.

Wallach, D., Goffinet, B., Bergez, J.-E., Debaeke, P., Leenhardt, D., Aubertot, J.-N., 2001. Parameter estimation for crop models: a new approach and application to a corn model. Agronomy Journal 93, 757–766.

Waller L.A., Smith, D., Childs, J.E., Real, L.A., 2003. Monte Carlo assessments of goodness-of-fit for ecological simulation models. Ecological Modelling 164, 49–63.

Willmott, C.J., 1981. On the validation of models. Physical Geography 2, 184–194.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M., 1987. Statistics for the evaluation and comparison of models. Journal of Geophysical Research 90, 8995–9005.

Exercises

1. Table 7 shows data for yield from 6 different plots. The data consist of the value of the single model explanatory variable x, the measured values Y and the values calculated using the model $f(x;\hat{\theta}) = \hat{a} + \hat{b}x$ with $\hat{a} = 0.5$ and $\hat{b} = 0.15$.

Table 7. Measured values of an output variable Y, value of explanatory variable x and calculated values $f(x;\hat{\theta})$ for 6 situations.

Y	x	$f(x;\hat{\theta})$
0.61	1.0	0.65
1.18	2.5	0.88
1.38	4.0	1.10
1.91	5.5	1.32
2.01	7.0	1.55
1.81	8.5	1.78

   (a) Plot calculated *versus* measured values. Plot the residues.
   (b) Calculate the measures of agreement presented in Table 2 for this model and these data. For the threshold measures calculate *TDI*(25%) and *CP*(0.2). Also, calculate the 3 terms in the decomposition of *MSE* from Eq. (12).
   (c) Suppose that instead of the model of Table 7 we use the model $f(X;\hat{\theta}) = \bar{Y}$, where $\bar{Y}$ is the average of the measured yields. Plot calculated *versus* measured values for this model. Plot the residues for this model.
   (d) Calculate the measures of agreement in Table 2 for the model $f(X;\hat{\theta}) = \bar{Y}$. For the threshold measures calculate *TDI*(25%) and *CP*(0.2). Also, calculate the 3 terms in the decomposition of *MSE* from Eq. (12).
   (e) What are your conclusions about these two models?

2. Under what conditions is *MSE* an unbiased estimator of the mean squared error of prediction?

3. In Example 2 in the text, suppose that $\mathrm{var}(x^{(3)}) = 4.0$ but that all the other conditions remain as stated in the example.

   (a) What are the values of Λ, Δ and $MSEP(\hat{\theta})$ now?
   (b) The model has not changed, so why has the mean squared error of prediction of the model changed compared to the value in Example 2?
   (c) Invent a real-world example that could cause such a change in $\mathrm{var}(x^{(3)})$.

4. Suppose that the true relation that gave rise to the data in Table 7 is $Y = \theta_1(1 - \mathrm{e}^{-\theta_2 x}) + \varepsilon$ with $\theta_1 = 2.0$, $\theta_2 = 0.4$, $E(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = 0.1$. Suppose that the target distribution consists of just the 6 situations of Table 7. Finally, suppose that the data have not been used during model development.

   (a) What are the values of Λ, Δ and $MSEP(\hat{\theta})$ for this model and target distribution?
   (b) What is the major source of error?
   (c) What is the value of $MSEP(\hat{\theta})$ estimated from the available data?

5. Suppose that in Exercise 4, the model is $f(x; \hat{\theta}) = \hat{a} + \hat{b}x$ but now the parameters $\hat{a}$ and $\hat{b}$ are adjusted to the data in Table 7 using the least squares criterion.
   - (a) What are the adjusted parameter values?
   - (b) What is the value of *MSE* for this adjusted model?
   - (c) Estimate $MSEP(\hat{\theta})$ for this model using cross-validation. What is the estimated value of $MSEP(\hat{\theta})$? Compare *MSE* and $MSEP(\hat{\theta})$ and explain.
6. Suppose that the measured Y values in an experiment have measurement error and that the variance of the measurement error is 0.3.
   - (a) If $MSEP(\hat{\theta})$ for measured Y is 0.4, what is $MSEP(\hat{\theta})$ for the true Y?
   - (b) Is it worthwhile to try to reduce the prediction error? Explain.
7. Suppose that a crop model has 5 explanatory variables that are measured with error. The errors are independent, and in each case the variance of the measurement error is 2.0. The vector of partial derivatives of the model with respect to the explanatory variables is $(x^{(1)}, x^{(2)}, 3 + x^{(3)}, x^{(4)2}, x^{(5)})^{\mathrm{T}}$. In the target population, $E(x^{(i)}) = 0$, $E(x^{(i)2}) = 3.0$ and $E(x^{(i)4}) = 5.0$ for $i = 1, \ldots, 5$.
   - (a) What is the contribution of the error in explanatory variables to $MSEP(\hat{\theta})$?
   - (b) Suppose that the variance of the explanatory variables were halved to 1.0. What would be the contribution to $MSEP(\hat{\theta})$ now?
8. When a model is biased, for example giving results that systematically under-predict, it is tempting to adjust the model by adding a constant that removes the bias. Bias removal adds one parameter (the estimated bias) to the model. We will apply bias removal to the model and data of Table 7.
   - (a) For the data and model in Table 7, what is the estimated bias?
   - (b) What is the modified model after bias removal? What is the bias of this modified model?
   - (c) What is the value of *MSE* for the modified model?
   - (d) Is $MSEP(\hat{\theta})$ for the modified model necessarily smaller than for the original model? Why?
   - (e) What is the value of $MSEP(\hat{\theta})$ for the modified model estimated by cross-validation? Compare with the estimated value of $MSEP(\hat{\theta})$ of the original model. Is bias removal worthwhile in this case?
9. We can consider bias removal in a more general context than for a particular model and data set. Very generally, we want to compare $E_{\hat{\theta}}[MSEP(\hat{\theta})]$ for a model before and after bias removal. Suppose that we have a data set with N measurements of Y, which represent a random sample from the target distribution. Suppose that we have some model, referred to as the *initial* model, which is independent of those data. The predictions of that model are noted $\hat{Y}_{\text{initial}}$. The model after bias removal will be referred to as the *unbiased* model, and the predictions according to this model will be noted $\hat{Y}_{\text{unbiased}}$.
   - (a) Let $\hat{bias}$ be the bias of the *initial* model estimated from the data. What is the expression for $\hat{bias}$? What is the expression for $\hat{Y}_{\text{unbiased}}$ in terms of $\hat{Y}_{\text{initial}}$

and $\hat{bias}$? Is the *unbiased* model independent of the data? If not, how many parameters in the *unbiased* model are estimated from the data. What are the expressions for the parameters in terms of the data?
(b) What is the expression for $\Lambda_{\text{unbiased}}$ in terms of Λ_{initial}? Does the fact of removing bias change the explanatory variables of the model?
(c) Let *bias* represent the true value of bias. What is the expression for *bias* in terms of Y and $\hat{Y}_{\text{initial}}$. If $\text{var}(Y) = \sigma^2$, what is the expression for $\text{var}(\hat{bias})$?
(d) In $\Delta_{E_{\hat{\theta}}[f(X;\hat{\theta})],\text{unbiased}}$, we will consider only the uncertainty in the extra parameter of the *unbiased* model compared to the *initial* model. Derive an expression for $\Delta_{E_{\hat{\theta}}[f(X;\hat{\theta})],\text{unbiased}}$ in terms of $\Delta_{E_{\hat{\theta}}[f(X;\hat{\theta})],\text{initial}}$, *bias* and $\text{var}(\hat{bias})$. (Hint: in the expression for $\Delta_{E_{\hat{\theta}}[f(X;\hat{\theta})],\text{unbiased}}$, add and subtract the term bias. You will then be able to write this as a sum of two terms. The expectation over the uncertain parameters will only concern the second term. The first term will have the general form $E\{[T - E(T)]^2\} = E(T^2) - [E(T)]^2$).
(e) Express $E_{\hat{\theta}}[MSEP_{\text{unbiased}}(\hat{\theta})]$ in terms of $E_{\hat{\theta}}[MSEP_{\text{initial}}(\hat{\theta})]$. Under what conditions does removing bias reduce prediction error? If the sample size N is increased, how is the effect of bias removal modified?

10. Suppose that there is a model which predicts $\hat{J}_i(d)$ values of 70, 96, 74, 65 and 64, respectively for the five situations in Table 6.

(a) What is the estimated value of the criterion C for this model?
(b) Based on the data in Table 7, and supposing that the objective is to use the model for management, which model would be preferred between this new model and the model which gave the results in the table? Why?

11. A scientist entertains two theories about root activity in young peach trees. One theory is that activity (represented by uptake per unit root length) is uniform from spring to autumn. The second is that root activity is the greatest in spring. How would you test these theories? What conclusions exactly could be drawn from the results?

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 3

Uncertainty and sensitivity analysis for crop models

H. Monod, C. Naud and D. Makowski

1. **Introduction** 55
2. **Ingredients of uncertainty and sensitivity analysis** 57
3. **Uncertainty analysis** 62
4. **Sensitivity analysis** 67
5. **Discussion** 93
References 94
Exercises 97

1. Introduction

A crop model is the result of a long and complex construction process, involving data at multiple stages for understanding basic processes, elaborating model structure, estimating parameters and evaluating prediction quality. In various stages of a model's life, however, there is a need to study the model on its own, with an emphasis on its behaviour rather than on its coherence with a given data set. This is where uncertainty analysis sensitivity analysis and related methods become useful for the modeller or model user.

Uncertainty analysis consists of evaluating quantitatively the uncertainty or variability in the model components (parameters, input variables, equations) for a given situation, and deducing an uncertainty distribution for each output variable rather than a misleading single value. An essential consequence is that it provides methods to assess, for instance, the probability of a response to exceed some threshold. This makes uncertainty analysis a key component of **risk analysis** (Vose, 1996).

The aim of **sensitivity analysis** is to determine how sensitive the output of a crop model is, with respect to the elements of the model which are subject to uncertainty or variability. This is useful as a guiding tool when the model is under development as well as to understand model behaviour when it is used for prediction or for decision support. For dynamic models, sensitivity analysis is closely related to the study of **error propagation**, i.e. the influence that the lack of precision on model input will have on the output.

Because uncertainty and sensitivity analysis usually relies on simulations, they are also closely related to the methods associated with **computer experiments**. A computer experiment is a set of simulation runs designed in order to explore efficiently the model responses when the input varies within given ranges (Sacks et al., 1989; Welch et al., 1992). The goals in computer experiments identified by Koehler and Owen (1996) include optimization of the model response, visualization of the model behaviour, approximation by a simpler model or estimation of the average, variance or probability of the response to exceed some threshold.

Within a given model, model equations, parameters and input variables are all subject to variability or uncertainty. First, choices have to be made on the model structure and on the functional relationships between input variables and state and output variables. These choices may sometimes be quite subjective and it is not always clear what their consequences will be. Martinez et al. (2001) thus perform a sensitivity analysis to determine the effects of the number of soil layers on the output of a land surface–atmosphere model. For spatial models, there is frequently a need to evaluate how the scale chosen for input variables affects the precision of the model output (see e.g. Salvador et al., 2001).

Second, parameter values result from estimation procedures or sometimes from bibliographic reviews or expert opinion. Their precision is necessarily limited by the variability and possible lack of adequacy of the available data. Some parameters may also naturally vary from one situation to another. The uncertainty and natural variability of parameters are the central point of many sensitivity analyses. Bärlund and Tattari (2001), for example, study the influence of model parameters on the predictions of field-scale phosphorus losses, in order to get better insight into the management model ICECREAM. Ruget et al. (2002) perform sensitivity analysis on parameters of the crop simulation model STICS, in order to determine the main parameters that need to be estimated precisely. Local sensitivity methods, based on model derivatives with respect to parameters, are commonly used for checking identifiability of model parameters (Brun et al., 2001).

Third, additional and major sources of variability in a model output are, of course, the values of its input variables. Lack of precision when measuring or estimating input variables needs to be quantified when making predictions from a model or when using it for decision support. Aggarwal (1995) thus assesses the implications of uncertainties in crop, soil and weather inputs in the spring wheat WTGROWS crop model. Rahn et al. (2001) compare contrasted input scenarios for the HRI WELL-N model on crop fertilizer requirements through a sensitivity analysis. They identify the main factors which need to be measured precisely to provide robust recommendations on fertilization. Contrasted settings of the input variables are used for performing sensitivity or uncertainty analyses assuming different scenarios by Dubus and Brown (2002).

Model structure, model parameters and input variables represent three basic sources of model uncertainty. It is often advisable to study their influence on a model simultaneously (Saltelli et al., 2000) and alternative groupings of uncertainty sources may then be more adequate. Rossing et al. (1994), for example, distinguish sources that can be controlled by more intensive data collection (model parameter estimates), and uncontrollable sources when predictions are made (daily temperature, white noise). Ruget et al. (2002), on the other hand, decompose the sensitivity analyses according to STICS sub-modules on, e.g. energy conversion, rooting or nitrogen absorption. Jansen et al. (1994) advocate to divide uncertainty sources into groups of parameters or input variables which can be considered to be mutually independent.

As shown by the examples above, uncertainty and sensitivity analysis may have various objectives, such as:

- to check that the model output behaves as expected when the input varies;
- to identify which parameters have a small or a large influence on the output;

- to identify which parameters need to be estimated more accurately;
- to detect and quantify interaction effects between parameters, between input variates or between parameters and input variates;
- to determine possible simplification of the model;
- to identify input variables which need to be measured with maximum accuracy.

Some of these objectives have close links with other methods associated with modelling, like model construction, parameter estimation or model use for decision support.

The diversity of motivations for performing sensitivity analysis is associated with a large choice of methods and techniques. In this chapter, we present a selection of approaches representative of this diversity. This selection, however, will be far from exhaustive. We refer to the book edited by Saltelli et al. (2000) for a recent and comprehensive exposition of sensitivity analysis methods and applications, and to Saltelli et al. (2004) for a more practical presentation.

In this chapter, Section 2 is dedicated to preliminary notions on the basic components of an uncertainty and sensitivity analysis. Section 3 covers several methods of uncertainty analysis. Methods of sensitivity analysis are presented in Section 4 – local and one-at-a time sensitivity analysis methods, and more global methods (variance-based sensitivity analysis) which enable to study simultaneously the influence of several model components.

2. Ingredients of uncertainty and sensitivity analysis

2.1. The crop model

The structure and properties of the crop model may influence the choice of the uncertainty and sensitivity analysis. One reason is that the objectives depend on the crop model capabilities and complexity.

More specifically, as remarked by Koehler and Owen (1996), the number of inputs (variables or parameters), the number of outputs and the speed with which the model f can be calculated may vary enormously in applications, and these quantities will obviously play an important role in the objectives of a sensitivity analysis and on the adequacy of the various available methods. Among the methods presented in the sequel, some are adapted to small numbers of model simulations (e.g. local and one-at-a-time methods, methods based on experimental designs), while others require a large number of simulations (methods based on Monte-Carlo sampling, for instance).

A price has to be paid while using more economical methods, and this price depends on the main model properties – it may be necessary to select a number of factors smaller than desired, or most interactions between factors may have to be assumed as negligible, or the investigation may be unable to detect model departures from linearity or near-linearity. It follows that some methods are well-adapted only if the model is well-behaved in some sense, while other methods are more "model-independent" (Saltelli et al., 1999), i.e. more robust to complex model behaviours such as strong non-linearity, discontinuities, non-monotonicity or complex interactions between factors.

2.2. *Input factors*

The model components whose influence on the output is to be investigated will be called the *input factors* of the sensitivity analysis. An input factor may be:

- either a set of alternative model structures or functional relationships within a sub-module of the model;
- or an uncertain or variable parameter θ_j;

A winter wheat dry matter model

A simple crop model will be used in this chapter to illustrate the different methods of uncertainty and sensitivity analysis. The model has a single state variable, the above-ground winter wheat dry matter, denoted by $U(t)$ with t the day number since sowing. This state variable is calculated on a daily basis as a function of cumulative degree-days $T(t)$ (above a baseline of 0°C) and of daily photosynthetically active radiation $\mathrm{PAR}(t)$. The model equation is:

$$U(t+1) = U(t) + E_{\mathrm{b}} E_{\mathrm{imax}} \left[1 - \mathrm{e}^{-K.\mathrm{LAI}(t)}\right] \mathrm{PAR}(t) + \varepsilon(t),$$

with E_{b} the radiation use efficiency, E_{imax} the maximal value of the ratio of intercepted to incident radiation, K the coefficient of extinction, $\mathrm{LAI}(t)$ the leaf area index on day t, and $\varepsilon(t)$ a random term representing the model error. In this chapter, we consider the deterministic part of the model only, so this model error will be assumed null in the simulations. $\mathrm{LAI}(t)$ is calculated as a function of cumulative degree-days $T(t)$, as follows (Baret, 1986):

$$\mathrm{LAI}(t) = L_{\max} \left\{ \frac{1}{1 + \mathrm{e}^{-A[T(t)-T_1]}} - \mathrm{e}^{B[T(t)-T_2]} \right\}.$$

The dry matter at sowing ($t = 1$) is set equal to zero: $U(1) = 0$. In addition, the constraint $T_2 = \frac{1}{B} \log[1 + \exp(A \times T_1)]$ is applied, so that $\mathrm{LAI}(1) = 0$.
We will assume that the *dry matter at harvest* $U(t_{\mathrm{H}})$ is the main output variable of interest, and denote

$$\begin{aligned} \hat{Y} &= U(t_{\mathrm{H}}) \\ &= \sum_{t=1}^{t_H - 1} E_{\mathrm{b}} E_{\mathrm{imax}} \left[1 - \mathrm{e}^{-K\mathrm{LAI}(t)}\right] \mathrm{PAR}(t) \end{aligned}$$

While presenting sensitivity analysis, it is convenient to consider the model in the form

$$\hat{Y} = f(X; \theta).$$

In this expression, $X = (T(1), \ldots, T(t_H), \mathrm{PAR}(1), \ldots, \mathrm{PAR}(t_H))$ denotes the daily climate input variables, and $\theta = (E_{\mathrm{b}}, E_{\mathrm{imax}}, K, L_{\max}, A, B, T_1)$ denotes the vector of parameters, with $L_{\max}$ the maximal value of LAI, T_1 a temperature threshold and A and B two additional parameters.

- or an input variable X_l;
- or a series of several related input variables X_l, e.g. annual series of daily climate variables in a given region.

The choice of the input factors depends on the objective of the sensitivity analysis. They must include, of course, the model components of direct interest in the study. But in many cases, the sensitivity of the model with respect to these components is likely to depend on additional components. For instance, the sensitivity of a crop model with respect to its main parameters is often highly dependent on the values of climate- or soil variables. Consequently, these variables must also be considered for inclusion in the list of input factors, unless, alternatively, separate sensitivity analyses are performed with different modalities of these variables.

Note that each input variable of the model may or may not be selected as an input factor of the sensitivity analysis. For instance, if a sensitivity analysis is performed for a given soil type, the input variables related to soil can be fixed. In this case, the soil input variables will not be included among the input factors of the sensitivity analysis. The term *input factor* is further reserved for factors of the sensitivity analysis.

Notation

The number of input factors will be denoted by s and the input factors will be denoted by $Z_1, \ldots, Z_s$, in order to distinguish them clearly from the *model input variables* X_l. An *input scenario* will be defined as a combination of levels $\mathbf{z} = (z_1, \ldots, z_s)$ of the sensitivity input factors. When several input scenarios need to be defined simultaneously, they will be denoted by $\mathbf{z}_k = (z_{k,1}, \ldots, z_{k,s})$, with subscript k identifying the scenarios.

Whatever the choice of the factors, it is assumed that for each input scenario $\mathbf{z}$, the other crop model components f, $\mathbf{x}$ and θ are completely determined so that the output $f(\mathbf{x}, \theta)$ can be calculated. We will keep the same notation f to identify the model expressed as a function of input variables $f(\mathbf{x}, \theta)$ or as a function of an input scenario $f(\mathbf{z}) = f(z_1, \ldots, z_s)$.

A winter wheat dry matter model (continued)

In the winter wheat dry matter model, the seven parameters have associated uncertainty and so they represent seven input factors for the uncertainty and sensitivity analyses. The other source of uncertainty to be considered in this example is that related to the input variables of the model. Instead of considering each input variable $\mathrm{PAR}(t)$ and $T(t)$ at each time t as a separate sensitivity input factor, a set of fourteen annual series of climate measurements in the region of interest will constitute the eighth factor of the sensitivity analysis.

Thus, there are eight factors: the seven parameters E_{b}, E_{imax}, K, $L_{\max}$, A, B, T_1 and the climate factor C. An input scenario is a vector

$$\mathbf{z} = (z_{E_{\mathrm{b}}}, z_{E_{\mathrm{imax}}}, z_K, z_{L_{\max}}, z_A, z_B, z_{T_1}, z_C)$$

specifying a combination of values of the input parameters. As this example shows, a factor may be quantitative – the seven parameters – or categorical – the climate series.

2.3. Uncertainty in input factors

For each input factor, the amount of uncertainty needs to be defined. The uncertainty in an input factor can be described in different ways. For a parameter, it is often given as the most likely value plus or minus a given percentage. Or it is specified through a continuous probability distribution over a range of possible values. The uncertainty about climate can either be summarized by series of climatic variable values measured during 10, 20 or 30 years, or be simulated by a climate generator (Racsko et al., 1991).

In this chapter, three main characteristics are considered for describing the uncertainty: nominal values, uncertainty domains and probability distributions.

The *nominal* value $z_{0,i}$ of an input factor Z_i represents the most standard setting of the corresponding model parameter or input variable in the conditions of the study. The *control scenario* $\mathbf{z}_0$ is defined as the input scenario with each input factor fixed at its nominal value. These notions are useful, in particular, for local sensitivity methods (see Section 4).

The *uncertainty range* represents the set of possible values for an input factor. Usually,

- for a parameter θ_j, it is an interval $[\theta_{\min(j)}, \theta_{\max(j)}]$ around the nominal value, representing the uncertainty range of the parameter values based on bibliography, expert opinion or experimental data;
- for a quantitative input variable X_l, it represents the range of variation $[x_{\min(l)}, x_{\max(l)}]$ under the conditions of the study; alternatively, it can be chosen to reflect the lack of precision when this variable is measured in a given field;
- for categorical factors, it is a set of modalities representative of the phenomenon under study; for climate series, typically, the domain of variation is a set of recently observed annual series in one or several sites.

Except for input factors with a negligible influence on model output, the influence of any given input factor will appear stronger if its uncertainty range is enlarged compared to other factors. Consequently, the uncertainty ranges must be tuned as finely as possible to the objectives and scales of the study.

Probability distributions must be specified for the methods of sensitivity analysis based on random sampling. The uniform distribution, which gives equal weight to each value within the uncertainty range, is commonly used in sensitivity analysis when the main objective is to understand model behaviour. In uncertainty analysis, more flexible probability distributions are usually needed to represent the input uncertainty (see Section 3). Practical methods to determine distributions from data or expert opinion are presented in Chapters 7 and 8 of Vose (1996).

Coding of input factors. It often simplifies presentation and calculation, when a common uncertainty range is used for all quantitative sensitivity factors. This may be done by coding the levels of the factors so that they vary between −1 and +1 or between 0 and 1. Coded values z_i^c of an input factor Z_i can easily be calculated from the uncoded values through the following relationship:

$$z_i^c = \frac{z_i - (z_{\min(i)} + z_{\max(i)})/2}{(z_{\max(i)} - z_{\min(i)})/2} \quad \text{for a } [-1, +1] \text{ range of variation,}$$

or

$$z_i^c = \frac{z_i - z_{\min(i)}}{z_{\max(i)} - z_{\min(i)}} \quad \text{for a [0, 1] range of variation.}$$

A winter wheat dry matter model (continued)

The chosen nominal values and uncertainty ranges are given in Table 1 for the parameters. These values come from past experiments, bibliography and expert knowledge. For the climate factor, a set of 14 annual series observed in the region of Grignon (France) was chosen. Note that for such a factor, there is no obvious nominal value.

Table 1. Uncertainty intervals for the parameters of the winter wheat dry matter models.

Parameter	Unit	Nominal value	Uncertainty range	
E_b	g/MJ	1.85	0.9	2.8
E_{imax}	–	0.94	0.9	0.99
K	–	0.7	0.6	0.8
L_{max}	–	7.5	3	12
T_1	C	900	700	1100
A	–	0.0065	0.003	0.01
B	–	0.00205	0.0011	0.003

To illustrate the coding of factors, let us consider the parameter E_b. The values z_{Eb}^c of E_b vary in the uncertainty range [0.9, 2.8]. By setting $z_{Eb}^c = (z_{Eb} - 1.85)/0.95$, we get coded values z_{Eb}^c which vary in $[-1, +1]$.

2.4. Methods of uncertainty and sensitivity analysis

An uncertainty analysis can be used to answer the question *What is the uncertainty in* $\hat{Y} = f(Z)$ *given the uncertainty in* Z*?* This type of analysis consists of four steps:

i. Definition of the distribution of the uncertain input factors.
ii. Generation of N scenarios of the input factors $\mathbf{z}_k = (z_{k,1}, \ldots, z_{k,s})$, $k = 1, \ldots, N$.
iii. Computation of the model output for each scenario, $f(\mathbf{z}_k)$, $k = 1, \ldots, N$.
iv. Analysis of the output distributions (computation of means, variances, quantiles ...).

These steps are discussed in details in Section 3.

Two types of sensitivity analysis are usually distinguished, *local* sensitivity analysis and *global* sensitivity analysis. Local SA focus on the local impact of the factors on the model outputs and is carried out by computing partial derivatives of the output variables with respect to the input factors. With this kind of methods, the factors are allowed to vary within small intervals around nominal values. These intervals are not related to the uncertainty in the factor values.

Global sensitivity analysis can be used to answer the question *How important are the individual elements of Z with respect to the uncertainty in $\hat{Y} = f(Z)$?* Like uncertainty analysis, global SA consists in (i) defining the distributions of the input factors, (ii) generating scenarios of input factors and (iii) computing output variables for each scenario. But the fourth step is different and consists of calculating a sensitivity index for each element of Z. These indices are computed varying the factors over their whole uncertainty ranges. Methods of global sensitivity analysis are very useful because they allow the crop modeller to identify the factors that deserve an accurate measure or estimation. Most of Section 4 is devoted to these methods.

3. Uncertainty analysis

3.1. Probability distributions for input factors

The first step of an uncertainty analysis is to define the probability distributions for the input factors. When performing an uncertainty analysis, attention must be paid choosing in adequate probability distributions. The range of input values usually has more influence on the output than the distribution shapes, but some characteristics such as the degree of symmetry or skewness may also play a role.

There is a large choice of probability distributions available. In this section, we give a brief overview and refer to Vose (1996) for a more detailed presentation. The uniform distribution puts equal weight on each value in the uncertainty range. In most cases, however, the extreme values of the uncertainty ranges are less likely than the middle values. Among symmetric distributions, the well-known Gaussian distribution is often convenient since it requires only the specification of a mean value and a standard deviation. In uncertainty analysis, it is often replaced by the truncated Gaussian distribution or by symmetric beta distributions, which give upper- and lower bounds to the possible values.

Sometimes the distribution should be asymmetric, for example if the input parameter or variable is positive and likely to be near zero. Then log-normal, gamma or beta distributions offer a large range of possibilities.

Finally the triangular distributions (or more general piecewise-linear distributions) are often convenient for a simple representation of subjective beliefs, because they are defined entirely by their uncertainty range and their most-likely value. The distribution is zero outside the uncertainty range, it is maximum at the most-likely value, and it is linear between the extreme values of the range and the most-likely value.

3.2. Generation of input factor values

Once the probability distributions have been specified, representative samples have to be drawn from these distributions. This is done most often by Monte Carlo sampling. In Monte Carlo sampling, the samples are drawn independently, and each sample is generated by drawing independently the value of each sensitivity factor Z_i. Note that many mathematical or statistical softwares include routines for quasi-random number generation, so that Monte Carlo samples are quite easy to generate. Provided the quasi-random generators are reliable, Monte Carlo sampling provides unbiased estimates of the expectation and variance of each output variable.

Latin hypercube, importance and LP_τ sampling (see Helton and Davis, 2000) are alternatives to Monte Carlo sampling. The basic principle of Latin hypercube sampling is briefly described here in the case of uniform distributions. First, the range of each factor is divided into P intervals of equal probability and one value is selected at random from each interval. Second, the P values obtained for the factor Z_1 are paired at random and without replacement with the P values obtained for the factor Z_2. The P pairs are then randomly combined without replacement with the P values obtained for the factor Z_3 and so on. The interest of Latin hypercube sampling is that it ensures the full coverage of the range of variation of each factor. A drawback of this method is that it gives biased estimates of the variance. According to Helton and Davis (2000), Latin hypercube sampling is useful when large samples are not computationally practicable and the estimation of very high quantiles is not required.

For illustration, we generated two samples of 10 values of a pair of independent and uniformly distributed random variables, $Z_1 \sim U(0, 1)$ and $Z_2 \sim U(0, 1)$. One sample was generated by Monte Carlo sampling (Fig. 1a) and the other one by Latin hypercube sampling (Fig. 1b). The results show that the values generated by Latin hypercube sampling cover the whole ranges of variation of the random variables. This is not necessarily the case when a Monte Carlo method is used as shown in Figure 1a.

It is necessary sometimes to consider correlations between some input parameters or variables. This requires generating samples from joint multivariate probability distributions. When the distributions are normal, the following method can be used. Assume that the vector $\mathbf{Z} = (Z_1, \ldots, Z_s)^T$ is distributed as $\mathbf{Z} \sim N(0, \Sigma)$, where Σ is a $(s \times s)$ variance–covariance matrix. Define U as an upper triangular matrix such that $\Sigma^{-1} = U^T U$ (Cholesky decomposition). The vector $U\mathbf{Z}$ is normally distributed with mean equal to zero and with a variance–covariance matrix equal to the identity matrix: $\text{var}(UZ) = UU^{-1}(U^T)^{-1}U^T = I$. A random value $\mathbf{z}$ of $\mathbf{Z}$ is obtained by generating a vector $\mathbf{d}$ including s values randomly generated from $N(0, 1)$ and then by calculating $U^{-1}\mathbf{d}$.

When the input factors are not normally distributed, the method proposed by Iman and Conover (1982) can be used to generate samples from joint multivariate probability distributions. Taking account of correlations is particularly important for series of climatic variables. As mentioned before, this case can be tackled by using past climatic series or climate generators.

3.3. Computation of the model output for each scenario

Once the sample of factor values, $\mathbf{z}_1, \ldots, \mathbf{z}_N$, have been generated, the corresponding model output values, $f(\mathbf{z}_1), \ldots, f(\mathbf{z}_N)$, must be computed. If the computation of the model output requires a lot of time, this step may be difficult to carry out. With some very complex models, the sample size N must be set equal to a small value due to the computation time. This problem is illustrated in Chapter 16. On the contrary, this step is straightforward for models that are less complex and computationally intensive as shown by Makowski et al. (2004) with the AZODYN model.

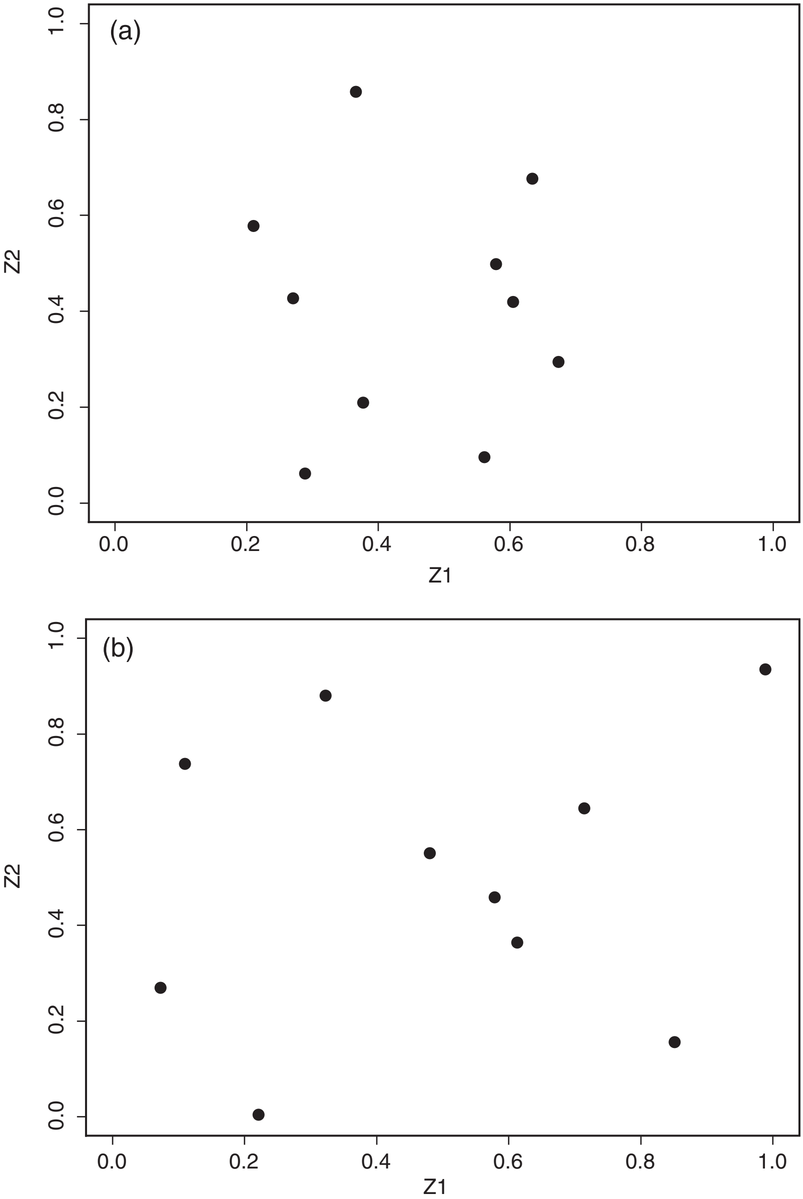


Figure 1. Samples of 10 values of two independent random variables. $Z_1 \sim U(0, 1)$ and $Z_2 \sim U(0, 1)$ obtained by Monte-Carlo sampling (a) and by Latin hypercube sampling (b).

3.4. Analysis of the output distribution

The last step of the analysis is to summarize the values of $f(\mathbf{z}_1), \ldots, f(\mathbf{z}_N)$. Different quantities can be easily calculated. For example, when $f(\mathbf{Z})$ is a single output variable, estimates of the expected value and variance of $f(\mathbf{Z})$ are given by $\bar{f} = \sum_{k=1}^{N} f(\mathbf{z}_k)/N$ and $1/(N-1)\sum_{j=1}^{N}\left[f(\mathbf{z}_j) - \bar{f}\right]^2$ respectively. It is also useful to estimate the quantiles associated to the distribution and the probabilities that $f(\mathbf{Z})$ is lower than some thresholds. Lacroix et al. (2005) used this method to study the risk of pollution of water by nitrate from crop model simulations. The probabilities are often plotted as a function of the threshold values and the resulting curve is called *cumulative probability distribution*.

The quantile q, defined by $P[f(\mathbf{Z}) < q] = \alpha$, can be estimated as follows. The first step is to order the output values. The ordered values are noted $f(\mathbf{z}_{(1)}), \ldots, f(\mathbf{z}_{(i)}), \ldots, f(\mathbf{z}_{(N)})$. The second step is to determine the value i such as $(i-2)/(N-1) \leq \alpha < (i-1)/(N-1)$. The quantile is then defined by $\hat{q} = s \times f(\mathbf{z}_{(i-1)}) + (1-s)f(\mathbf{z}_{(i)})$ where $s = i - 1 + (1-N)\alpha$.

A histogram representation of the output variable values can also provide interesting information as shown in the following example.

A winter wheat dry matter model (continued)

The winter wheat dry matter model was used to compare uncertainty analyses with the uniform distribution and with a symmetric bell-shaped distribution. For sake of simplicity, we considered the Beta distribution with both shape parameters equal to 5, denoted by Beta(5,5). This distribution is symmetric and bounded between 0 and 1, and it puts more weight on the middle values of the [0, 1] interval (see Fig. 2). By applying the transformation $z = z_{\min(i)} + B \times (z_{\max(i)} - z_{\min(i)})$ where B follows a Beta(5,5) distribution, it yields a similar distribution over the uncertainty range of Z_i.

In the second stage of the uncertainty analyses, $N = 5000$ scenarios were generated, using the generators of quasi-random numbers implemented in the *R* software (www.r-project.org) for uniform or Beta distributions. For each scenario, the climatic year was chosen at random with equi-probability. The values of the seven parameters were generated one after another, assuming independence between factors.

In the third stage, the biomass at harvest was calculated with the model for each simulated scenario.

The fourth stage here included a histogram representation of the output and the calculation of basic statistics. The histograms of the model responses are shown in Figure 2. When input data was generated assuming a uniform distribution, combinations of parameter values very unlikely in practice appeared quite frequently, giving extreme output values. By contrast, the Beta(5,5) distribution made samples with extreme values of several factors very rare and the output distribution was much less flat. Some statistics on the simulated output distributions are given in Table 2.

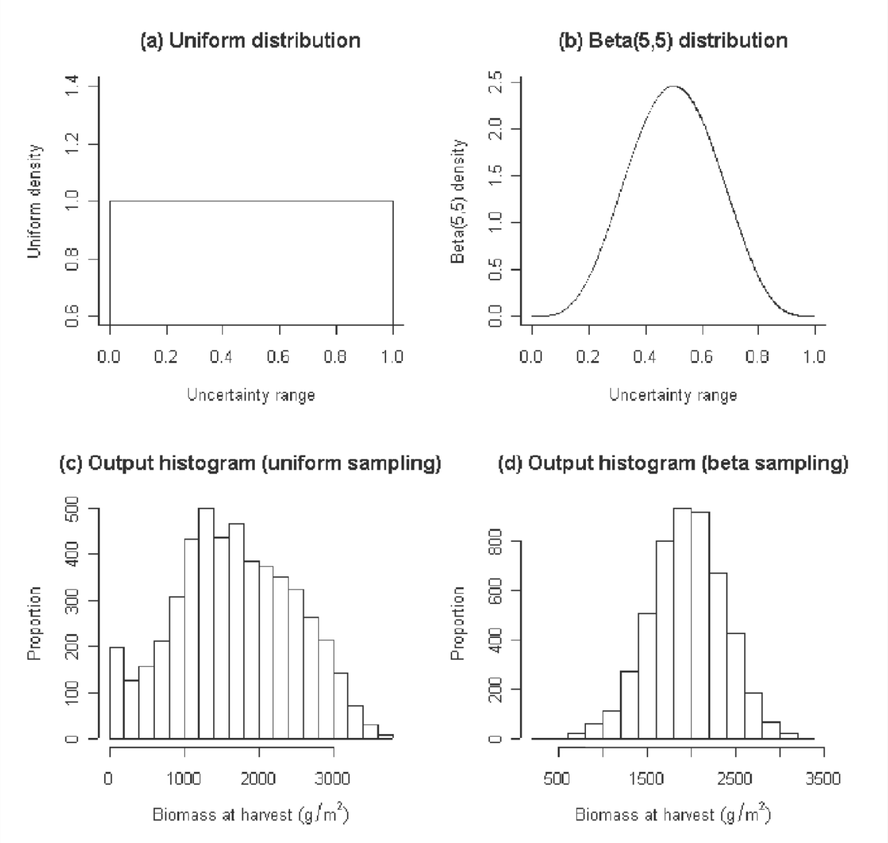


Figure 2. Density functions of the standard uniform (a) and Beta(5, 5) distributions (b). Histograms of the winter wheat model output (biomass at harvest) from samples of size 5000, generated assuming the uniform (c) or Beta(5, 5) (d) distributions.

Table 2. Some statistics on the biomass distributions (g/m^2) resulting from the uncertainty analyses.

Sampling	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Uniform	0	1119	1636	2257	3785
Beta(5,5)	134	1665	1955	2233	3305
	Mean	Standard deviation			
Uniform	1669	785			
Beta(5,5)	1946	420			

4. Sensitivity analysis

4.1. Overview of the different methods

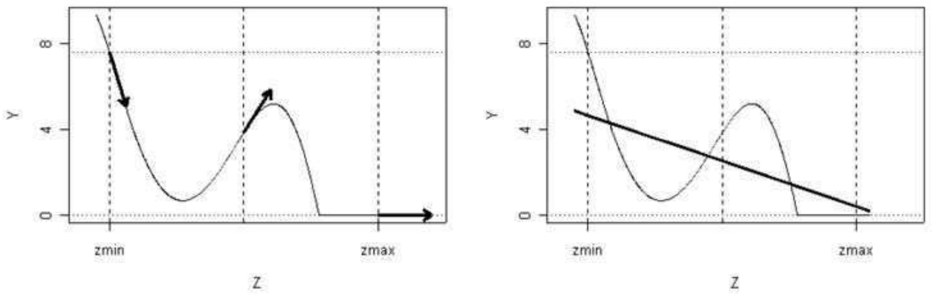
There are many different ways to define sensitivity of a model with respect to its inputs. This section presents the main approaches, without detailing precise criteria. The sensitivity with respect to a single input factor is first considered, then the sensitivities with respect to several factors.

4.1.1. One input factor

Figure 3 illustrates the basic approaches to measure sensitivity from the relationship between a single input factor Z and a model output $\hat{Y} = f(Z)$.

Local sensitivity analysis is based on the local derivatives of output $\hat{Y}$ with respect to Z, which indicate how fast the output increases or decreases *locally* around given values of Z. The derivatives can sometimes be calculated analytically, but they are usually calculated numerically for complex models. Problems may arise if the derivative of the model does not exist at some points. In addition, the derivatives may depend strongly on the Z-value. This problem is illustrated in Figure 3a where three derivatives are reported.

(a) Derivatives **(c) Average slope**



(b) Range **(d) Variance**

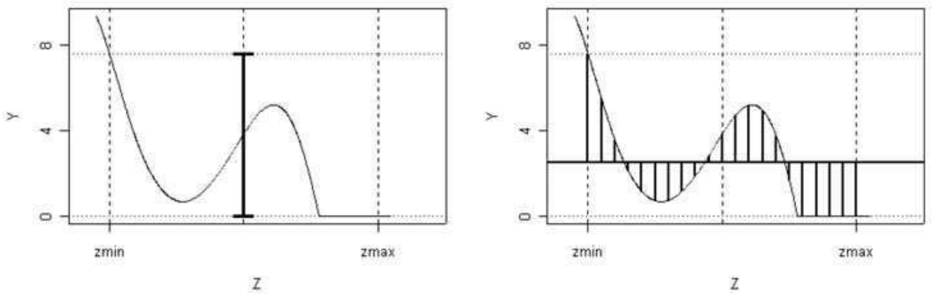


Figure 3. Bases for defining sensitivity criteria of model output $\hat{Y}$ with respect to input factor Z.

The local (first-order) sensitivity coefficient $S_i^{\text{local}}(\mathbf{z}_k)$ is defined as the partial derivative of the output variable $\hat{Y}$ with respect to factor Z_i, calculated at the scenario $\mathbf{z}_k$:

$$S_i^{\text{local}}(\mathbf{z}_k) = \left.\frac{\partial f(\mathbf{Z})}{\partial Z_i}\right|_{\mathbf{z}_k}$$

This criterion is equivalent to the slope of the calculated model output in the parameter space. The $S_i^{\text{local}}(\mathbf{z}_k)$ criterion is an absolute measure of sensitivity, which depends on the scales or measurement units of $\hat{Y}$ or Z_i. A standardized version, called the relative sensitivity, is defined by:

$$S_i^{\text{local}}(\mathbf{z}_k) = \left.\frac{\partial f(\mathbf{Z})}{\partial Z_i}\right|_{\mathbf{z}_k} \times \frac{z_{k,i}}{f(\mathbf{z}_k)}$$

Local sensitivity analysis can be used to study the role of some parameters or input variables in the model. But this method is less useful than global sensitivity analysis when the purpose of the analysis is to study the effect of uncertainty of several factors on model outputs. A more detailed description of local sensitivity analysis is given by Turányi and Rabitz (2000).

A winter wheat dry matter model (continued)

For illustration, the local sensitivity coefficient for parameter E_{b} is defined by

$$S_{Eb}^{\text{local.r}}(\mathbf{z}) = \sum_{t=1}^{t_H-1} z_{E_{\text{imax}}}\left[1 - \mathrm{e}^{-z_K z_{\text{LAI}(t)}}\right] z_{\text{PAR}(t)}.$$

In *global sensitivity analysis* (Fig. 3b–d), on the other hand, the output variability is evaluated when the input factors vary in their whole uncertainty domains. This provides a more realistic view of the model behaviour when used in practice. There are several methods to perform global sensitivity analyses and the whole Section 4.2 is concerned with their description, while the book edited by Saltelli et al. (2000) is a comprehensive reference.

The global degree of association between Z and $\hat{Y}$ over the interval $[z_{\text{min}}, z_{\text{max}}]$ can first be measured through a model approximation. For instance, if the crop model is approximated by a linear relationship between Z and $\hat{Y}$ (Fig. 3c), sensitivity can be measured by the squared regression coefficient or by the linear correlation between Z and $\hat{Y}$. This approach is described in Section 4.2.3. It is a simple and often efficient way to measure sensitivity, provided the model approximation is adequate.

The approaches illustrated in Figures 3b and d are different since they do not rely on a model approximation, in principle at least. They are called model-independent in the sense of Saltelli et al. (1999), because they measure variation of $\hat{Y}$ independently of how this variation is distributed along the Z-axis. The sensitivity criterion illustrated in Figure 3b is simply based on the range of the model output when Z runs within $[z_{\min}, z_{\max}]$ and it will be briefly discussed in Section 4.2.1 on one-at-a-time methods. In the approach illustrated in Figure 3d, sensitivity is measured by the variance of $\hat{Y}$ over $[z_{\min}, z_{\max}]$. This approach will be described in Sections 4.2.2, 4.2.4 and 4.2.5.

4.1.2. Several input factors

Figure 4 presents an *interaction plot* between two input factors Z_1 and Z_2: in this plot, the relationship between input Z_1 and output $\hat{Y}$ is represented for several distinct values of Z_2. The numerical values shown in Figure 4 are also presented in Table 3. If the effects of Z_1 and Z_2 on $\hat{Y}$ were additive, then the curves would be parallel. On the contrary, Figure 4 shows that there are strong interaction effects on $\hat{Y}$ between factors Z_1 and Z_2. The interaction plot shows clearly that, in case of an interaction between Z_1 and Z_2, the sensitivity of $\hat{Y}$ to Z_1 depends on the value of Z_2 and vice-versa. This situation occurs with most crop models, because crop models are not simply additive functions of parameters and input variables.

It is common practice to measure sensitivity for each input factor Z_i separately, with all other factors fixed at their single nominal values. However, this prevents interactions from being detected and quantified, whereas taking interactions into account is a key aspect of most global sensitivity methods. We discuss below the interest of several criteria with respect to their ability to take into account interactions between factors.

Consider for instance, the variance criterion $\mathrm{var}(\hat{Y})$ illustrated in Figure 3d, and suppose now that there are several input factors. Let us denote by $\mathrm{var}(\hat{Y}|Z_j = z_j, j \neq i)$ the

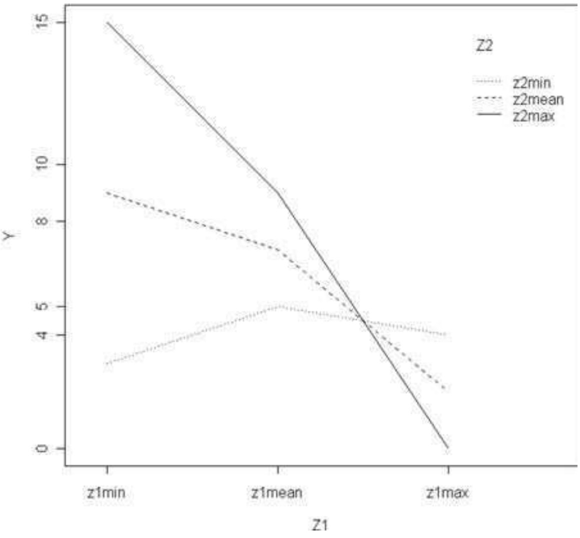


Figure 4. Two-factor interactions graphics: the output $\hat{Y}$ is represented as a function of the input factor Z_1, for three distinct values of Z_2.

Table 3. Output values $\hat{Y}$ for two interacting factors Z_1 and Z_2 and calculation of variance-based criteria for the first factor.

Z_1	Z_2	$\hat{Y}$	$E(\hat{Y}\mid Z_1)$	$\mathrm{var}(\hat{Y}\mid Z_1)$	$\mathrm{var}[E(\hat{Y}\mid Z_1)]$	$E[\mathrm{var}(\hat{Y}\mid Z_1)]$
1	1	3				
1	2	9	9	24		
1	3	15				
2	1	5				
2	2	7	7	8/3	26/3 ≈ 8.67	142/9 ≈ 15.78
2	3	9				
3	1	4				
3	2	2	2	8/3		
3	3	0				

variance of $\hat{Y}$ when z_i varies within its uncertainty domain and all other factors Z_j are fixed at given values z_j. Clearly, this variance gives information on the sensitivity of $\hat{Y}$ with respect to Z_i.

In a strict one-at-a-time approach, the criterion $\mathrm{var}(\hat{Y}|Z_j = z_j, j \neq i)$ is calculated at the nominal values $z_{0,j}$ only: it is equal to $\mathrm{var}(\hat{Y}|Z_j = z_{0,j}, j \neq i)$. If there are interactions between factors, however, then $\mathrm{var}(\hat{Y}|Z_j = z_j, j \neq i)$ depends on the z_j values and a more synthetic sensitivity criterion is preferable.

One possibility consists in using the variance of $\hat{Y}$ averaged over the z_js, rather than calculated at specific values $z_{0,j}$. Thus $\mathrm{var}(\hat{Y}|Z_j = z_{0,j}, j \neq i)$ is replaced by $\mathrm{var}[\mathrm{E}(\hat{Y}|Z_i = z_i)]$, where $\mathrm{E}(\hat{Y}|Z_i = z_i)$ denotes the expected (or average) model output when factor Z_i takes a given value z_i and the other factors vary within their uncertainty domains. The variance calculated in this way was called the *top marginal variance* by Jansen et al. (1994). It corresponds to the *main-effect* in an analysis of variance or to the *first-order* index in some sensitivity analysis methods. When this is applied to the example of Figure 4 (see Table 3), the first-order sensitivity to Z_1 is equal to $26/3 \approx 8.67$.

A second possibility consists in considering the expected value of $\mathrm{var}(\hat{Y}|Z_j = z_j, j \neq i)$ over all possible values of the z_js, for $j \neq i$, $\mathrm{E}[\mathrm{var}(\hat{Y}|Z_j = z_j, j \neq i)]$. The variance calculated in this way was called the *bottom marginal variance* by Jansen et al. (1994). By analogy to definitions given in Saltelli et al. (1999), we call such criteria *total* sensitivity criteria. When this is applied to the example of Figure 4 (see Table 3), the total sensitivity to Z_1 is equal to 142/9 ≈ 15.78.

The total sensitivity of $\hat{Y}$ to Z_i can be interpreted as the expected remaining uncertainty in $\hat{Y}$ if all other input factors were determined exactly. In the example, the total sensitivities of both factors are larger than their main-effect sensitivities. This is a general property, and this difference between the total and main-effect sensitivities is entirely due to interactions between the factors.

Thus, total sensitivity gives a comprehensive measure of the influence of an input factor. This measure can be decomposed into main-effects and interactions, and this decomposition usually gives more insight on the model behaviour. As a conclusion, both types of criteria are useful and complementary. They will be illustrated, with the winter wheat model, in the sections on design of experiments and on sampling-based and variance-based methods.

4.2. Methods of global sensitivity analysis

4.2.1. One-at-a time methods

The most intuitive method to conduct a sensitivity analysis is to vary one factor at a time, while the other factors are fixed at their nominal values. The relationship between the values z_i of factor Z_i and the responses $f(z_{0,1} \ldots z_{0,i-1}, z_i, z_{0,i+1}, \ldots z_{0,s})$ determines a one-at-a-time response profile. Drawing response profiles is often useful, at least in preliminary stages. However, we have already argued that more global methods are preferable, because they take account of and quantify interactions between input factors.

In practice, each input factor Z_i takes k equispaced values from $z_{\min,i}$ to $z_{\max,i}$, with increments $\delta = (z_{\max,i} - z_{\min,i})/(k-1)$. The model responses $f(z_{0,1}, \ldots z_{0,i-1} z_i, z_{0,i+1} \ldots z_{0,s})$ are then calculated for the k discretized values z_i. Figure 5 represents the simulated scenarios when this procedure is applied to three input factors.

If the number of sensitivity factors is not too large, graphical representations are the best way to summarize the response profiles. Alternatively, summary quantities may be calculated for each factor's profile, and compared between factors. Bauer and Hamby (1991), for instance, proposed using the following index

$$I_i^{\mathrm{BH}} = \frac{\max_{z_i} f(z_{0,1} \ldots z_{0,i-1}, z_i, z_{0,i+1} \ldots z_{0,s}) - \min_{z_i} f(z_{0,1} \ldots z_{0,i-1}, z_i, z_{0,i+1} \ldots z_{0,s})}{\max_{z_i} f(z_{0,1} \ldots z_{0,i-1}, z_i, z_{0,i+1} \ldots z_{0,s})}$$

This index can be approximated by the difference between the maximum and minimum simulated values.

The number k of values per profile must be chosen carefully when the model is nonlinear and particularly when it is non-monotonic. Provided k is odd, the number of model simulations to calculate all profiles is equal to $s(k-1)+1$. When k is small and the

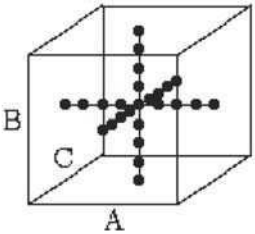


Figure 5. Sampled points when three factors are studied through one-at-a-time sensitivity profiles.

model is non-linear, non-linear effects, as well as maxima or minima, may be undetected, which may lead to under-estimating sensitivity indices such as the index of Bauer and Hamby (1991). When k is large, the computing time may become too large if there are many input factors and the model is complex. In that case, it is more efficient to reserve computing time to more global methods of sensitivity analysis.

A winter wheat dry matter model (continued)

For the winter wheat dry matter model, no highly non-linear phenomena was expected, so that a small number of discretized values was considered sufficient. Besides, between-year variability was expected to have an influence on sensitivity. Consequently, one-at-a-time profiles were calculated with respect to each parameter and for each annual climate series.

The average profiles and their ranges over the climate series are represented in Figure 6. The results show that parameters E_b, A and B have a stronger influence on the simulated biomass value than the other parameters. They show that between-year variability depends on the values of the input factors. However, they give no information on the interactions between parameters.

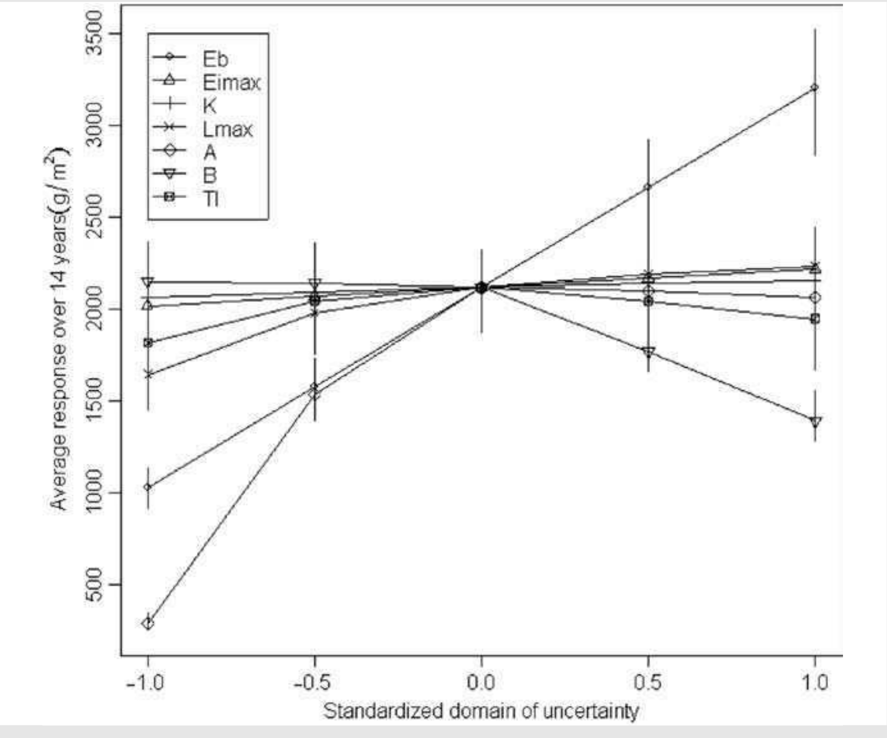


Figure 6. One-at-a-time profiles for the winter wheat dry matter at harvest simulated by the model over 14 climatic series. Points indicate the average simulated values over the climatic series and vertical bars indicate the ranges of variation.

In its most restricted application, one-at-a-time sensitivity analysis is applied at the nominal values of the sensitivity factors only. In that case, it gives information on the model only in a small neighbourhood of the nominal values. However, more global sensitivity analyses may be obtained by calculating one-at-a-time local sensitivity criteria for a lot of different input scenarios. This idea is exploited by Morris (1991). Morris defines the elementary effect of the ith factor for a given scenario $\mathbf{z}_0 = (z_{0,1}, \ldots, z_{0,s})$ as

$$d_i(\mathbf{z}_0) = \frac{f(z_{0,1} \ldots z_{0,i-1}, z_{0,i} + \Delta, z_{0,i+1} \ldots z_{0,s}) - f(z_{0,1} \ldots z_{0,i-1}, z_{0,i}, z_{0,i+1} \ldots z_{0,s})}{\Delta}$$

where $z_{0,i} + \Delta$ is a perturbed value of $z_{0,i}$. The principle of Morris' method is to sample a series of scenarios $\mathbf{z}_0 = (z_{0,1}, \ldots, z_{0,s})$ in the s-dimensional space defined by the values $[z_{\min(i)}, z_{\min(i)} + \delta, z_{\min(i)} + 2\delta, \ldots, z_{\max(i)}]$, $i = 1, \ldots, s$ and to calculate $d_i(\mathbf{z}_0)$ for each sampled value. The resulting distribution of the elementary effects of the ith factor is then characterized by its mean and variance. A high mean indicates a factor with an important influence on the output. A high variance indicates either a factor interacting with another factor or a factor whose effect is non-linear.

4.2.2. Factorial design and analysis of variance

The sensitivity analysis of a crop model is similar to an experiment where nature is being replaced by the simulated crop model. It follows that the classical theory of experimental design provides very useful tools for sensitivity analysis. In particular, factorial designs make it possible to evaluate simultaneously the influence of many factors, with possibly a very limited number of runs. An additional practical advantage is that the methods of analysis are available in general statistical packages.

Despite the analogy between natural experiments and sensitivity analyses, some differences must be pointed out. First, there is nothing like measurement error in simulated experiments, at least when the model is deterministic. As a consequence, there is no residual variance and it is unnecessary to replicate the same scenarios and introduce blocking, whereas replication and blocking are the key components of designed experiments. The second difference is that the number of runs may quite often be much larger in simulation studies than in real experiments.

Many books are dedicated to the design of experiments. A very good reference on factorial designs and response surface methods is Box and Draper (1987).

4.2.2.1. Complete factorial designs

With s input factors and m modalities per factor, there are m^s distinct input scenarios. The (unreplicated) complete m^s factorial design consists of running simulations for each of these scenarios exactly once.

The common point between the complete factorial design and the one-at-a-time profiles is that each factor is studied at a restricted number of levels. However, the major difference is that the emphasis in factorial designs is on making all factors vary simultaneously. This implies that the global "input space" of the model is much better investigated, as can be seen by comparing Figure 7 with Figure 5.

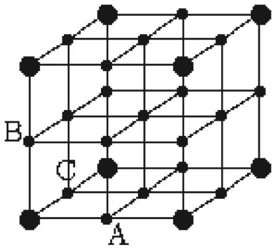


Figure 7. Sampled points when three factors are studied through a complete factorial design with two modalities per factor (large dots) or three modalities per factor (small and large dots).

Table 4. Number of runs for complete m^s factorial designs.

s	$m = 2$	$m = 3$	$m = 4$	$m = 5$
5	32	243	1024	3125
10	1024	59049	1048576	9765625
20	1.05e + 06	3.49e + 09	1.10e + 12	9.54e + 13

A less favourable consequence is that the complete factorial design requires many runs when the number of factors under study is large, as Table 4 shows. For this reason, the 2^s and 3^s factorial designs are the most frequently used complete factorial designs when the number of factors is large. These designs are very useful to quantify interactions between factors.

Factorial decomposition of the model response The analysis of variance (ANOVA) is based on the decomposition of the response variability between contributions from each factor and from interactions between factors. This decomposition is related to the statistical theory of the linear model.

Consider a model with two input factors Z_1 and Z_2, and let $\hat{Y}_{ab} = f(a, b)$ denote the model response when $z_1 = a$ and $z_2 = b$. In a complete m^2 factorial design, there are m possible values for a and m possible values for b and so there are m^2 distinct $\hat{Y}_{ab}$ values. Let $\hat{Y}_{\bullet\bullet}$ denote their general mean, $\hat{Y}_{a\bullet}$ denote the mean when $z_1 = a$, and $\hat{Y}_{\bullet b}$ the mean when $z_2 = b$. Then, when restricted to the m^2 design scenarios, the model can be decomposed into

$$\hat{Y}_{ab} = \mu + \alpha_a + \beta_b + \gamma_{ab}, \tag{1}$$

where $\mu\hat{Y}_{\bullet\bullet}$ is the general mean, $\alpha_a = \hat{Y}_{a\bullet} - \mu$ is called the main effect of factor Z_1 when $z_1 = a$, $\beta_b = \hat{Y}_{\bullet b} - \mu$ is the main effect of factor Z_2 when $z_2 = b$, and $\gamma_{ab} = \hat{Y}_{ab} - (\mu + \alpha_a + \beta_b)$ is the interaction between Z_1 and Z_2 when $z_1 = a$ and $z_2 = b$. The factorial effects satisfy the properties $\sum_a \alpha_a = 0$, $\sum_b \beta_b = 0$, and $\sum_a \gamma_{ab} = \sum_b \gamma_{ab} = 0$. The number of free ANOVA parameters $(\alpha_a, \beta_b, \gamma_{ab})$ associated with each factorial term is called its "degrees of freedom" number. There are $(m - 1)$ degrees of freedom for the main effects of Z_1 and Z_2 and $(m - 1)^2$ degrees of freedom for their interaction.

The response variability can be decomposed into factorial terms as follows:

$$\underbrace{\sum_{ab}(\hat{Y}_{ab}-\mu)^2}_{\mathrm{SS}_T} = \underbrace{m\sum_{a}\alpha_a^2}_{\mathrm{SS}_1} + \underbrace{m\sum_{b}\beta_b^2}_{\mathrm{SS}_2} + \underbrace{\sum_{a,b}\gamma_{ab}^2}_{\mathrm{SS}_{12}}, \tag{2}$$

where SS_T measures the total variability in the model responses, SS_1 is the sum of squares associated with the main effect of Z_1, SS_2 is the sum of squares associated with the main effect of Z_2, and SS_{12} is the sum of squares associated with the interaction between Z_1 and Z_2.

With s factors at m levels, the complete ANOVA decomposition is a sum of $(2^s - 1)$ factorial terms:

$$SS_T = \sum_i SS_i + \sum_{i<j} SS_{ij} + \cdots + SS_{1\ldots s}, \tag{3}$$

including main effects (SS_i) and interactions between up to s factors ($SS_{1\ldots s}$). The number of degrees of freedom for an interaction between q factors is equal to $(m-1)^q$. Note that for a 2^s factorial design, all factorial terms have just one degree of freedom.

ANOVA results and sensitivity indices For the sensitivity analysis of a deterministic model, the main interest lies in comparing the contributions of the factorial terms to the total variability, while formal testing of hypotheses has no real meaning since there is no residual variability. It follows that the most useful information lies in the sums of squares. By dividing the sums of squares by the total variability, the following "ANOVA" sensitivity indices can be easily calculated:

- main effects sensitivity indices $S_1 = \frac{SS_1}{SS_T}$, $S_2 = \frac{SS_2}{SS_T}$;
- interaction sensitivity indices $S_{12} = \frac{SS_{12}}{SS_T}$;
- total sensitivity indices such as $TS_1 = \frac{SS_1+SS_{12}}{SS_T}$ or $TS_2 = \frac{SS_2+SS_{12}}{SS_T}$, which summarize all factorial terms related to a particular factor.

A winter wheat dry matter model (continued)

For the eight sensitivity factors of the "winter wheat dry matter model" example, a 2^8 complete factorial design would require the number of climatic series to be limited to two. We applied instead a $2^7 \times 14$ complete factorial design, where the 14 climatic series were crossed with the 2^7 scenarios based on the minimum and maximum values of the parameter uncertainty ranges. There were thus a total of $1792 = 2^7 \times 14$ simulations of the crop model. The analysis of variance on the simulation results can be performed assuming the complete factorial model, including interactions between up to eight factors. In practice, simpler models are often sufficient to capture the most interesting sensitivity features.

For illustration, the results presented in Table 5 were calculated assuming a model with 8 main effects (7 parameters + climate) and interactions between two factors only. The sums of squares in Table 5 are given by many statistical software packages, but not the sensitivity column which was calculated by dividing the sum-of-squares column by the total variability SS_T of the data. In this analysis, the quantities associated with the residuals correspond to all terms which were not included in the model, that is here, interactions between three or more factors. The coefficient of determination R^2 of a model is, by definition, the percentage of the total variability explained by the model. Here, it is equal to 0.94, indicating that only 6% of the variability in the simulated model output is accounted for by interactions between more than three factors.

Sensitivities are represented graphically in Figures 8 (factorial indices) and 9 (total indices). For these figures, the complete factorial model was used, but the differences with the model with main effects and two-factor interactions were small. The most influential factors are the parameters E_b, A and B, which confirms results of the one-at-a-time profiles (Figure 6). The figures also show that the influence of interactions is high, which could not be detected by the one-at-a-time profiles.

Table 5. Analysis of variance table of the complete factorial design applied to the winter wheat dry matter model; the table was calculated for the model including main effects and two-factor interactions. Sensitivities smaller than 0.01 are not displayed.

	SS	Sensitivity index
E_b	777 593 320	0.33
E_{imax}	6686 674	
K	3662 758	
L_{max}	80 732 881	0.03
A	520 104 586	0.22
B	309 742 948	0.13
TI	551 495	
$YEAR$	7330 246	
$E_b{:}E_{imax}$	1763 250	
$E_b{:}K$	965 855	
$E_b{:}L_{max}$	21 288 948	
$E_b{:}A$	137 149 566	0.06
$E_b{:}B$	81 678 016	0.04
$E_b{:}TI$	145 427	
$E_b{:}YEAR$	1932 958	
$E_{imax}{:}K$	8306	
$E_{imax}{:}L_{max}$	183 068	
$E_{imax}{:}A$	1179 375	
$E_{imax}{:}B$	702 365	
$E_{imax}{:}TI$	1251	
$E_{imax}{:}YEAR$	16 622	
$K{:}L_{max}$	823 631	
$K{:}A$	82 704	
$K{:}B$	635 271	

continued

Table 5.—Cont'd.

	SS	Sensitivity index
K:TI	395	
K:*YEAR*	6643	
$L_{\max}$:A	60 448	
$L_{\max}$:B	17 116 469	
$L_{\max}$:TI	35 467	
$L_{\max}$:*YEAR*	145 584	
A:B	193 147 537	0.08
A:TI	28 101 635	0.01
A:*YEAR*	2586 798	
B:TI	1425 195	
B:*YEAR*	1178 019	
TI:*YEAR*	2471 694	
Residuals	128 829 265	

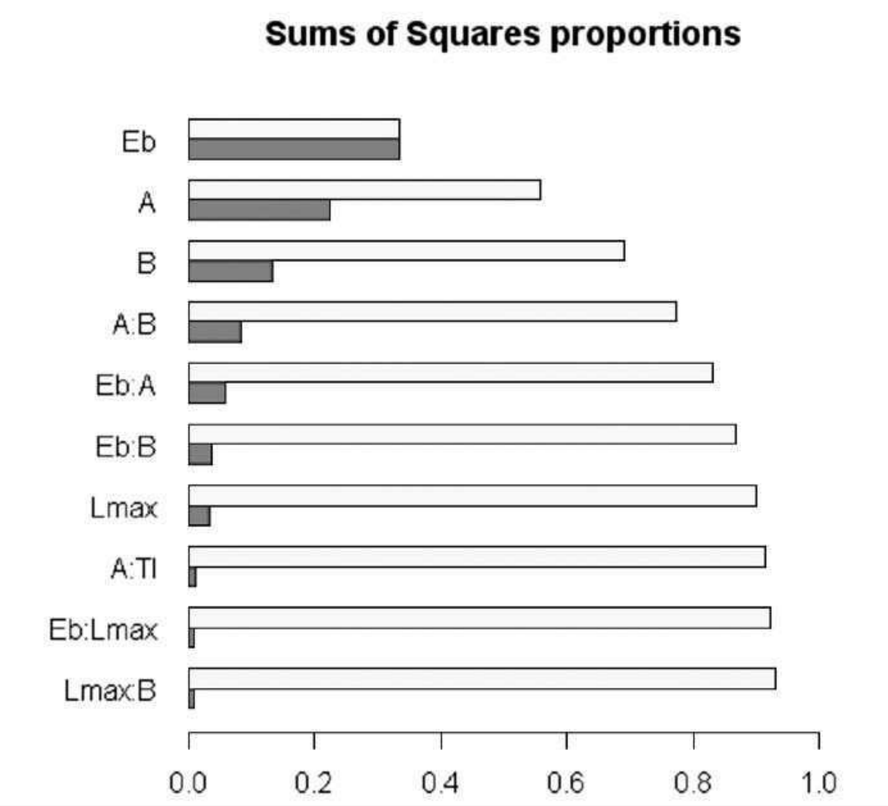


Figure 8. The eight largest factorial sensitivity indices based on the $2^s \times 14$ factorial design and its analysis of variance with a complete factorial model, for the winter wheat crop model; the upper bars show cumulative indices.

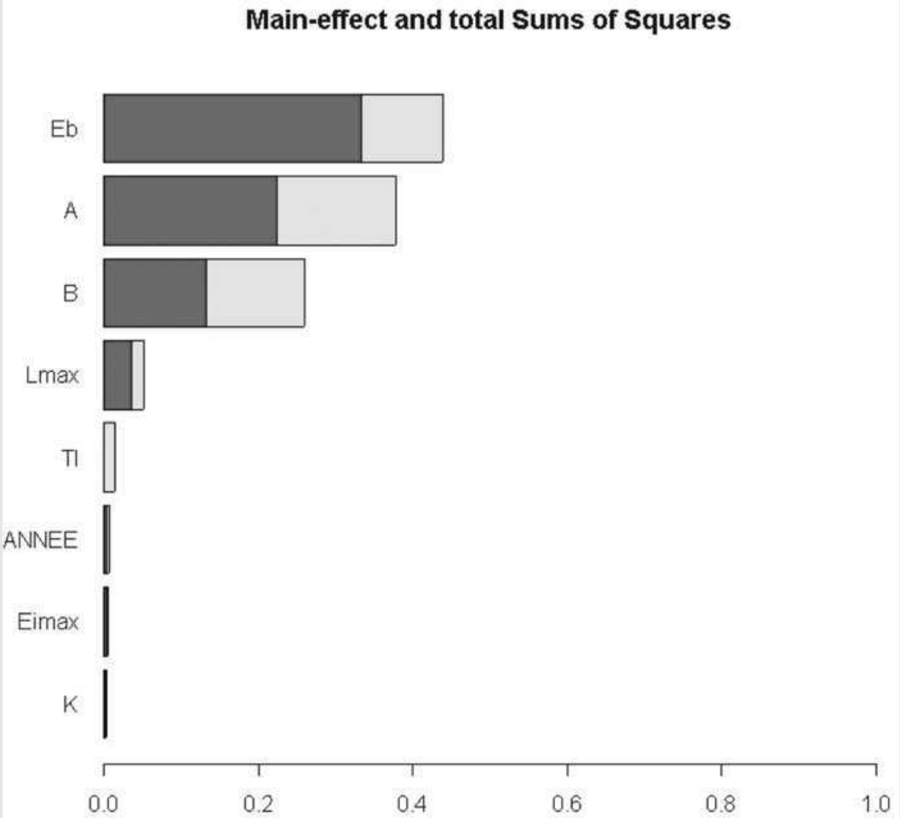


Figure 9. Main-effect (first part of the bars) and total (full bars) sensitivity indices based on the $2^s \times 14$ factorial design and its analysis of variance, for the winter wheat model.

4.2.2.2. Fractional factorial designs

When there is a large number of factors, the factorial main effects and low-order interactions can usually be estimated quite accurately by running only a fraction of the complete factorial design. When applied to sensitivity analysis, fractional factorial designs are very useful for screening a large number of factors and for detecting the most influential ones, with a relatively small number of runs. This requires, however, the assumption that higher-order interactions are negligible. It also requires that the fraction be carefully chosen. This can be done through algebraic methods of construction (see Box and Draper, 1987; Kobilinsky, 1997).

Consider, for example, the fractional design for seven factors at two modalities ±1 given in Table 6. This is a complete factorial design for factors A, B and C, which are called the basic factors of the fraction. The modalities of the basic factors have been used to calculate those of four additional factors D, E, F and G.

Consider first the design restricted to factors A, B, C and G. This is a half-fraction of the 2^4 complete factorial design, with eight runs instead of 16. Because this is an

Table 6. Complete 2^3 factorial design and fractional design defined by $ABC = 1$.

A	B	C	$D = AB$	$E = AC$	$F = BC$	$G = ABC$	Y
-1	-1	-1	$+1$	$+1$	$+1$	-1	Y_1
-1	-1	$+1$	$+1$	-1	-1	$+1$	Y_2
-1	$+1$	-1	-1	$+1$	-1	$+1$	Y_3
-1	$+1$	$+1$	-1	-1	$+1$	-1	Y_4
$+1$	-1	-1	-1	-1	$+1$	$+1$	Y_5
$+1$	-1	$+1$	-1	$+1$	-1	-1	Y_6
$+1$	$+1$	-1	$+1$	-1	-1	-1	Y_7
$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	Y_8

incomplete factorial design, not all factorial terms can be estimated. However, there are quite simple rules to determine which terms can be estimated. Thus, the relationship $G = ABC$ which was used for defining G implies that the main-effect of G is confounded with the $A{:}B{:}C$ interaction. The two effects cannot be estimated separately, but if $A{:}B{:}C$ is assumed to be negligible, then the main-effect of G can be estimated. By multiplying both sides of the $G = ABC$ equality with factor letters and by adopting the convention that $A^2 = B^2 = C^2 = G^2 = 1$, other confounding rules can be obtained. For example, multiplying by A yields $AG = A^2BC$ which gives $AG = BC$ after simplification. This implies that the interactions $A{:}G$ and $B{:}C$ are confounded. More generally, there is one confounding relationship associated with each factorial effect between the basic factors. Here, the confounding relationships are:

$$\begin{aligned}
1 &= ABCG\\
A &= BCG\\
B &= ACG\\
C &= ABG\\
AB &= CG\\
AC &= BG\\
BC &= AG\\
ABC &= G,
\end{aligned}$$

where 1 indicates the general mean. The resolution of a fractional design is, by definition, the minimum order among the interactions confounded with the general mean. Here, there is only one fourth-order interaction confounded with the mean, so the fraction has resolution IV (by convention, the resolution is often written with roman numbers). With a fraction of resolution IV, the general mean is confounded with the four-factor interaction, main effects are confounded with three-factor interactions, and two-factor interactions are mutually confounded. Assuming that three- and four-factor interactions are negligible, all main effects can be estimated.

Consider now the design with the seven factors A to G. This is a $1/(2^4)$ fraction of the complete factorial design, with 8 runs instead of 128! Now, instead of being confounded by pairs, factorial effects are confounded by groups of size 16. For example, the confounding relationships involving the general mean are

$$\begin{aligned} 1 &= ABD = ACE = BCF = ABCG = BCDE = ACDF = CDG = ABEF \\ &= BEG = AFG = DEF = ADEG = BDFG = CEFG = ABCDEFG \end{aligned}$$

and the confounding relationships involving the main effect of A are

$$\begin{aligned} A &= BD = CE = ABCF = BCG = ABCDE = CDF = ACDG = BEF \\ &= ABEG = FG = ADEF = DEG = ABDFG = ACEFG = BCDEFG. \end{aligned}$$

This is a resolution III fractional design, and with such a design, the main effects are confounded with interactions between two and more factors. Thus the main effects can be estimated provided all interactions are considered negligible.

Both examples given above can be generalized to more factors and more modalities per factors. With 2^n simulations, it is possible to study up to $2^n - 1$ factors in a resolution III fraction, and up to 2^{n-1} factors in a resolution IV design. Of course, higher resolutions should be preferred when possible. The main difficulty is to find the most appropriate confounding relationships when defining new factors from the basic ones. Tables are given in Kobilinsky (1997). The PROC FACTEX procedure of the SAS QC© module can generate fractional designs automatically.

4.2.2.3. Other experimental designs

In a 2^s complete or fractional factorial design, all information on each quantitative factor Z_i is based on the model behaviour at only two levels per factor. This is optimal when, for any setting of the other factors, the model is a linear or near-linear function of z_i. It often remains efficient when the model is monotonous. However, 2^s designs do not allow one to detect and quantify non-linear relationships between a sensitivity factor and the output.

In that case, it is necessary to consider designs with more levels per factor. One may use 3^s, 4^s complete or fractional designs, which ensure that quadratic effects may be detected as well as linear effects. Flexible fractional designs exist also for these designs, in fact for all m^s designs where m is a prime number or a power of a prime.

The response surface methodology (see Box and Draper, 1987) offers an alternative approach to study the influence of quantitative factors on a response function. It is based on an approximation of the crop model by a polynomial function of degree one or two of the input factors, and on convenient designs to estimate their parameters. This approach has been applied to the STICS crop model (Ruget et al., 2002) and we refer to this article for a detailed presentation in the context of crop models.

4.2.3. Intensive sampling and correlation criteria

In the sensitivity analysis methods presented in Section 4.2.2, the sampled modalities of the input factors are precisely defined by the factorial design. Another approach consists in randomly generating factor values by Monte Carlo sampling. The principle is to randomly generate N scenarios of the input factors $\mathbf{z}_k = (z_{k,1}, \ldots, z_{k,i}, \ldots, z_{k,s})$ $k = 1, \ldots, N$, and to compute the model output for each scenario, $f(\mathbf{z}_k)$ $k = 1, \ldots, N$, in a similar way to what is done for an uncertainty analysis. The statistical methods related to regression (see e.g. Venables and Ripley, 1999) are then used to represent and to measure the sensitivity of the output variables with respect to the input factors. These methods are presented below.

Correlation coefficients can be used to quantify the relationships between input factors and output variables. Let $s_{\hat{Y}}^2 = \frac{1}{N}\sum_{k=1}^{N}[f(\mathbf{z}_k) - \bar{f}]^2$ and $s_{Z_i}^2 = \frac{1}{N}\sum_{k=1}^{N}(z_{k,i} - \bar{z}_i)^2$ denote the empirical variances of $\hat{Y} = f(\mathbf{Z})$ and Z_i in the simulations, and let $\hat{\text{cov}}(\hat{Y}, Z_i) = \frac{1}{N}\sum_{k=1}^{N}[f(\mathbf{z}_k) - \bar{f}][z_{k,i} - \bar{z}_i]$ denote their covariance. Then the PEAR (Pearson Product Moment Correlation Coefficient) coefficient between Z_i and $\hat{Y}$ is defined by

$$r_{Z_{i,\hat{Y}}} = \frac{\hat{\text{cov}}(\hat{Y}, Z_i)}{s_{\hat{Y}} s_{Z_i}}.$$

It varies between -1 and $+1$ and it measures the degree of linear association between the variations of Z_i and those of $\hat{Y}$. Some non-linear associations may remain undetected and underestimated by the PEAR coefficient. An alternative is the Spearman correlation coefficient, which is calculated on the ranks of Z_i and Y. The Spearman correlation coefficient is more adequate in case of strongly non-linear, but still monotonous, relationships.

With the PEAR or Spearman coefficients, no account is taken of the possible effects of input factors other than Z_i. In contrast, the partial correlation coefficient (PCC) aims at measuring the association between Z_i and $\hat{Y}$ after eliminating possible effects due to other input factors Z_j, $j \neq i$. The PCC coefficient is similar to the PEAR correlation coefficient, but it is calculated with $f(\mathbf{z}_k)$ and $z_{k,i}$ replaced by the residuals of the following two regression models

$$f(\mathbf{z}_k) = b_0 + \sum_{j \neq i} b_j z_{k,j} + \varepsilon_k, \quad z_{k,i} = c_0 + \sum_{j \neq i} c_j z_{k,j} + \varepsilon'_k,$$

where b_js and c_js are regression coefficients to be estimated.

Regression models give a general framework for studying the influence of all input factors simultaneously. By approximating the crop model under study, they make it possible to evaluate the influence of each input factor. Consider for instance the regression model with first-order effects only:

$$f(\mathbf{z}_k) = b_0 + \sum_{i=1}^{s} b_i z_{k,i} + \varepsilon''_{ik}, \tag{4}$$

where b_i are the regression coefficients to be estimated and ε''_{ik} is the approximation error term. The regression coefficients are estimated by least-squares. The quality of the adjustment is synthesized typically by calculating the model coefficient of determination R^2, that is, the percentage of output variability explained by the model.

The estimated regression coefficients $\hat{b}_i$ can be considered as sensitivity measures associated with the factors Z_i, provided they are standardized with respect to the variability in $\hat{Y}$ and in Z_i. The standardized regression coefficients (SRC) are defined as the quantities $\hat{b}_i(s_{Z_i}/s_{\hat{Y}})$.

Many more principles and techniques of regression are useful for sensitivity or uncertainty analysis, but it is out of the scope of this chapter to present them all. However, a few remarks can be made:

- the regression model in Eq. (4) can be extended in order to incorporate interactions between input variables, qualitative as well as quantitative factors, quadratic as well as linear effects. This is useful in particular if the regression coefficient of determination is small;
- when the number of terms in the model is large, model selection techniques (stepwise regression for instance) may become a precious aid to interpretation, since they can eliminate factors with negligible influence;
- the regression techniques presented here are good essentially at capturing linear effects between the Z_is and the Ys. Alternative methods should be considered when non-linear relationships are suspected;
- polynomial regression is one of the basic approaches in response surface methodology. It can be used on randomly selected simulations as described here, but also on simulations based on factorial or response surface design (Ruget et al., 2002).

A winter wheat dry matter model (continued)

$N = 5000$ scenarios were generated, using the generators of quasi-random numbers implemented in the *R* software (www.r-project.org) for Uniform and Beta distributions. Figure 10 shows scatterplots of the model simulations. A scatterplot is a representation of the points $\left[z_{k,i}, f(\mathbf{z}_k)\right]$, where $z_{k,i}$ is the value of Z_i in the k-th simulation and $f(\mathbf{z}_k)$ is the simulated response. In order to get a better visualisation, only 500 points have been represented in the plots of Figure 10. Non-parametric smoothing lines, based on local regressions, have been added to the plots in order to better visualize the relationship between $f(\mathbf{z}_k)$ and $z_{k,i}$. Figure 10 reveals a negative correlation between biomass at harvest and parameter B, and a positive correlation between the model output and parameters E_{b} and A.

PEAR and SRC coefficients for the parameters of the winter wheat dry matter model are given in Table 7. They have been calculated with the linear model function of the statistical package R, from the 5000 simulations. The results are very similar to those obtained with analysis of variance, with E_{b}, A and B the most influential parameters. The difference between the SRC and PEAR coefficients is small because the data set is large (5000 samples) and so the input factors are nearly orthogonal (the maximum empirical correlation between input factors is 0.037). There is a larger difference between the input sampling distributions, with a stronger sensitivity to E_{b} when the beta distribution is used.

The coefficient of determination of the model with only first-order effects (uniform case) is $R^2 = 0.78$. This shows that interactions account for more than 20% of the output variability.

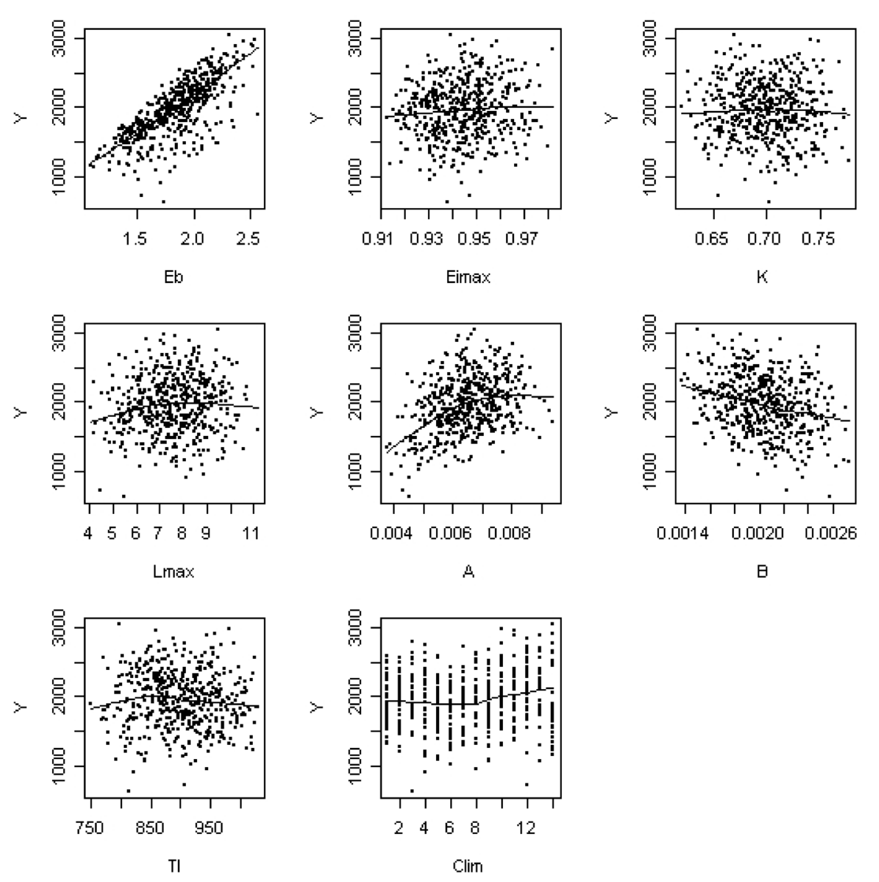


Figure 10. Scatter plots between the simulated values of biomass at harvest (g/m^2) and each input factor over its range of uncertainty, based on 500 simulations.

Table 7. PEAR and SRC coefficients for the winter wheat dry matter model, estimated from 5000 Monte Carlo samples.

Parameter	Uniform sampling		Beta sampling	
	PEAR	SRC	PEAR	SRC
E_b	0.62	0.63	0.71	0.73
E_{imax}	0.10	0.06	0.04	0.06
K	0.04	0.03	0.04	0.03
L_{max}	0.15	0.17	0.15	0.16
A	0.47	0.49	0.36	0.39
B	−0.33	−0.34	−0.30	−0.32
TI	0.04	0.03	0.04	0.04

4.2.4. Intensive sampling and variance-based sensitivity analysis

4.2.4.1. Variance-based measures of sensitivity

In the approaches based on experimental design followed by analysis of variance or on Monte Carlo sampling followed by regression, sensitivity analysis is based on an approximation of the crop model by a simpler linear model. In the variance-based methods described in this section, the principle is to decompose the output variability $D = \mathrm{Var}(\hat{Y})$ globally, without an intermediate simplified model.

Sobol decomposition of the model The methods are based on model and variance decompositions that are very similar to those encountered in analysis of variance. To emphasize the similarities and differences, we adopt a presentation which parallels that in Section 4.2.2.

Consider two quantitative input factors Z_1 and Z_2, and let $\hat{Y}_{ab}$ denote the model response when $z_1 = a$ and $z_2 = b$. The Sobol decomposition of the crop model f (Sobol, 1993) is given by

$$\hat{Y}_{ab} = \mu + f_1(a) + f_2(b) + f_{12}(a, b). \tag{5}$$

This decomposition is quite similar to the decomposition in Eq. (1), but, in contrast to Section 4.2.2, a and b are now assumed to vary continuously within the uncertainty interval [0, 1]. It follows that the general mean of the crop model f is now defined by

$$\mu = \int_0^1 \int_0^1 f(z_1, z_2) dz_1 dz_2.$$

The main effect of factor Z_1 is defined by the function

$$f_1(a) = \int_0^1 f(a, z_2) dz_2 - \mu$$

of a. Similarly, the main effect of B is defined by

$$f_2(b) = \int_0^1 f(z_1, b) dz_1 - \mu.$$

Finally, the interaction between Z_1 and Z_2 is defined by

$$f_{12}(a, b) = f(a, b) - f_1(a) - f_2(b) + \mu.$$

The factorial effects thus defined satisfy orthogonality properties which make the decomposition unique and give it a lot of nice properties. In particular, these properties yield an orthogonal decomposition of the response variability into factorial terms:

$$\underbrace{\int_0^1\int_0^1 (\hat{Y}_{z_1 z_2}-\mu)^2 dz_1 dz_2}_{\mathrm{Var}(\hat{Y})} = \underbrace{\int_0^1 f_1(a)^2 da}_{D_1} + \underbrace{\int_0^1 f_2(b)^2 db}_{D_2} + \underbrace{\int_0^1\int_0^1 f_{12}(a,b)^2 da db}_{D_{12}}, \tag{6}$$

where D_1 is the variability associated with the main effect of Z_1, D_2 is the variability associated with the main effect of Z_2 and D_{12} is the variability associated with the interaction between Z_1 and Z_2.

With s quantitative factors, the decomposition of the variance $\mathrm{Var}(\hat{Y})$ generalizes to:

$$\mathrm{var}(\hat{Y}) = \sum_{i=1}^{s} D_i + \sum_{i<j} D_{ij} + \cdots + D_{1\ldots s}. \tag{7}$$

In the decomposition Eq. (7), D_i corresponds to the main or first-order effect of Z_i denoted by $\mathrm{var}[\mathrm{E}(\hat{Y}|Z_i = z_i)]$ in Section 4.2.2. The terms $D_{ij}, \ldots, D_{1\ldots s}$ of Eq. (7) correspond to the interactions between the input factors. This is very similar to the analysis of variance. However, $\mathrm{var}(\hat{Y})$ now represents the variability of $\hat{Y}$ with respect to the overall uncertainty in the input factors, and not only over a limited number of experimental design points. This makes it more adequate for taking account of irregular and non-linear effects.

In probabilistic terms, D_i is the variance of the conditional expectation $\mathrm{E}(\hat{Y}\,|Z_i = z_i)$. If $\hat{Y}$ is sensitive to Z_i, $\mathrm{E}(\hat{Y}\,|Z_i = z_i)$ is likely to vary a lot when z_i changes and so D_i is likely to be large. This is why D_i is also called an "importance measure" in the vocabulary of sensitivity analysis.

Sensitivity indices Sensitivity indices are derived from the decomposition Eq. (7) by dividing the importance measures by $\mathrm{var}(\hat{Y})$:

$$S_i = D_i/\mathrm{var}(\hat{Y})$$

$$S_{ij} = D_{ij}/\mathrm{var}(\hat{Y})$$

$$\ldots$$

Consequently, the sensitivity indices satisfy

$$S_1 + \cdots + S_s + S_{1,2} + \cdots + S_{1,2,\ldots s} = 1$$

and can be interpreted as the proportions of $\mathrm{var}(\hat{Y})$ explained by the various factorial terms.

As explained in Section 4.1.2, two main types of sensitivity indices can be defined for each factor Z_i. The first-order sensitivity index S_i is useful for measuring the average influence of factor Z_i on the model output, but it takes no account of the interaction effects involving Z_i. The second useful index is the total sensitivity index of Z_i, equal to the sum of all factorial indices involving Z_i:

$$TS_i = S_i + \sum_{j \neq i} S_{ij} + \cdots + S_{1 \ldots s}.$$

Note that TS_i is also equal to $1 - S_{-i}$, where S_{-i} denotes the sum of all indices where Z_i is not involved.

4.2.4.2. Estimation based on Monte Carlo sampling

In order to estimate the first-order sensitivity index S_i, the basic idea is to evaluate the model response at N randomly sampled pairs of scenarios $sc_{A,k}$ and $sc_{B,k}$ defined by

$$\left.\begin{array}{l} \mathbf{z}_{A,k} = (z_{k,1}, \ldots, z_{k,i-1}, z_{k,i}, z_{k,i+1}, \ldots, z_{k,s}) \\ \mathbf{z}_{B,k} = (z'_{k,1}, \ldots, z'_{k,i-1}, z_{k,i}, z'_{k,i+1}, \ldots, z'_{k,s}) \end{array}\right\}, \quad k = 1, \ldots, N$$

with the same level $z_{k,i}$ of Z_i and all other levels sampled independently. Let D denote $\text{var}(\hat{Y})$, then

$$\hat{f}_0 = \frac{1}{2N} \sum_{k=1}^{N} [f(\mathbf{z}_{A,k}) + f(\mathbf{z}_{B,k})]$$

$$\hat{D} = \frac{1}{2N} \sum_{k=1}^{N} [f(\mathbf{z}_{A,k})^2 + f(\mathbf{z}_{B,k})^2] - \hat{f}_0^2$$

$$\hat{D}_i = \frac{1}{N} \sum_{k=1}^{N} f(\mathbf{z}_{A,k}) \cdot f(\mathbf{z}_{B,k}) - \hat{f}_0^2$$

are unbiased estimators of, respectively, the average value of $\hat{Y}$, its total variance, and the main-effect of Z_i. An obvious estimator of S_i is then $\hat{S}_i = \hat{D}_i / \hat{D}$.

The procedure just described requires $2N$ model simulations for the estimation of each first-order index. When the first-order indices of all s factors must be calculated, the following procedure is more efficient computationally than performing s independent sets of $2N$ simulations:

- generate N input scenarios by Monte Carlo sampling, and store them in a $N \times s$ matrix M; the rows in M will form the $\mathbf{z}_{A,k}$ scenarios for all factors;
- generate N more input scenarios by Monte Carlo sampling, and store them in a $N \times s$ matrix M'; the rows in M' will be used to form the $\mathbf{z}_{B,k}$ scenarios;
- calculate the responses $f(\mathbf{z}_{A,k})$ for each scenario in M;
- for each factor Z_i calculate the responses $f(\mathbf{z}_{B,k})$ where $\mathbf{z}_{B,k}$ is determined by row k of M' for all factors different from Z_i and by row k of M for factor Z_i;
- apply the formulae given above for the calculation of $\hat{S}_i$.

This algorithm requires $N(s+1)$ model simulations for the calculation of the first-order sensitivity indices of s factors. An even more efficient sampling scheme, the winding stairs, was proposed by Jansen (1994). It is not described here for the sake of brevity.

A winter wheat dry matter model (continued)

Figure 11 displays results of a sampling-based sensitivity analysis. A Monte Carlo sample of size 1000 was used to generate a winding stairs set of simulations. Because there were eight factors (seven parameters + climate) in the model and we chose a basis of 1000 Monte Carlo samples, the number of model simulations needed to estimate first-order and total indices was equal to 9 × 1000. In order to show the variability of the estimates due to sampling,

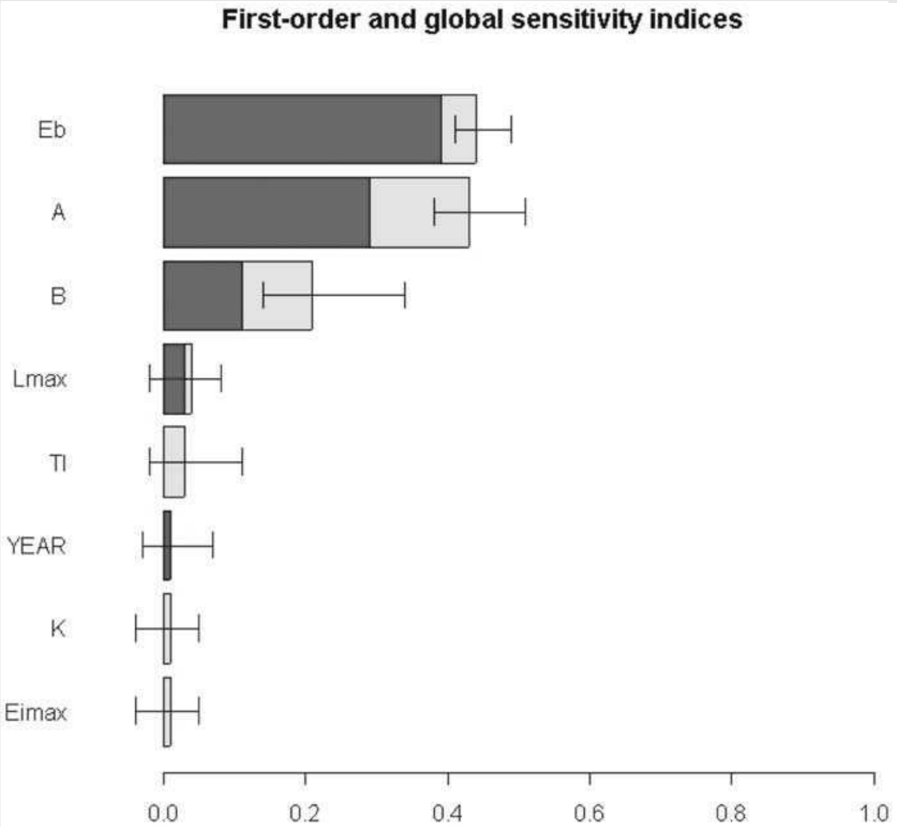


Figure 11. First-order and total Sobol sensitivity indices estimated from Latin hypercube sampling combined with winding stairs; there were 20 runs with 9 × 1000 model simulations for each run; the first part of the bars corresponds to the average (over the 20 runs) estimate of the first-order index, the full bars indicate average estimates of total indices, while the lines indicate extreme estimates of total indices.

this procedure was repeated 20 times, and the ranges of the estimates over the 20 series of simulations are displayed.

The results are different but quite consistent with those obtained with a designed experiments. This is not very surprising because the model behaves quite linearly and so the more intensive sampling-based method does not bring much more information on the model behaviour.

The sampling-based methods give unbiased estimates of the sensitivity indices, but the estimates can be quite variable and even take negative values, as Figure 11 shows. Homma and Saltelli (1996) propose a corrective term to improve this problem. Nevertheless, it remains important to evaluate the precision of the sensitivity indices by repeating the procedure a few times as we did.

The same principle can be generalized to the estimation of second-order or higher effects and to the estimation of total sensitivity indices. For estimating the interaction sensitivity S_{ij}, for instance, the model responses have to be calculated for pairs of scenarios $\mathbf{z}_{A,k}$ and $\mathbf{z}_{B,k}$ with the same levels of Z_i and Z_j. For estimating total sensitivity, the model responses have to be calculated for pairs of scenarios $\mathbf{z}_{A,k}$ and $\mathbf{z}_{B,k}$ with the same levels of all factors except Z_i. This allows the sensitivity index S_{-i} to be estimated, and TS_i is then estimated by $\hat{TS}_i = 1 - \hat{S}_{-i}$.

4.2.5. FAST method for sampling and estimating variance-based criteria

The Fourier amplitude sensitivity test (FAST) is another method for estimating variance-based measures of sensitivity. It is inspired by the Fourier decomposition of a time series in signal theory and was developed initially for analysing the sensitivity of chemical reaction systems to rate coefficients (Cukier et al., 1973, 1975; Schaibly and Shuler, 1973). Recently, its use has been generalized to many domains of applications and new developments have been proposed. The presentation below is limited to the main principles. More details can be found in Chan et al. (2000).

4.2.5.1. FAST sampling

In the FAST method, all input factors are assumed to be quantitative and coded so that their domain of variation is $[0, 1]$. Then the possible scenarios belong to the multidimensional input space $[0, 1]^s$. With Monte Carlo sampling, the simulated scenarios are selected at random within $[0, 1]^s$. With the FAST method, they are selected systematically (or almost systematically) along a search trajectory which is specifically designed to explore efficiently the input space. This is illustrated, in the simple case of two factors, in Figure 12. Figure 12a shows a set of $N = 100$ scenarios sampled according to the FAST principles. These scenarios were generated by regular sampling along the curve visible in Figure 12b.

In the design of a FAST sampling scheme, an integer ω_i is associated with each input factor Z_i. This integer is called the frequency of Z_i and its choice will be explained below. The levels of the input factors Z_i for the simulated scenarios $\mathbf{z}_k (k = 1, \ldots, N)$,

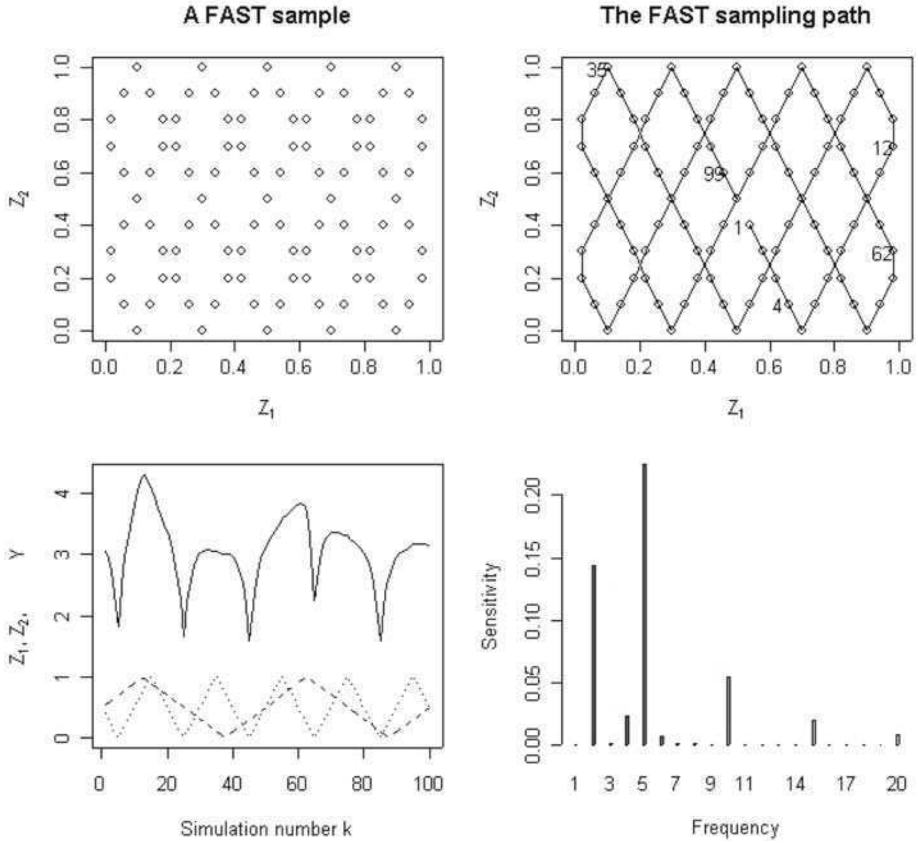


Figure 12. Illustration of FAST principles for two input factors Z_1 and Z_2. (a) Samples of (Z_1, Z_2) values with $\omega_1 = 2$, $\omega_2 = 5$, $\phi_1 = \phi_2 = 0$ and $N = 100$. (b) FAST sampling path indicating the order of the generated scenarios (the numbers 1, 4 ... indicate the first, fourth ... generated scenarios). (c) Values of Z_1, Z_2, and of a response $\hat{Y}$ for N scenarios in ascending order of simulation. (d) Sensitivity indices obtained for several frequencies.

are given by

$$z_{k,i} = G(\sin(\omega_i u_k + \phi_i)),$$

where the scalars

$$u_k = -\pi + \frac{2k-1}{N}\pi$$

form a regularly-spaced sample of the interval $(-\pi, +\pi)$ and can be interpreted as coordinates on the search curve; $G(u)$ is a transformation function from $[-1, 1]$ to $[0, 1]$; and the

ϕ_is are optional random phase-shift parameter taking values in $[0, 2\pi)$. The transformation function

$$G(u) = \frac{1}{2} + \frac{1}{\pi}\arcsin(u),$$

proposed by Saltelli et al. (1999), ensures that the levels of each factor are uniformly, or almost uniformly, sampled.

In Figure 12, $\omega_1 = 2$, $\omega_2 = 5$ and $\phi_1 = \phi_2 = 0$. As can be verified in Figure 12b, each ω_i corresponds to the frequency with which the curve comes back to its starting value for the levels of factor Z_i. Figure 12b shows that the sampling path goes through each value of Z_1 exactly twice (in a given direction). Similarly, the sampling path goes through each value of Z_2 exactly five times.

4.2.5.2. Principles of FAST sensitivity estimation

The principle of FAST is that, if the response $\hat{Y}$ is sensitive to a given factor Z_i, then $\hat{Y}$ and Z_i should vary simultaneously over the scenario index k. In Figure 12c, the variations of Z_1, Z_2, and a putative response $\hat{Y} = f(Z_1, Z_2)$ are displayed as a function of k. This figure shows that the oscillations of $\hat{Y}$ and those of the factor Z_2 are quite simultaneous. This result indicates that Z_2 has a strong influence on the response. With the FAST method, the sensitivity of the output to the factors is quantified by estimating a sensitivity index for a series of frequency (Figure 12d). If the factor Z_i has a strong influence on the model output, the index takes high values for $\omega = \omega_i$ and for its higher harmonics $(2\omega_i, 3\omega_i, \ldots)$. Figure 12d shows that the sensitivity index is higher for $\omega = \omega_2 = 5$ than for $\omega = \omega_1 = 2$. This result reveals that the model output is more sensitive to Z_2 than to Z_1.

4.2.5.3. Spectral decomposition of $f(z_{k,1}, z_{k,2})$ variability

The variability of $f(z_{k,1}, z_{k,2})$ is decomposed into components associated with each frequency ω from 1 to $N - 1$, defined by:

$$D_{[\omega]} = A_\omega^2 + B_\omega^2,$$

where

$$A_\omega = \frac{1}{2\pi} \sum_{k=1}^{N} f(z_{k,1}, z_{k,2})\cos(\omega u_k)$$

$$B_\omega = \frac{1}{2\pi} \sum_{k=1}^{N} f(z_{k,1}, z_{k,2})\sin(\omega u_k).$$

The scalar $D_{[\omega]}$ is called the spectral component of $\hat{Y}$ at frequency ω, while A_ω and B_ω are called the Fourier coefficients of $\hat{Y}$ at frequency ω. They are theoretically defined as integrals over $[-\pi, +\pi]$, but they are shown here in the discrete summation form imposed by the finite number of simulations. The scalar $S_{[\omega]} = D_{[\omega]}/\left(\sum D_{[\omega]}\right)$ can then be considered as the proportion of variability of $f(z_{k,1}, z_{k,2})$ associated with frequency ω. The values of $S_{[\omega]}$ are presented in Figure 12d.

4.2.5.4. The classical FAST method

The original FAST (Cukier et al., 1973, 1975; Schaibly and Shuler, 1973) is a method for estimating essentially the first-order sensitivity indices (or main effects) of the factors Z_i. The frequencies of the different factors are chosen so that the spectral components $D_{[\omega]}$ of $\hat{Y}$ at frequency ω_i and at its first higher harmonics depend on the effects of input factor Z_i only. It follows that the sensitivity index of Z_i can be estimated by

$$S_i = \sum_{p=1}^{M} S_{[p\omega_i]},$$

where M is the number of harmonics taken into account and is usually set to $M = 4$.

Adequate sets of frequencies have been proposed by Cukier et al. (1973) for up to 19 factors. In FAST, there is a minimum number of simulations which is equal to $2M\max(\omega_i) + 1$. For example, when there are $s = 8$ factors, the frequencies given by Cukier et al. (1973) are $23, 55, 77, 97, 107, 113, 121, 125$, and so the minimum number of simulations is equal to $8 \times 125 + 1 = 1001$.

4.2.5.5. The extended FAST method

The extended FAST method (Saltelli et al., 1999) allows the estimation of the first-order *and* the total sensitivity indices. In a simulation study on a crop model, it appeared more efficient than the Monte Carlo approach to estimate first and total sensitivity indices (Makowski et al., 2004). As opposed to the classical FAST, it requires separate sets of simulations for each input factor Z_i of interest.

In the simulations dedicated to factor Z_i, the frequency ω_i must satisfy: $\omega_i \geq 2M\max(\omega_j)$ where $\max(\omega_j)$ denotes the largest frequency associated with a factor other than Z_i. As for classical FAST, there is a minimum number N_0 of simulations, equal to $2M\omega_i + 1$. In practice, N_0 is usually chosen first, ω_i is then chosen as the largest integer satisfying $2M\omega_i + 1 \leq N_0$ and the other frequencies ω_j are chosen to satisfy the constraint $\omega_i \geq 2M\max(\omega_j)$ as well as a few other favourable properties.

The first-order sensitivity index of Z_i is estimated by

$$S_i = \sum_{p=1}^{M} S_{[p\omega_i]},$$

as in classical FAST. The total sensitivity index of Z_i is estimated by

$$TS_i = 1 - \sum_{\omega=1}^{M\max(\omega_j)} S_{[\omega]},$$

since all frequencies lower than $M\omega_{\max(j)}$ correspond to the factorial terms not involving Z_i.

Example on the winter wheat dry matter (continued):

Figure 13 displays results of an extended FAST sensitivity analysis for the model, repeated twenty times. For each replication and each input factor, a FAST sample of size 1000 was generated. Thus, the number of model simulations needed to estimate first-order and total indices was equal to 8×1000 per replication. For each replication, the phase-shift parameters ϕ were drawn at random, and the frequencies were randomly allocated to all factors except the one under study. The ranges of the estimates over the 20 series of simulations are displayed in Figure 13.

The results are very coherent with the Sobol estimates (Fig. 11). However they show much less variability between replications and a practical advantage is that the sensitivity indices are always positive, as expected.

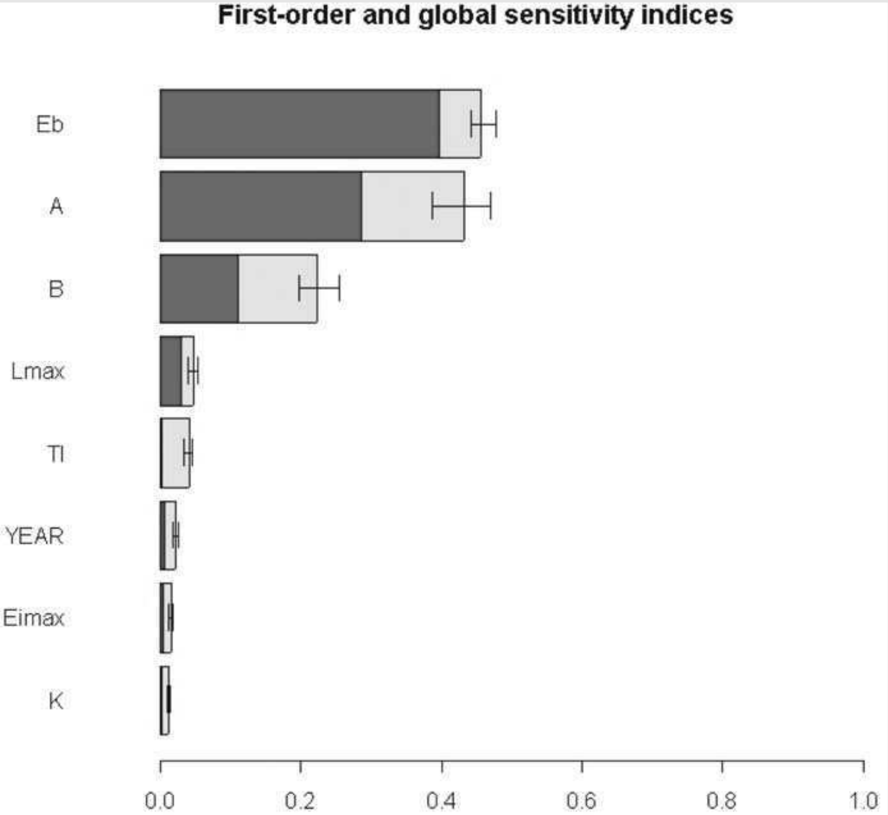


Figure 13. First-order and total sensitivity indices estimated by the extended FAST; there were 20 runs with 8 × 1000 model simulations for each run; the first part of the bars corresponds to the average (over the 20 runs) estimate of the first-order index, the full bars indicate average estimates of total indices, while the lines indicate extreme estimates of total indices.

5. Discussion

Which method to choose?

As the previous sections have shown, there is a large diversity of methods to perform sensitivity analyses. When choosing which one to use for a specific problem, important points to consider are the objectives of the study, the number of input factors to vary, the degree of regularity of the model and the computing time for individual model simulations.

If the objective is to screen for the most influential ones among a large number of input factors, the method of Morris or factorial designs are well adapted. Factorial designs with two-level factors are very efficient, but they give information only on the linear trends associated with each input factor. The method of Morris, by contrast, allows the investigation on the whole uncertainty ranges of the input factors.

When the objective is to quantify the influence of several input factors, experimental designs are very flexible, but once again, they give information on the model behaviour only for specific values of the input factors. Thus, it is necessary to assume, often implicitly, that the model is well-behaved and quite regular (for example, linear or near-linear if factors take two levels; near-quadratic if the factors take three levels, etc.). Methods based on intensive sampling, such as those described in the section on variance-based methods, have the advantage of being "model-free", that is, they do not rely on model approximations and they explore the full uncertainty ranges of the input factors. However, they require a large number of simulations.

In fact, there is no best method for all situations, and the differences between methods are less crucial than the accurate description of the uncertainty sources. A good understanding of the techniques and the ability to adapt them to one's situations is another key element.

Additional aspects of sensitivity analysis

Some key aspects of sensitivity analysis have been mentioned only briefly above but can be of great importance for a crop model.

The ability to take correlations into account between input factors, when generating scenarios, can make simulations much more representative of the phenomena under study. It was shown that such correlations can be taken into account in an uncertainty analysis. This is much more difficult for sensitivity analysis. There is a need to develop methods of sensitivity analysis that would take such correlations into account when interpreting simulation results.

It is often of great interest to consider the sensitivity of a response to a whole group of input factors (climatic/soil variables, or parameters associated with a specific growth stage). For most methods presented above, this can be done by summing the factorial indices associated with all factors within the group under consideration. The analogue of a first-order index is then the sum of all factorial indices involving only factors within the group. The analogue of a total index is the sum of all factorial indices involving at least one factor within the group. Note that this is not equal to the sum of the total indices of the factors within the group, because interactions are counted several times within a sum of total indices.

The sensitivity analysis of dynamic responses $\hat{Y}(t)$ has not been considered explicitly in this chapter. The methods described above can be applied to time t separately, and it may then be interesting to follow how sensitivity indices change with time. However, it is often more useful to perform sensitivity analyses on meaningful characteristics of the response time series. These characteristics can be either chosen by the modeller or determined by applying multivariate techniques such as the Principal Components Analysis or the Partial Least Squares to the simulated response time series (Campbell et al., 2004).

For all the methods considered until now, only one level of uncertainty was considered for each factor. However, it happens quite frequently that distinct levels of uncertainty need to be considered: for example, climate uncertainty at a local scale *versus* a regional scale; or uncertainty in parameters at present and after further experiments; or simply uncertainties in the true levels of uncertainty on some parameters. An application in forestry is presented by Gertner et al. (1996).

Software

General statistical packages make it possible to implement the methods of analysis based on experimental design, analysis of variance and regression. But it is necessary to be aware of some interpretations: the meaning of a significance test is dubious when the responses come from a perfectly deterministic model. The SAS QC© (SAS/QC© User's guide, 1999) module includes procedures to construct factorial and optimal designs (proc factex, proc optex).

Some software packages for general modeling or for risk analysis include methods for sensitivity or uncertainty analysis (Crystall Ball©, Risk©). There exist also softwares dedicated to sensitivity analysis. These software packages are restricted to the calculation of local sensitivity, but one exception is Simlab, which includes the main methods of global sensitivity analysis (see Saltelli et al., 2004).

References

Aggarwal, P.K., 1995. Uncertainties in crop, soil and weather inputs used in growth models: implications for simulated outputs and their applications. Agricultural Systems 48, 361–384.

Baret, F., 1986. Contribution au suivi radiométrique de cultures de céréales. Ph.D. Dissertation, Université Orsay, Orsay, France.

Bauer, L.R., Hamby, K.J., 1991. Relative sensitivities of existing and novel model parameters in atmospheric tritium dose estimates. Radiation Protection Dosimetry 37, 253–260.

Bärlund, I., Tattari, S., 2001. Ranking of parameters on the basis of their contribution to model uncertainty. Ecological Modelling 142, 11–23.

Box, G.E.P., Draper, N.R., 1987. *Empirical Model Building and Response Surfaces*. Wiley, New York.

Brun, R., Reichert, P., Künsch, H.R., 2001. Practical identifiability analysis of large environmental simulation models. Water Resources Research 37, 1015–1030.

Campbell, K., McKay, M.D., Williams, B.J., 2004. Sensitivity analysis when model outputs are functions. In *Proceedings of the Fourth International Conference on Sensitivity Analysis of Model Output (SAMO)*, March 8–11, 2004, Santa Fe, USA.

Chan, K., Tarantola, S., Saltelli, A., Sobol, I.M., 2000. Variance-based methods. In *Sensitivity Analysis*, A. Saltelli, K. Chan, and E.M. Scott (eds.), 167–197. Wiley, New York.

Colbach, N., Meynard, J-M., 1995. Soil tillage eyespot: influence of crop residue distribution on disease development and infection cycles. European Journal of Plant Pathology 101, 601–611.

Cukier, R.I., Fortuin, C., Shuler, K.E., Petshek, A.G., Schaibly, J.H., 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I. Theory. The Journal of Chemical Physics 59, 3873–3878.

Cukier, R.I., Schaibly, J.H., Shuler, K.E. 1975. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. III. Analysis of the approximations. The Journal of Chemical Physics 63, 1140–1149.

Dubus, I.G., Brown, C.D., 2002. Sensitivity and first-step uncertainty analyses for the preferential flow model MACRO. Journal of Environmental Quality 31, 227–240.

Gertner, G., Parysow, P., Guan, B., 1996. Projection variance partitioning of a conceptual forest growth model with orthogonal polynomials. Forest Science 42, 474–486.

Helton, J.C., Davis, F.J., 2000. Sampling-based methods. In *Sensitivity Analysis*, A. Saltelli, K. Chan, and E.M. Scott (eds.), 101–153. Wiley, New York.

Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis of model output. Reliability Engineering and systems Safety 52, 1–17.

Iman, R.L., Conover, W.J., 1982. A distribution-free approach to inducing rank correlation among input variables. Commun. Statist. Simul. Comput. B, 311–334.

Jansen, M.J.W., Rossing, W.A.H., Daamen, R.A., 1994. Monte Carlo estimation of uncertainty contributions from several independent multivariate sources. In: Gasman, J., van Straten, G. (eds), *Predictability and Non-Linear Modelling in Natural Sciences and Economics*, 334–343. Kluwer, Dordrecht.

Kobilinsky, A., 1997. Les plans fractionnaires. In J.-J. Droesbeke, J. Fine, and G. Saporta (eds.), 69–209. *Plans d'expériences: applications á l'entreprise*, Paris: Technip.

Koehler, J.R., Owen, A.B., 1996. Computer experiments. In *Handbook of Statistics, Vol. 13, Design and Analysis of Experiments*, S. Ghosh and C.R. Rao (eds.), 261–308. Elsevier, Amsterdam.

Lacroix, A., Beaudoin, N., Makowski, D., 2005. Agricultural water nonpoint pollution control under uncertainty and climate variability. Ecological Economics 53, 115–127.

Makowski, D., Naud, C., Monod, H., Jeuffroy, M.-H., Barbottin, A., 2004. Global sensitivity analysis for calculating the contribution of genetic parameters to the variance of crop model prediction. In *Proceedings of the Fourth International Conference on Sensitivity Analysis of Model Output (SAMO)*, 8–11 March, 2004, Santa Fe, USA.

Martinez, J.E., Duchon, C.E., Crosson, W.L., 2001. Effect of the number of soil layers on a modeled surface water budget. Water Resources Research 37, 367–377.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33, 161–174.

Rahn, C., Mead, A., Draycott, A., Lillywhite, R., Salo, T., 2001. A sensitivity analysis of the prediction of the nitrogen fertilizer requirement of cauliflower crops using the HRI WELL-N computer model. Journal of Agricultural Science 137, 55–69.

Racsko, P., Szeil, L., Semenov, 1991. Ecological Modelling: a serial approach to local stochastic weather models. Elsevier, Amsterdam.

Rossing, W.A.H., Daamen, R.A., Jansen, M.J.W., 1994. Uncertainty analysis applied to supervised control of aphids and brown rust in winter wheat Part 2. Relative importance of different components of uncertainty. Agricultural Systems 44, 449–460.

Ruget, F., Brisson, N., Delécolle, R., Faivre, R., 2002. Sensitivity analysis of a crop simulation model, STICS, in order to choose the main parameters to be estimated. Agronomie 22, 133–158.

Sacks, J., Welch, W., Mitchell, T., Wynn, H.P., 1989. Design and analysis of computer experiments. Statistical Science 4, 409–435.

Saltelli, A., Tarantola, S., Chan, K., 1999. A quantitative model-independent method for global sensitivity analysis of model output. Technometrics 41, 39–56.

Saltelli, A., Chan, K., Scott, E.M., 2000. *Sensitivity Analysis*. Wiley, New York.

Saltelli, A., Tarantola, S., Campolongo, F., 2000. Sensitivity analysis as an ingredient of modelling. Statistical Science 15, 377–395.

Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., 2004. *Sensitivity Analysis in Practice*. Wiley, New York.

Salvador, R., Piñol, J., Tarantola, S., Pla, E., 2001. Global sensitivity analysis and scale effects of a fire propagation model used over Mediterranean shrublands. Ecological Modelling 136, 175–189.

SAS/QC User's Guide 1999, version 8, SAS Publishing, Cary, 2098 p.

Schaibly, J.H., Shuler, K.E., 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. II. Applications. The Journal of Chemical Physics 59, 3879–3888.

Sobol, I.M., 1993. Sensitivity analysis for non-linear mathematical models. Mathematical Modelling and Computer Experiments 1, 407–414.

Turányi, T., Rabitz, H., 2000. Local Methods. In *Sensitivity Analysis*, A. Saltelli, K. Chan, and E.M. Scott (Eds), 81–99. Wiley, New York.

Venables, W.N., Ripley, B.D., 1999. *Modern Applied Statistics with S-PLUS*. Springer, Berlin.

Vose, D., 1996. *Quantitative Risk Analysis*. Wiley, New York.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., Morris, M.D., 1992. Screening, predicting and computer experiments. Technometrics 4, 15–25.

Exercises

Uncertainty and sensitivity analysis with a model predicting the percentage of diseased plants

We consider a model simulating the percentage of plants with eyespot pathogen (*Pseudocercosporella hypotrichosises*) in a field in function of cumulative degrees–days since sowing. The model is defined by

$$\hat{Y}(t) = 100 \times \frac{1 - \exp\left[-\left(c_1 + c_2\right)t\right]}{1 + \frac{c_2}{c_1}\exp\left[-\left(c_1 + c_2\right)t\right]}$$

where $\hat{Y}(t)$ is the predicted percentage of diseased plants when the cumulative degrees–days is equal to t, and $\theta = (c_1, c_2)$ are the model parameters.

Here, the objective is to predict the percentage for a field located in the Paris Basin at $t = 2300$C°day. The values of the two model parameters were studied previously (Colbach and Meynard, 1995) but accurate results are not known. In this study, we consider that the uncertainty ranges of c_1 and c_2 are $[4.5 \cdot 10^{-8}, 3.5 \cdot 10^{-4}]$ and $[4 \times 10^{-4}, 6.5 \times 10^{-3}]$ respectively. The nominal values of c_1 and c_2 are equal to 1.75×10^{-4} and 3.5×10^{-3}, respectively.

1. We assume that the uncertainty in c_1 and c_2 is modelled by uniform distributions over the parameter uncertainty ranges. A sample of ten values of $\theta = (c_1, c_2)$ is generated by Monte Carlo sampling and is reported in Table 8. Each value of θ defines an input scenario.
   - (a) Calculate $\hat{Y}(2300)$ for each one of the ten scenarios presented Table 8.
   - (b) Estimate the expected value and standard deviation of $\hat{Y}(2300)$ from the ten computed values of $\hat{Y}(2300)$.

Table 8. Ten values of c_1 and c_2 generated by Monte Carlo sampling.

c_1	c_2
1.71×10^{-4}	6.42×10^{-3}
1.25×10^{-4}	2.52×10^{-3}
9.65×10^{-5}	1.67×10^{-3}
3.38×10^{-4}	4.79×10^{-3}
2.97×10^{-4}	4.39×10^{-3}
4.88×10^{-5}	5.51×10^{-3}
1.36×10^{-4}	5.94×10^{-4}
2.99×10^{-5}	1.11×10^{-3}
1.97×10^{-4}	3.36×10^{-3}
3.17×10^{-4}	5.93×10^{-4}

(c) Estimate the probability $P[\hat{Y}(2300) \geq 80\%]$ from the ten computed values of $\hat{Y}(2300)$.
(d) The procedure described above is repeated 5 times leading to a 5 samples of 10 values of $\theta = (c_1, c_2)$. Each sample is used to estimate the expected value of $\hat{Y}(2300)$. The 5 estimated values of $E[\hat{Y}(2300)]$ are 93.59, 88.47, 95.28, 92.02, 96.48, 79.03. How do you explain this large variability?
(e) Define a procedure to choose the size of the sample of values of θ in order to estimate accurately $E[\hat{Y}(2300)]$ and $P[\hat{Y}(2300) \geq 80\%]$?

2. (a) Perform a local sensitivity analysis of $\hat{Y}(2300)$ with respect to c_1 and c_2 at the nominal parameter values. Which parameter has the highest relative sensitivity?
   (b) Calculate five equispaced values of c_1 and c_2 from the minimal to the maximal parameter values.
   (c) Set c_2 equal to its nominal value and calculate $\hat{Y}(2300)$ for the five values of c_1 defined above. Then, set c_1 equal to its nominal value and calculate $\hat{Y}$ (2300) for the five values of c_2.
   (d) Calculate the sensitivity index of Bauer and Hamby (1991) (see textbook) for each parameter from the computed values obtained in 2.(c). Which parameter has the highest index?
   (e) Calculate the sensitivity index of Bauer and Hamby (1991) for c_2 when c_1 is set equal to its minimal value. Compare this index value with the value obtained in 2.(d).

3. Consider a complete factorial design with three modalities per factor.
   (a) How many distinct scenarios (i.e. values of $\theta = (c_1, c_2)$) are included in this design?
   (b) Define a complete factorial design with three modalities per factor using only the minimal, nominal and maximal parameter values.
   (c) Calculate the general mean of $\hat{Y}_i(2300)$ where $\hat{Y}_i(2300)$ is the value of $\hat{Y}(2300)$ obtained with the ith scenario.
   (d) The total variability of $\hat{Y}(2300)$ can be measured by $\text{var}[\hat{Y}(2300)] = \frac{1}{N}\sum_{i=1}^{N}[\hat{Y}_i(2300) - \bar{Y}]^2$ where $\bar{Y}$ is the mean of $\hat{Y}_i(2300)$ and N the number of scenarios in the factorial design. Calculate $\text{var}[\hat{Y}(2300)]$.
   (e) Estimate $E[\hat{Y}(2300)|c_1]$ for each value of c_1 considered in the factorial design.
   (f) Estimate $E[\hat{Y}(2300)|c_2]$ for each value of c_2 considered in the factorial design.
   (g) Estimate $\text{var}\{E[\hat{Y}(2300)|c_1]\}$ and then $\frac{\text{var}\{E[\hat{Y}(2300)|c_1]\}}{\text{var}[\hat{Y}(2300)]}$.
   (h) Estimate $\text{var}\{E[\hat{Y}(2300)|c_2]\}$ and then $\frac{\text{var}\{E[\hat{Y}(2300)|c_2]\}}{\text{var}[\hat{Y}(2300)]}$.
   (i) To which sensitivity indices do $\frac{\text{var}\{E[\hat{Y}(2300)|c_1]\}}{\text{var}[\hat{Y}(2300)]}$ and $\frac{\text{var}\{E[\hat{Y}(2300)|c_2]\}}{\text{var}[\hat{Y}(2300)]}$ correspond?
   (j) Estimate $\text{var}[\hat{Y}(2300)|c_1]$ for each value of c_1 considered in the factorial design.
   (k) Estimate $\text{var}[\hat{Y}(2300)|c_2]$ for each value of c_2 considered in the factorial design.
   (l) Estimate $E\{\text{var}[\hat{Y}(2300)|c_1]\}$ and then $\frac{E\{\text{var}[\hat{Y}(2300)|c_1]\}}{\text{var}}[\hat{Y}(2300)]$.

(m) Estimate $E\{\mathrm{var}[\hat{Y}(2300)|c_2]\}$ and then $\frac{E\{\mathrm{var}[\hat{Y}(2300)|c_2]\}}{\mathrm{var}[\hat{Y}(2300)]}$.

(n) To which sensitivity indices do $\frac{E\{\mathrm{var}[\hat{Y}(2300)|c_1]\}}{\mathrm{var}[\hat{Y}(2300)]}$ and $\frac{E\{\mathrm{var}[\hat{Y}(2300)|c_2]\}}{\mathrm{var}[\hat{Y}(2300)]}$ correspond?

(o) Compare the indices calculated in 3.(n) and those calculated in 3.(i). Are they different? Why?

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 4

Parameter estimation for crop models

D. Makowski, J. Hillier, D. Wallach, B. Andrieu
and M.-H. Jeuffroy

1. Introduction **101**
2. Basic notions **102**
3. Standard nonlinear regression **106**
4. Estimation of parameters of complex crop models **116**
5. Discussion **138**
References **139**
Exercises **141**

1. Introduction

A large number of studies deal with parameter estimation in regression, but crop models have a number of characteristics which make much of this work inapplicable. The basic problem is that crop models usually have many parameters, and often more parameters than the number of data. Thus, it is generally numerically impossible to estimate all the parameters. On the other hand, there is usually information in addition simply to field data. Crop models are based on equations which describe the processes involved in crop growth and development, and there is in general information about these processes. For example, there might be information about the thermal time to flowering, which comes from controlled environment experiments, or information about maximum rate of root elongation from specific experiments on this aspect of crop growth. Thus the problem of parameter estimation for crop models is not a straightforward regression problem, involving estimating all the model parameter values from a set of field data. The problem is rather that of using both field data and information about growth and development to estimate model parameters.

Another important characteristic of crop models concerns the criteria for judging them, and therefore for judging any proposed method of parameter estimation. The quality of model predictions is often the very explicit criterion for judging a model (see the chapter on model evaluation). The practical consequence is that we are interested in statistical methods that are expected to perform well when predictive quality is the major criterion. Most statistical procedures aim rather at providing "good" parameter values (we will define this rigorously below). Of course, this is also of interest for crop models because their parameters often have a biological or physical meaning, and it is important for agronomists to determine accurately the values of such parameters. Nevertheless, the criteria of good parameter values and of good predictions are related but not identical.

One further specificity of crop models that needs to be kept in mind is the structure of the data. In order to have a sufficiently large and representative data set, one generally uses data from several different experiments for crop model parameter adjustment. The result is a data set which can be quite inhomogeneous, with different variables, and different measurement times, in different fields. Thus, we need a method of parameter estimation that is adapted to data sets with a complex structure.

The result of these characteristics is that parameter estimation for crop models is still a rather open field. It is not then surprising that there is no general consensus on the best approach to parameter estimation for these models. It is, however, rather surprising that there has been little discussion of the issue in the literature. The problem is very often noted, but few solutions have been proposed.

In this chapter we begin by a presentation of several important basic statistical concepts. We review successively the concepts of *population*, *samples of data*, *model error*, *parameter*, and *estimation*.

Next, we present nonlinear regression in the case of simple models when all parameters can be estimated from data. We discuss several practical problems like the heterogeneity of model error variances, correlation of model errors, and accuracy of parameter estimates. Several estimation methods (least squares, maximum likelihood...) are described and illustrated with nonlinear models predicting the kinetics of organ growth.

We then discuss the case of complex dynamic crop models that have many parameters. Here one decides to estimate only a subset of the model parameters, and then those parameters are estimated from data. This is actually a family of approaches, depending on how one chooses which parameters to estimate and how they are estimated. Several methods are described for selecting parameters and estimating their values from data and prior information. The value of Bayesian methods is emphasized.

2. Basic notions

2.1. Population and samples of data

In standard regression, model parameters are all estimated from a sample of measurements. Statisticians consider that this sample is taken from a target population defined by the set of all the different possible observed values. Of course, it is impossible to know the whole population. The sample of data is used to represent the target population and to estimate model parameters for this population.

For illustration, a population can be defined, for instance, by the set of all yield values that might be observed in wheat fields in northern France under standard crop management. In this case, a sample of observations consists of N values of wheat yield measurements taken from the population.

The size of the sample of data and the experimental design will depend on the cost of the measurements and on the time available to perform the experiments. Both the sample size and the experimental design have an influence on the accuracy of the parameter estimates.

2.2. Model error

We have explained in the first chapter of this book that a crop model can be thought of as a response function defined by $f(x;\theta)$ where f is the model equation, x is a vector of model inputs, and θ is a vector of model parameters. This function never predicts perfectly the response variables. The error, noted as $\varepsilon = y - f(x;\theta)$, is the difference between the prediction and the observed response y for a given situation. ε is a sum of two components, one corresponding to the measurement error and the other due to inadequacy of the response function $f(x;\theta)$.

The distribution of ε can be described by using a statistical model. The error ε is then defined as a random variable and the probability distribution of ε depends on one or several parameters. For example, when $\varepsilon \sim N(0, \sigma^2)$, the error distribution depends on only one parameter, σ^2, that represents the variance of the errors. The error distribution will usually include more parameters because the errors associated with different output variables often have heterogeneous variances or/and are correlated. For example, the errors associated with the values of biomass predicted by a model at different dates for a given field usually have different variances and are often correlated. In such a case, the error distribution will depend on a large number of parameters. In general, these parameters are unknown and must be estimated from data at the same time as the crop model parameters.

The definition of a realistic statistical model for the error distribution is an important step in the parameter estimation process. We will see below that, the choice of an inappropriate statistical model can have bad consequences on parameter estimates. For several reasons, it is often necessary to define very complex statistical models to describe realistically the distribution of the crop model errors.

2.3. Parameters

The objective of an estimation method is to estimate the values of the model parameters θ from a sample of data. A parameter is a numerical value that is not calculated by the model and is not a measured or observed input variable. The same quantity may or may not be a parameter depending on circumstances. For example, initial soil mineral nitrogen may be measured, in which case it is an input variable. In other cases it may not be measured, in which case it is a parameter that has somehow to be estimated.

In Section 2.2, we have defined θ as a vector of parameters. What is the exact meaning of these parameters? They represent the *true* parameter values. This notion is not very easy to understand because the true parameter values are unknown in practice. To understand its meaning, it is necessary to make some assumptions on the model errors. Here, we assume that the crop model equations f were chosen such that $E(\varepsilon) = 0$, where $E(.)$ is the expectation operator. This is realistic when f is a correct representation of the system. This assumption means that the model errors are centered on zero. Thus, the average value of all the possible observations y for a given set of input variables x is equal to the model prediction, i.e. $E(y|x) = f(x;\theta)$ where $E(.|.)$ is the conditional expectation operator. Now, we see that the true parameter value θ represents the parameter value leading to $E(y|x) = f(x;\theta)$ for all values of the input variables x taken from the population.

2.4. Estimation

Parameter estimation is an important subject in statistics. It is useful to distinguish two approaches, the frequentist and the Bayesian. The frequentist uses estimation methods to approximate the true parameter values θ by using only a sample of data. For the frequentist, parameters are not random variables but are fixed. Prior information on parameter values are not taken into account. Different types of frequentist methods (maximum likelihood, least squares...) were developed in the 1920s and 1930s by R.A. Fisher, J. Neyman, and E. Pearson notably. The application of a frequentist method to a particular dataset gives a *point estimate* of the model parameters and the function that relates point estimates to datasets is called an *estimator*.

The Bayesian estimates parameters from two different types of information, a sample of data (like the frequentist) and prior information about parameter values. The result of the application of a Bayesian method is a probability distribution of parameter values. All Bayesian methods proceed in two steps. The first step is to define a parameter probability distribution based on literature or expert knowledge. This distribution is called *prior parameter distribution* and reflects the initial state of knowledge about parameter values. The prior distribution can be, for example, a uniform distribution with lower and upper bounds derived from expert knowledge or a normal distribution. The second step consists in calculating a new parameter probability distribution from both the prior distribution and the available data. This new distribution, called *posterior parameter distribution*, is computed by using the Bayes theorem. The posterior distribution can be used in different ways. Point estimates of parameters can be taken as the expected value or, alternatively, the mode of the posterior distribution. The posterior parameter distribution can also be used for generating the probability distribution of the model outputs, for instance, the distribution of yield (see Chapter 3).

Bayesian methods are older than frequentist methods; the original paper of Rev. Thomas Bayes was published in 1763. But the Bayesian approach was neglected by the statisticians of the early twentieth century for two reasons. First, classical statisticians argued that the results obtained with Bayesian methods depend on the prior distribution and, so, are quite subjective. Second, the application of Bayesian methods requires the evaluation of complex integrals. It was almost impossible to perform these calculations until the recent access to inexpensive and fast computing.

Bayesian methods are attractive when parameters have a biological or physical meaning. In crop models, prior information on parameter values can be obtained from past studies carried out to analyze crop growth and development. As an example, consider the radiation use efficiency, noted E_b. Most crop models use E_b as a parameter and numerous studies has been carried out to determine the value of this parameter for different crops. For instance, according to Jeuffroy and Recous (1999), E_b for wheat is in the range 1.09–3.8 g MJ^{-1} depending on the cultivar and on the method of measurement. Bayesian methods allow agronomists to combine such information with data in order to estimate parameter values.

But the definition of a prior distribution is not always straightforward. This is particularly difficult when the sources of information are heterogeneous. In an analysis of parameter values of SUCROS87 reported in the literature, Metselaar (1999) found coefficients of variation for several parameters above 100%. The large variability in parameter values reported in the literature is due to heterogeneity of the methods used to perform measurements, to variability of parameter values between sites–years and to differences between expert opinions. It is clearly difficult to define a unique prior distribution for parameter values in such situations.

2.5. Criteria for choosing a method of estimation

Suppose that you consider two or more methods for estimating the parameters of your model. What is the best method? The first approach is to apply each method and to evaluate the performances of the crop model by using successively the different sets of parameter estimates. In this case, the criteria used to compare the methods of estimation are those presented in Chapter 2. For example, if the objective is to predict accurately a particular output variable, a natural criterion for comparing estimation methods is the mean squared error of prediction (*MSEP*) value for this output variable. The best estimation method will be the method leading to the smallest *MSEP*. Other criteria should be considered when the practical objective is decision support and not only prediction (Wallach et al., 2001; Makowski and Wallach, 2002).

The second approach is to evaluate the accuracy of the parameter estimates obtained with the different methods. To do that, it is necessary to define new criteria. We give here only the definitions of the criteria. Methods to calculate these criteria will be explained later in the chapter. The first criterion is the *mean squared error* of the parameter estimator. This criterion is equal to the expected value of the squared difference between the true parameter values and their estimated values. The mean squared error of the estimator $\hat{\theta}$ is defined by $E[(\hat{\theta} - \theta)^2]$. The expectation is over all the samples of data drawn from the target population. Mean squared error gives information on the average difference between the true parameter values and the values estimated from various data sets of the same size and structure. A low mean squared error value indicates that the estimated parameter values tend to be close to the true parameter values.

Mean squared error can be partitioned into two components, namely bias and variance:

$$E[(\hat{\theta} - \theta)^2] = [E(\hat{\theta}) - \theta]^2 + \text{var}(\hat{\theta}) = \text{bias}^2 + \text{variance}.$$

The first component is the squared value of the *bias*, $E(\hat{\theta}) - \theta$. The bias is a useful criterion for evaluating the quality of an estimation method. A bias different from zero reveals a systematic error. For instance, a positive bias indicates that the values taken by $\hat{\theta}$ for various datasets are, on the average, larger than the true parameter value. The second component of the mean squared error is called *variance of the estimator*. This is also an interesting criterion for evaluating the quality of an estimation method. It measures the variability of the estimated values across data sets. A large value of $\text{var}(\hat{\theta})$ indicates that a change in the dataset can have a strong effect on the estimated values. Bias and variances influence the accuracy of the model predictions (Miller, 1990). Consequently, it is important to use a method leading to small bias and variance to obtain accurate model predictions.

Finally, it is useful to introduce another criterion: the *correlation* between estimators of different parameters. The variance of an estimator gives information on the variability of the estimated parameter value across data sets but does not provide information on the variation of a parameter relative to other parameters. Correlation is often useful for studying the relationship between parameter estimates. In general, high correlation between parameter estimators is a problem.

3. Standard nonlinear regression

When the models include a small number of parameters (i.e. ≤ 10), it is often possible to estimate all parameter values from data by standard nonlinear regression. Statistical methods for nonlinear models have been presented in details in many books (e.g. Seber and Wild, 2003). In this section, we give only a brief overview of the most important techniques.

3.1. Examples of simple nonlinear models

We present here a nonlinear model simulating distinct biological growth phases for laminae, sheaths, and internodes of maize as a function of thermal time (degree days after emergence). This model will be used in the next sections to illustrate several methods. See Fournier and Andrieu (2000) for more details on multi-phase models.

The model contains a single exponential phase for the meristematic growth and the establishment of the elongation zone Eq. (1), followed by a linear phase for stationary extension Eq. (2), and a plateau Eq. (3) when extension is complete. It is as follows:

$$f(x;\theta) = L_{\mathrm{MIN}} \mathrm{e}^{R_1(x-T_0)} \quad \text{if} \quad T_0 < x \leq T_1 \tag{1}$$

$$f(x;\theta) = \alpha + \beta(x - T_1) \quad \text{if} \quad T_1 < x \leq T_2 \tag{2}$$

$$f(x;\theta) = L_{\mathrm{MAX}} \quad \text{if} \quad T_2 < x \tag{3}$$

with constraints

$$\alpha = L_{\mathrm{MIN}} \mathrm{e}^{R_1(T_1-T_0)} \tag{4}$$

$$L_{\mathrm{MAX}} = \alpha + \beta(T_2 - T_1) \tag{5}$$

to ensure continuity of the function at the phase transitions. The input variable x is cumulative degree-days. The model defined by Eqs. (1–5) is referred to as Model 2 (2 phases). We may or may not wish to ensure the continuity of the 1st derivative at the transition between the exponential and linear phases Eqs. (1–2), which would give the additional constraint

$$\beta = L_{\mathrm{MIN}} R_1 \mathrm{e}^{R_1(T_1-T_0)} \tag{6}$$

The model defined by Eqs. (1–6) is referred to as Model 2-C. L_{MIN} may be fixed arbitrarily to be the length at which the modelling begins, which means that there are 5 parameters in

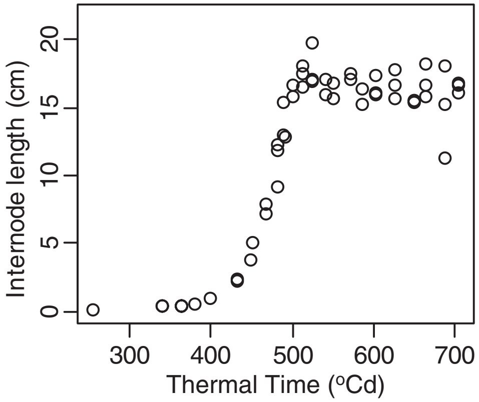


Figure 1. The length of the internode of the 8th phytomer as a function of thermal time.

Model 2 (R_1, T_0, T_1, T_2, β) and 4 parameters in Model 2-C (R_1, T_0, T_1, T_2). Models 2 and 2-C are nonlinear because they cannot be expressed as linear combinations of parameters. The numbers of parameters of these models are relatively small and standard nonlinear regression techniques can be used to estimate parameter values.

Parameters of models 2 and 2-C can be estimated for each organ from measurements of organ lengths. Figure 1 shows a fairly typical example of kinetics of organ extension, in this case the internode for the 8th phytomer of a maize plant (phytomer = leaf + internode + node). The measurements were obtained from plants 2–3 times a week from sowing to harvest. The extension initiates at a certain date, after which the length increases relatively slowly at first before the organ experiences a phase of rapid growth, followed by a plateau when growth is completed. An interesting observation which may be made from this figure is that the variance in the data appears to increase as a function of the organ size. For example, the variance calculated from replicates is equal to 0.00173 cm^2 at 363°C days, 0.845 cm^2 at 449°C days, and 2.09 cm^2 at 490°C days. A good understanding of this important property of the variance can improve the quality of the estimation of the parameters, as we shall see later.

3.2. Least squares estimation

3.2.1. Ordinary least squares

Suppose that a sample of N measurements, (x_i, y_i) $i = 1, \ldots, N$, is available for estimating the model parameter θ. The simplest method for estimating the model parameters is to apply ordinary least squares (OLS). The ordinary least squares estimate of θ minimizes the sum of the squared differences between the measurements and the model predictions

$$Z_{\text{OLS}}(\theta) = \sum_{i=1}^{N} [y_i - f(x_i;\theta)]^2 \tag{7}$$

In the case of linear models, it is possible to calculate analytically ordinary least squares estimates. The general form of a linear model is $f(x_i;\theta) = x_i^{\mathrm{T}}\theta$ where T is the transpose matrix operator. For this type of model, the OLS estimator of θ is $\hat{\theta}_{\mathrm{OLS}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ where $Y = (y_1,\dots,y_N)^{\mathrm{T}}$ and X is a matrix including the measured values of the input variables. The lines of X are $x_1^{\mathrm{T}},\dots,x_i^{\mathrm{T}},\dots,x_N^{\mathrm{T}}$. $\hat{\theta}_{\mathrm{OLS}}$ is related to the observations Y through a simple analytical function. This function can directly be used to calculate the values of the parameter estimates from data.

This is generally impossible with nonlinear models; parameter estimates cannot be expressed as functions of observations. The usual approach is to use an iterative algorithm for minimizing $Z_{\mathrm{OLS}}(\theta)$. We describe below one of these algorithm, the Gauss–Newton method (e.g. Seber and Wild, 2003).

Consider the nonlinear model defined by $f(x;\theta)$. The objective is to calculate the value of θ minimizing $Z_{\mathrm{OLS}}(\theta)$. The following algorithm is used to solve this problem:

(a) Define an initial value for θ, noted as $\hat{\theta}_0$.

(b) Linearize the model in order to replace the nonlinear model by a linear model. At this step, the linear Taylor expansion is calculated as follows:

$$f(x;\theta) \approx f(x;\hat{\theta}_0) + \sum_{j=1}^{p} \left.\frac{\partial f(x;\theta)}{\partial\theta_j}\right|_{\hat{\theta}_0} (\theta_j - \hat{\theta}_0 j),$$

where p is the total number of parameters and $\left.\frac{\partial f(x;\theta)}{\partial\theta_j}\right|_{\hat{\theta}_0}$ is the derivative of $f(x;\theta)$ relatively to the jth parameter taken at $\hat{\theta}_0$.

(c) Calculate the ordinary least squares estimate of θ for the linear model $f(x;\hat{\theta}_0) + \sum_{j=1}^{p} \left.\frac{\partial f(x;\theta)}{\partial\theta_j}\right|_{\hat{\theta}_0} (\theta_j - \hat{\theta}_0 j)$. This estimate is obtained by minimizing $\sum_{i=1}^{N}\left\{y_i - \left[f(x_i;\hat{\theta}_0) + \sum_{j=1}^{p} \left.\frac{\partial f(x_i;\theta)}{\partial\theta_j}\right|_{\hat{\theta}_0} (\theta_j - \hat{\theta}_{0j})\right]\right\}^2$. As the model is now linear, an analytical expression for the estimator can be derived as explained above. We obtain $\hat{\theta}_1 = \hat{\theta}_0 + \Delta_0$, where $\hat{\theta}_1$ is the new parameter estimates, $\Delta_0 = (F.^{\mathrm{T}}F.)^{-1}\ F.^{\mathrm{T}}[Y - F]$, $F = [f(x_1;\hat{\theta}_0),\dots,f(x_N;\hat{\theta}_0)]^{\mathrm{T}}$, and $F.$ is a $(N \times p)$ matrix whose elements are defined by $\left.\frac{\partial f(x_i;\theta)}{\partial\theta_j}\right|_{\hat{\theta}_0}$.

(d) Replace $\hat{\theta}_0$ by $\hat{\theta}_1$ and return to step *a*.

For illustration, consider the simple nonlinear model, $f(x;\theta) = \mathrm{e}^{\theta x}$, including only one parameter θ and one input variable x. At step *b*, the linear Taylor expansion for this model is $\mathrm{e}^{\theta x} \approx \mathrm{e}^{\hat{\theta}_0 x} + (\theta - \hat{\theta}_0)x\mathrm{e}^{\hat{\theta}_0 x}$ and, at step *c*, $\hat{\theta}_1$ is calculated as

$$\hat{\theta}_1 = \hat{\theta}_0 + \frac{\sum_{i=1}^{N} x_i \mathrm{e}^{\hat{\theta}_0 x_i}(Y_i - \mathrm{e}^{\hat{\theta}_0 x_i})}{\sum_{i=1}^{N} x_i^2 \mathrm{e}^{2\hat{\theta}_0 x_i}}.$$

Different parameter values $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ are successively calculated with this algorithm. It can be shown that these values converge to the ordinary least squares estimate of θ. The algorithm stops when the difference $\sum_{i=1}^{N} [y_i - f(x_i; \hat{\theta}_{k+1})]^2 - \sum_{i=1}^{N} [y_i - f(x_i; \hat{\theta}_k)]^2$ is lower than a small threshold value defined in the algorithm. The last parameter value generated by the algorithm is used as an estimate of θ.

The Gauss–Newton algorithm and related algorithms (e.g. Gauss–Marquardt) can be applied with commercial software like SAS (PROC NLIN, SAS/STAT User's Guide, 1990) or S-PLUS (*NLS* function, SPLUS 6, 2001). It must be emphasized that numerical problems often arise when the parameters to estimate are numerous (>10–15). In some cases, the algorithms do not converge to the true minimal value of $Z_{\mathrm{OLS}}(\theta)$ but to a local minimum. This problem is illustrated in Figure 2 where values of $Z_{\mathrm{OLS}}(\theta)$ are reported as a function of the parameter value for a hypothetical model. There are two local minima and one global minimum (the optimal solution). Depending on the initial value θ_0, a Gauss–Newton type algorithm may or may not converge to the global minimum. With some initial values, the algorithm will converge to a local minimum and the resulting parameter value will differ from the least square estimate. To avoid this kind of problem, it is important to run the algorithm with different starting values successively. Another solution consists in using global optimization algorithm like, for instance, simulated annealing (Goffe et al., 1994). These algorithms are less sensitive to initial values than Gauss–Newton type algorithm and are more likely to converge to the optimal solution. An important drawback of these algorithms is that they often require a long calculation time. An application of simulated annealing to a crop model is described by Mavromatis et al. (2001).

The OLS method is simple and is commonly used by crop modelers. See Grimm et al. (1993) for an application to a crop model. This method has good properties if the model errors are normally distributed and are independent with constant variance and zero mean for all x values (e.g. Huet et al., 1992). Under these assumptions, the OLS estimators

- converge toward true parameter values when the number of observations is large;
- are unbiased;
- have minimum variances $\mathrm{var}(\hat{\theta})$ among all unbiased estimators.

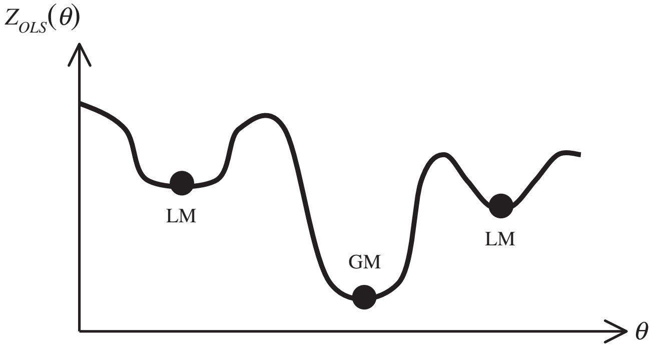


Figure 2. Local minima (LM) and global minimum (GM).

In many situations, the variance of the observations is not constant. For example, the variances of the measurements displayed in Figure 1 increase with the organ size. In this situation, the model errors are likely to have heterogeneous variances and an application of the ordinary least squares method to our organ growth models will lead to estimators which are unbiased but which do not have minimum variance. Another common problem is that the model errors are not independent. In this case also, ordinary least squares estimators do not have minimum variance.

3.2.2. Generalized least squares

Variance heterogeneity and correlation between model errors can be taken into account by using the method of generalized least squares (Seber and Wild, 2003). The function to be minimized is now

$$Z_{\mathrm{GLS}}(\theta) = [Y - F(\theta)]^{\mathrm{T}} V^{-1} [Y - F(\theta)] \tag{8}$$

where Y is a vector including the N observations, $Y = (y_1, \ldots, y_N)^{\mathrm{T}}$, $F(\theta)$ is a vector including the N model predictions, $F(\theta) = [f(x_1; \theta), \ldots, f(x_N; \theta)]^{\mathrm{T}}$, and V is $(N \times N)$ variance–covariance matrix of the model errors. The diagonal elements of V are the variances of the model errors and the off-diagonal elements are the covariances of the model errors. Ordinary least squares is a special case in which V is the identity matrix (the diagonal elements are all equal to 1 and the off-diagonal elements are all equal to zero).

Another special case is when V is diagonal (the diagonal elements are different from 1 and the off-diagonal elements are all equal to zero). Then, $Z_{\mathrm{GLS}}(\theta)$ is equal to

$$\sum_{i=1}^{N} \frac{[y_i - f(x_i; \theta)]^2}{\sigma_i^2} \tag{9}$$

where σ_i^2 is the variance of the model error ε_i, $i = 1, \ldots, N$. Equation (9) is the weighted sum of the squared differences between model predictions and observations. The value of θ minimizing Eq. (9) is often referred to as the weighted least squares estimate.

When the matrix V is known, the minimization of Eq. (8) can be performed by using the same algorithms as those described for ordinary least squares. The first step is to express the matrix V as $V = R^{\mathrm{T}} R$ (Cholesky decomposition) where R is an upper triangular matrix. At the second step, the data are transformed as follows:

$$Y_* = (R^{\mathrm{T}})^{-1} Y \quad \text{and } F_*(\theta) = (R^{\mathrm{T}})^{-1} F(\theta)$$

where Y_* and $F_*(\theta)$ are the vectors including the transformed observations and the transformed predicted values respectively. Finally, the last step consists in calculating the value of θ by minimizing $[Y_* - F_*(\theta)]^{\mathrm{T}}[Y_* - F_*(\theta)]$. That is, when V is known, the problem can be transformed to an OLS problem. As a result, all the OLS properties hold.

The drawback of this method is that it requires the knowledge of the matrix V. In some simple cases, it is possible to estimate the variances and covariances from replicates.

Vold et al. (1999) describe an application to a model simulating the dynamics of carbon and nitrogen in the soil. Their dataset includes three replicates at each date of measurement and these replicates are used to estimate a variance–covariance matrix.

The estimation of the elements of V from replicates is often impossible in practice. In the dataset presented in Figure 1, we see that only one measurement was performed at some dates. In this case, it is impossible to estimate variances from replicates for all dates of measurements. An alternative is to estimate θ and V iteratively (Gallant, 1987). A first estimate of θ is obtained with an initial value of V (e.g. an identity matrix). Then, the value of V is updated using the variance–covariance matrix of the model residues. This procedure is repeated until the difference between two successive estimates is negligible. An alternative is to use the maximum likelihood method described below.

3.3. Maximum likelihood

Suppose that N observations, $Y = (y_1, \ldots, y_N)^{\mathrm{T}}$, are available for estimating parameters. Let E denote the vector of the N model error terms, $E = (\varepsilon_1, \ldots, \varepsilon_N)^{\mathrm{T}}$. We assume that E is normally distributed, $E \sim N(0, V)$ where V is $(N \times N)$ variance–covariance matrix of the model errors. Under this assumption, the likelihood of θ and V is defined by

$$\begin{aligned} L(y_1, \ldots, y_N; \theta, V) &= P(y_1, \ldots, y_N|\theta, V) \\ &= (2\pi)^{-N/2}|V|^{-1/2}\exp\{-½[Y - F(\theta)]^{\mathrm{T}}V^{-1}[Y - F(\theta)]\} \end{aligned}$$

where $F(\theta)$ is a vector including the N model predictions, $F(\theta) = [f(x_1; \theta), \ldots, f(x_N; \theta)]^{\mathrm{T}}$, and $P(.)$ is a probability density function. The likelihood represents the probability of the observations for given values of θ and V. The maximum likelihood estimates of θ and V are the values maximizing this probability. When V is known, the value of θ maximizing $L(y_1, \ldots, y_N; \theta, V)$ is equal to the value minimizing $[Y - F(\theta)]^{\mathrm{T}}V^{-1}[Y - F(\theta)]$. Thus, in this case, the generalized least squares estimator is equal to the maximum likelihood estimator.

Numerical methods were developed for maximizing the likelihood with nonlinear models e.g. GNLS in R and SPLUS (R, http://www.r-project.org; Ripley, 2001; SPLUS6, 2001). These algorithms find the local maximum of the likelihood function if an initial solution is proposed, via the iterative solution of linearized regression problems.

3.4. Parametric modeling of the error variance

The size of the variance–covariance matrix V is large when the measurements are numerous. The estimation of a big matrix V can lead to numerical problems and to inaccurate estimates. In such cases, it is useful to simplify the matrix V. A first approach is to set some elements equal to zero. For example, when the model errors are independent, the off-diagonal elements of the variance–covariance matrix can be fixed to zero. Then, only the diagonal elements are estimated from data. Another approach is to describe the matrix V using just few parameters that need to be estimated, rather than estimating each element of the matrix.

For illustration, we consider the estimation of the parameters of the multi-phase models defined in Section 3.1. First, we assume that the model errors are independent i.e. $\text{cov}(\varepsilon_i, \varepsilon_{i'}) = 0$, $i \neq i'$. As a consequence, the variance–covariance matrix of the model errors is diagonal. Second, we define a parametric model for the error variances in order to reduce further the number of elements of the matrix. This approach is described below.

Figures 1 and 3 show that the variance of the measurements tends to increase with organ length. Consequently, it seems realistic to define the following statistical model for the error variance: $\text{var}(\varepsilon) = \sigma^2 f(x;\theta)^\tau$ (Huet et al., 1996). With this model, the variances of the errors increase with length $f(x;\theta)$ if $\tau > 0$ and the diagonal elements of the variance–covariance matrix depend only on θ and on two additional parameters, namely σ and τ.

The maximum likelihood method is implemented for estimating θ, σ, and τ from data. To apply this method, we assume that the model errors ε are normally and independently distributed. The maximum likelihood estimates of θ, σ, and τ are then obtained by maximizing the following likelihood function:

$$\begin{aligned} L(y_1,\ldots,y_N;\theta,\sigma,\tau) &= P(y_1,\ldots,y_N|\theta,\sigma,\tau) \\ &= P(y_1|\theta,\sigma,\tau)\times\cdots\times P(y_N|\theta,\sigma,\tau) \\ &= \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2 f(x_i;\theta)^\tau}}\exp\left\{-\frac{[y_i - f(x_i;\theta)]^2}{2\sigma^2 f(x_i;\theta)^\tau}\right\} \end{aligned} \tag{10}$$

where y_i is the organ length measured at time x_i. The likelihood Eq. (10) is the product of the probabilities of observing each of the measured values given the values of θ, σ, and τ.

We used GNLS to maximize the likelihood function Eq. (10) for estimating the four parameters of model 2-C for the lamina of the phytomer 9 of a maize plant. Figure 3 shows the organ length measurements used for parameter estimation and the fitted model. It is informative to plot on both linear and log scales so that the fit for all phases can be assessed.

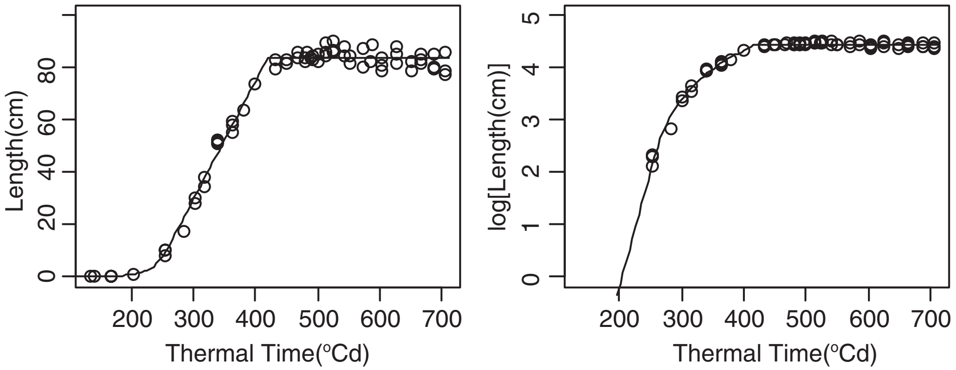


Figure 3. Length of lamina of the 9th phytomer as a function of thermal time after emergence. Fitted with Model 2-C. Results are shown in linear (left) and log scales (right).

3.5. Measuring accuracy of parameter estimates

The accuracy of the parameter estimates can be evaluated by calculating the variance of the estimator across data samples, noted $\text{var}(\hat{\theta})$. Most of the software developed for estimating the parameters of nonlinear models are able to return estimates of $\text{var}(\hat{\theta})$. This can be very useful, for example to compare parameter estimates obtained for different datasets corresponding to different experimental treatments.

Sometimes, we are not only interested in the model parameters θ but also in composite functions of the model parameters which we may wish to compare between experimental treatments. For example, in the organ growth models presented in Section 3.1, we are interested in the model parameters and also in the functions defined in Eqs. (5) and (6). Although the formulae for these functions permit their calculation as a function of the model parameters θ, it is not possible to obtain a parametric estimate of their standard errors. The bootstrap is a nonparametric resampling method offering a solution to this problem (e.g. Efron and Gong, 1983).

The bootstrap is based on the idea that, working with the available data only and using re-sampling techniques, one can simulate other datasets which might have been obtained from the same field trial. These simulated datasets can then be used to obtain a distribution of values for the model parameter estimates. Once this is done, quantities such as the standard error of the parameter estimator may be calculated from the series of the estimated values obtained with the different samples.

In general, if the dataset contains N datapoints, then the standard re-sampling method is to simply select N datapoints randomly, with repetition, from this dataset. In our case, where each datapoint is in fact a (time, organ length) pair, there are possible problems associated with this method, since it is likely that the resampled distribution does not reproduce the original sampling regime, i.e. it does not respect the number of samples at a given point in time. Aside from the fact that this is contrary to the stated aim of bootstrap resampling above (to simulate alternative datasets which might be obtained with the same sampling regime), there is also the possibility that any given simulated dataset actually contains fewer dates than the original, which may prevent a numerical solver (e.g. GNLS) from converging. We thus present a commonly used variant of the bootstrap, which involves resampling the residuals as opposed to the data pairs.

This alternative method, which is outlined in Efron and Tibshirani (1988), is illustrated here with our multi-phase model. We first fit the model to data and obtain estimated values of θ, σ, and τ. Then, supposing the errors are normally distributed at any given point in time, we may use the variance model to obtain a new set of errors, which may then be added to the model in order to simulate a new data set. Assuming that the variance model is appropriate for the original data this yields simulated datasets with, on average, the same properties as the original with respect to the residuals.

Suppose that the function $g(\theta)$ is a composite of the model parameters (e.g. Eq. (5)). The stages of the method are given below for the organ growth models:

(1) Obtain estimated parameter values $\hat{\theta}$, $\hat{\sigma}$, $\hat{\tau}$ using a nonlinear model fitting procedure

(2) Obtain M sets of simulated errors $E_1, \ldots, E_k, \ldots, E_M$ using the estimated values $\hat{\theta}$, $\hat{\sigma}$, $\hat{\tau}$: E_k is the set of ε_{ik} sampled randomly from the normal distribution $N[0, \hat{\sigma}^2 f(x_i; \hat{\theta})^{\hat{\tau}}]$ for all $i = 1, \ldots, N$, where N is the number of datapoints.

3. Add the simulated errors to the model to obtain new datasets, $D_1, \ldots, D_k, \ldots, D_M$ where $D_k = \{Y_{ik} = \varepsilon_{ik} + f(x_i; \hat{\theta})\}$.
4. Repeat model fitting for each of these datasets to obtain M solution vectors $\hat{\theta}_1, \ldots, \hat{\theta}_k, \ldots, \hat{\theta}_M$.
5. Calculate the standard error of the function, g, of the parameters using the following formula:

$$SE[g(\theta)] = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}\left[g(\hat{\theta}_j) - \frac{1}{M}\sum_{k=1}^{M} g(\hat{\theta}_k)\right]^2}.$$

The model 2-C was fitted to measurements of lamina length obtained for the phytomers 4–15 of maize plants. Plants were taken from two fields, one field sown at a normal plant density (9.5 pl.m^{-2}), and another field sown at a high plant density (30.5 pl.m^{-2}). Parameters of model 2-C were estimated by maximum likelihood for each phytomer and each density. In Figure 4, we compare the estimated values for 4 different functions g, namely final lamina length $L_{\mathrm{MAX}} = \alpha + \beta(T_2 - T_1)$, parameter R_1, slope in linear phase

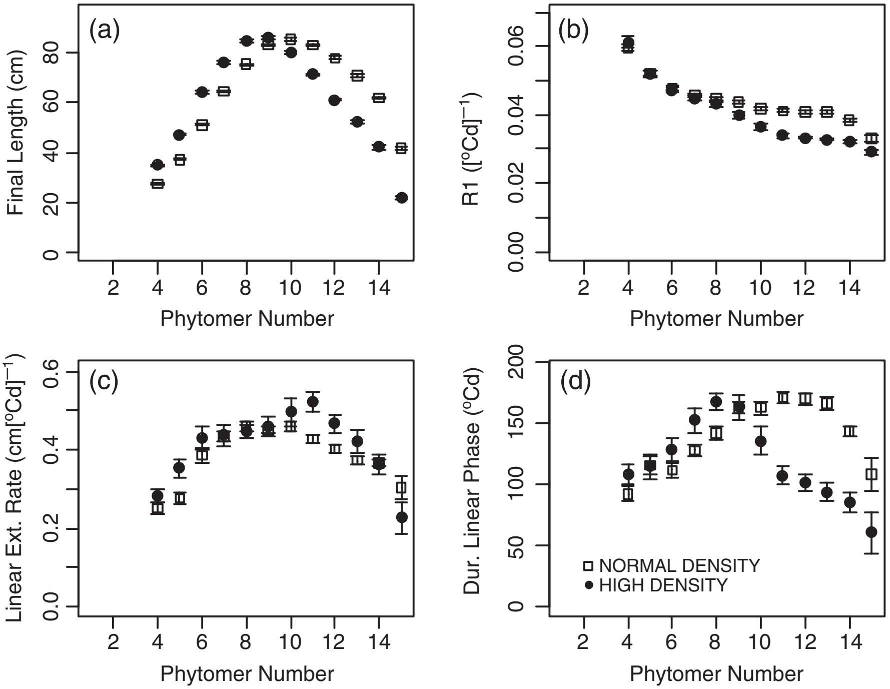


Figure 4. Estimated values for final length $L_{\mathrm{MAX}} = \alpha + \beta(T_2 - T_1)$ (a), R_1 (*b*), slope in linear phase $\beta = L_{\mathrm{MAX}} R e^{R_1(T_1 - T_0)}$ (c), and duration of linear phase $T_2 - T_1$ (d) for the laminae in normal density and high density maize crops. Points and squares indicate average parameter estimates and error bars represent ±2 standard errors, as calculated from the bootstrap with 500 samples of data.

$\beta = L_{\mathrm{MIN}} R e^{R_1(T_1 - T_0)}$, and duration of linear phase $T_2 - T_1$. In all cases, the standard errors have been calculated via the bootstrap method, using $M = 500$ samples of data. From these results, it becomes possible to

- observe the evolution of parameter values as a function of phytomer number;
- compare differences in parameter values of the corresponding organ between treatments (normal density *vs.* high density).

3.6. Convergence of the iterative procedures

All the methods discussed above all require the minimization or the maximization of a parameter function (e.g. $Z_{\mathrm{OLS}}(\theta)$). As explained before, these functions can rarely be minimized or maximized analytically and so an iterative procedure for finding the optimal parameter values is usually required. When the number of parameters is high relatively to the number of available data, these algorithms often fail to converge to the true optimum.

We encountered this problem with Model 2. We considered the estimation of the 5 parameters of Model 2 for the sheath of the 9th phytomer of a maize plant from 50 length measurements. We tested 100 initial vectors, ran GNLS with these initial values, and compared the 100 resulting series of parameter estimates. Figure 5 shows the 100 estimated values of parameters R_1 and T_0. The result reveals that the estimated parameter values depend strongly on the initial parameter values.

This problem may arise with numerical solvers which operate by starting with a proposed initial solution and then converging to a maximum of a likelihood function or to a minimum of a least squares function. Nonconvergence of numerical solvers is usually indicative of an intrinsic problem of matching the model to the data. The presence of a rather flat likelihood or least squares function is commonly due to the fact that the ratio of parameters to data points is too high, in other words, the model has too many parameters. So in this case all the solutions should be viewed with caution.

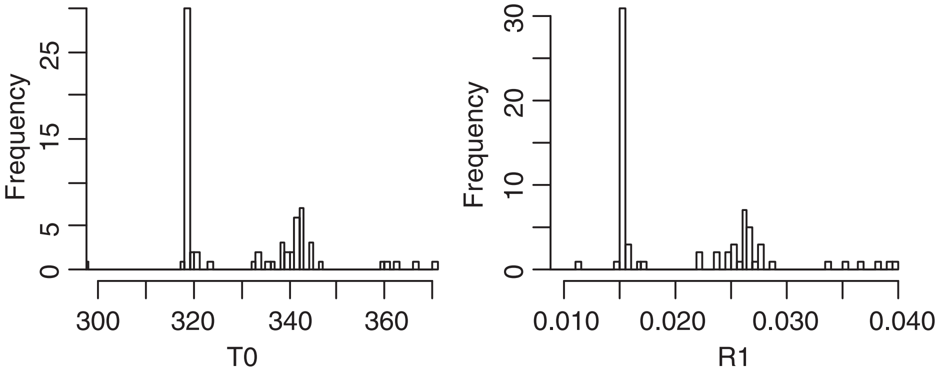


Figure 5. The values of R_1 and T_0 returned by GNLS from 100 different initial vectors randomly generated.

It is also important to note that parameter estimates can have very high variance when numerous parameters are estimated simultaneously from a small number of data. High variances of parameter estimates can lead to high *MSEP* values (i.e. inaccurate model predictions) as shown in Chapter 2. Thus, it is important to keep the ratio of the number of estimated parameters to the number of observations relatively low. According to Harrel (2001), the ratio of the number of estimated parameters to the number of observations should not be higher than about 1/20.

4. Estimation of parameters of complex crop models

Complex dynamic crop models include many parameters. For example, the STICS crop model (Brisson et al., 1998) includes more than 200 parameters. The standard statistical methods presented in Section 3 cannot be directly applied to these models. The estimation of a large number of parameters by least squares or maximum likelihood leads to inaccurate estimated values (high variances of the parameter estimators) and to inaccurate model predictions (high *MSEP*). This problem is often called *overparametrization*. Another problem, already encountered in Section 3.6, is that the numerical solvers presented in Section 3 often fail to converge to the optimum when the parameters are too numerous. Finally, in many crop models, it is impossible to estimate simultaneously all the parameters because several parameters are *unidentifiable* due to the structure of the model equations. Lack of identifiability occurs when several sets of parameters lead to the same model prediction.

For all these reasons, a common practice is to select a subset of parameters, to estimate those parameters from data by least squares, and to set the others equal to predefined values. The implementation of this approach requires one to decide which among all the parameters will be adjusted to the data and to choose a method for estimating the values of the selected parameters. These two issues are discussed below. This is one way of combining data and prior information for parameter estimation. An alternative is a Bayesian approach which is also discussed below.

4.1. Selection of parameters to estimate

Four methods are proposed here for selecting parameters:

- selection based on the literature;
- selection to avoid identifiability problems;
- sensitivity analysis;
- statistical choice of parameters to estimate.

4.1.1. Selection based on the literature

The selection of parameters is often based on literature. The principle is to estimate with data only the parameters that were not sufficiently studied in the literature. Wellknown parameters are fixed to values provided by the literature and the others are estimated using experimental data. This approach was applied by Bonesmo and Bélanger (2002) to estimate 17 parameters of a crop model. These authors set four wellknown parameters equal to values provided by the literature. The others were estimated by using a dataset.

An important drawback of this method is that it is not easy to determine if the information provided by the literature is sufficient or not. Therefore, selection of parameters based on literature has strong subjective elements. Note that this approach does not provide any protection against overparametrization.

4.1.2. Selection to avoid identifiability problems

Identifiability problems occur when there is no unique solution to the parameter estimation problem. A careful analysis of the model equations is often useful to avoid this kind of problem.

This approach can be illustrated with a submodel of the AZODYN crop model (Jeuffroy and Recous, 1999). This submodel includes three output variables, namely, above-ground winter wheat dry matter (*DM*, kg ha^{-1}), leaf area index (*LAI*, m^2 m^{-2}), and nitrogen uptake (*NU*, kg ha^{-1}). These variables are simulated daily starting at the end of winter until flowering in the absence of nitrogen stress, water stress, pests, and diseases. Dry matter is calculated as follows:

$$DM_j = DM_{j-1} + EBMAX \times ft_{j-1} \times EIMAX[1 - \exp(-K \times LAI_{j-1})] \times C \times gr_{j-1}$$

where DM_j and DM_{j-1} represent dry matter on days j and $j-1$ respectively, ft_{j-1} is a function taking into account the temperature on day $j-1$, gr_{j-1} is the global radiation on day $j-1$ (MJ ha^{-1}). *C*, *EBMAX* (kg MJ^{-1}), *EIMAX*, and *K* are four parameters. Before flowering, LAI_{j-1} is calculated in function of the critical nitrogen uptake level on day $j-1$ (NUC_{j-1}, kg ha^{-1}) as follows:

$$LAI_{j-1} = D \times NUC_{j-1}$$

where *D* is a parameter. Consequently, DM_j is related to NUC_{j-1} by

$$DM_j = DM_{j-1} + EBMAX \times ft_{j-1} \times EIMAX[1 - \exp(-K \times D \times NUC_{j-1})] \times C \times gr_{j-1}.$$

The last equation shows that, when only dry matter measurements are available, it not possible to estimate simultaneously *K* and *D*, and it is also impossible to estimate simultaneously *EBMAX*, *C*, *EIMAX*. Only the products $EBMAX \times C \times EIMAX$ and $K \times D$ can be estimated because simulated dry matter depends only on these two products. A numerical application shows that many sets of parameter values give identical values of $EBMAX \times C \times EIMAX$ and $K \times D$: we obtain $K \times D = 0.02016$ with $K = 0.72$ and $D = 0.028$ but also with $K = 0.6$ and $D = 0.0336$. Only two parameters can be estimated from dry matter measurements, one parameter among (*EBMAX*, *C*, *EIMAX*) and one parameter among (*K*, *D*). When both dry matter and *LAI* measurements are available, it is possible to estimate one more parameter. As *D* has an influence on the *LAI*, it is possible to estimate simultaneously *K* and *D* from dry matter and *LAI* measurements. However, it is still impossible to estimate *EBMAX*, *C* and *EIMAX*. Therefore, it is necessary to fix two of *EBMAX*, *C*, and *EIMAX* and to estimate the remaining parameter.

4.1.3. Selection by sensitivity analysis

Another method for selecting parameters is to perform a sensitivity analysis. The principle is to calculate a sensitivity index for each parameter and to select parameters with high sensitivity index values. This method allows modelers to identify the parameters that have a strong influence on the output variables of a model. Only these parameters are estimated from data and others are fixed to values provided by the literature. The implementation of this method requires the definition of a threshold of sensitivity. A common approach consists in defining the number of parameters to estimate before performing the sensitivity analysis. An interesting application is presented in Harmon and Challenor (1997). Chapter 3 provides more detail on this subject. Even though this method protects against estimating parameters that would be very hard to estimate from data, it does not protect against overparametrization.

4.1.4. Selection to minimize error of prediction

We present here a statistical method for selecting the parameters to estimate (Wallach et al., 2001). With this method, parameters are selected in order to minimize the errors of prediction of the crop model. The method is very similar to forward regression. It is implemented in three steps:

Step 1: Definition of a method for estimating model parameters and definition of a criterion for evaluating the accuracy of the model predictions. Wallach et al. (2001) consider the crop model 2CV. This dynamic model includes 26 parameters and three output variables, namely, maize yield, biomass and *LAI*. The method chosen for estimating parameters is weighted least squares. The criterion minimized with this method is a weighted sum of squared model errors denoted as Z_{WLS}. The criterion chosen for evaluating the accuracy of the model predictions is *MSEP* estimated by cross validation, noted here as $MSEP_{\mathrm{cv}}$ (see Chapter 2 for further explanations about *MSEP*).

Step 2: Selection of the parameters to estimate. First, each parameter is adjusted to data individually to minimize Z_{WLS}, while the other parameters all keep their initial values (provided by the literature). The parameter that leads to the smallest value of Z_{WLS} is the first parameter selected. Next, all combinations of the best single parameter and one of the remaining parameters are adjusted to minimize Z_{WLS}. The best second parameter is then selected, and so on. At the end of this step, the best models with 1, 2, 3, 4, ... adjusted parameters are known. Note that this procedure of selection could be replaced by sensitivity analysis.

Step 3: How many parameters? The values of $MSEP_{\mathrm{cv}}$ are calculated for the models from Step 2 (with 1, 2, 3, ... adjusted parameters). The model finally chosen is the model with the smallest value of $MSEP_{\mathrm{cv}}$.

The results obtained for the 2CV crop model are presented in Table 1. The square root of $MSEP_{\mathrm{cv}}$ ($RMSEP_{\mathrm{cv}}$) calculated for yield is reported as a function of the number of

Table 1. Root mean squared prediction error values estimated by cross validation for yield ($RMSEP_{\mathrm{cv}}$) when 0, 1, 2, 3, or 4 parameters of the crop model 2CV are estimated from experimental data (Wallach et al., 2001).

Number of estimated parameters	0	1	2	3	4
$RMSEP_{\mathrm{cv}}$ (t ha^{-1})	2.48	2.17	1.68	1.50	1.57

adjusted parameters. The model with zero adjusted parameter is the 2CV model with all parameters set equal to their initial values. In this case, $RMSEP_{\mathrm{cv}}$ is equal to 2.48 t ha^{-1}. Table 1 shows that $RMSEP_{\mathrm{cv}}$ is lower when some of the model parameters are adjusted to data. The smallest value of $RMSEP_{\mathrm{cv}}$ is obtained when three parameters are adjusted to data. The model predictions are less accurate when more than three parameters are estimated. In this example, the optimal solution is to estimate only three parameters from data and to fix the other 23 parameters to initial values.

The method described above is very attractive but is not easy to apply. The calculation time required to perform cross validation can be very long. Moreover, when data are not numerous, cross validation can give inaccurate results. A solution is to replace $MSEP_{\mathrm{cv}}$ in Step 3 by a simpler criterion like the Akaike Information Criterion (AIC) (Sakamoto et al., 1986) or the Bayesian Information Criterion (BIC) (Schwarz, 1978) defined by:

$$\mathrm{AIC} = -2 \times \log Lik + 2 \times P$$

$$\mathrm{BIC} = -2 \times \log Lik + \log(N) \times P$$

where log *Lik* is the logarithm of the model likelihood (see Section 3.3), *P* is the number of estimated parameters and *N* is the number of data. A simulation study was carried out by Tremblay and Wallach (2004) to compare performances of the various criteria for a particular crop model. The results showed that, for this model, the best criterion is not $MSEP_{\mathrm{cv}}$ but a version of BIC corrected for small samples.

A major advantage of this statistical approach is that it protects against overparametrization. This method is in fact designed to minimize *MSEP* by limiting the number of parameters to estimate.

4.1.5. Bias due to parameter selection

When only a subset of parameters is estimated from data, the estimators are likely to be biased. This problem is often called *compensation*. Two types of bias can occur. The first one is called *omission bias* (Miller, 1990). This bias is due to the fact that some of the model parameters are not estimated from data but are fixed to some pre-defined values. The omission bias is illustrated here with the following model:

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon$$

where x_1 and x_2 are two explanatory variables, and α_1 and α_2 are two parameters. We now assume that the modeler decides to set α_2 equal to zero and to estimate α_1 from N measurements y_i, $i = 1, \ldots, N$. In this case, the ordinary least squares estimate of α_1 is equal to

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{N} x_{1i} y_i}{\sum_{i=1}^{N} x_{1i}^2}.$$

This estimator is biased because

$$E(\hat{\alpha}_1) = \alpha_1 + \frac{\sum_{i=1}^{N} x_{1i} x_{2i}}{\sum_{i=1}^{N} x_{1i}^2} \alpha_2$$

where α_1 and α_2 are the true parameter values. We see that $E(\hat{\alpha}_1) \neq \alpha_1$ if $\alpha_2 \neq 0$. The extent of the bias of $\hat{\alpha}_1$ depends on the true value of the second parameter (α_2). The bias is large when α_2 differs strongly from zero i.e. from the value at which the second parameter was fixed by the modeler. More generally, the omission bias depends on the differences between the values at which the nonestimated parameters are fixed and the true values of these parameters.

The second type of bias is called *selection bias*. This bias occurs only when the same data are used twice, first to select the parameters and then to estimate the parameter values. This is the case for the selection method described in Section 4.1.4. The extent of the selection bias is not well known but, potentially, this bias can have an influence on the accuracy of model predictions (Miller, 1990).

4.2. Application of least squares to dynamic crop models

Once a subset of parameters selected, it is possible to estimate parameter values by least squares, as shown for standard nonlinear regression. Here, we show how to apply this method for estimating the parameters of the dynamic crop model AZODYN. This example is used to discuss different important practical issues.

4.2.1. Example

We consider a version of the AZODYN crop model that includes 18 parameters and three output variables, namely above-ground dry matter (kg ha^{-1}), leaf area index (*LAI*), nitrogen uptake (kg ha^{-1}). These variables are simulated daily between the end of winter and flowering. Table 2 shows the initial values of the parameters. The objective here is to estimate four parameters, namely, *EBMAX* (radiation use efficiency), *D* (ratio of *LAI* to critical level of nitrogen uptake) *K* (radiation extinction coefficient), and *VMAX* (maximal rate of nitrogen uptake). The ranges of possible values of these parameters as deduced from the literature are displayed in Table 2. The other parameters are set equal to values found in the agronomic literature (Jeuffroy and Recous, 1999) (Table 2). Thus, the parameters of the critical nitrogen concentration function (*E*, *F*, *G*, *H*, *L*, *M*, *N*, *P*),

Table 2. Model parameters, initial values, and ranges of variation.

Parameter	Definition	Initial value	Range of variation
EBMAX	Radiation use efficiency	3.3 g MJ^{-1}	1.8–4
K	Radiation extinction coefficient	0.72	0.6–0.8
D	LAI/critical nitrogen uptake	0.028	0.02–0.045
VMAX	Maximal rate of nitrogen uptake	0.5 kg $\mathrm{ha}^{-1}.^{\circ}\mathrm{Cd}^{-1}$	0.2–0.7
C	Photosynthetically active radiation/ global radiation	0.48	
Tmin	Minimal temperature for photosynthesis	0°C	
Topt	Optimal temperature for photosynthesis	15°C	
Tmax	Maximal temperature for photosynthesis	40°C	
EIMAX	Ratio of intercepted to incident radiation	0.96	
Tep-flo	Sum of temperature between earing and flowering	150°Cd	
E	Parameter of the critical nitrogen concentration function	1.55 t ha^{-1}	
F	Parameter of the critical nitrogen concentration function	4.4%	
G	Parameter of the critical nitrogen concentration function	5.35%	
H	Parameter of the critical nitrogen concentration function	−0.442	
L	Parameter of the maximal nitrogen concentration function	2 t ha^{-1}	
M	Parameter of the maximal nitrogen concentration function	6%	
N	Parameter of the maximal nitrogen concentration function	8.3%	
P	Parameter of the maximal nitrogen concentration function	−0.44	

the parameters affecting photosynthesis (*TMIN*, *TOPT*, *TMAX*), and the parameters *C* and *Tep-flo* were previously estimated in past studies and are not reestimated here.

The four parameters are estimated by using dry matter, nitrogen uptake, and *LAI* measurements obtained in Grignon (Paris Basin, France) during six years (1992, 1995, 1996, 1998, 1999, and 2001). Measurements were obtained each year at 8–18 different dates between end of winter and flowering. Three replicates were performed at each date of measurements. The variances of the dry matter measurements obtained in 1995 are displayed in Figure 6. Each variance was calculated from three replicates. Figure 6 shows that the variances are very heterogeneous and tend to increase with time. Figure 7 presents the measurements averaged over replicates for the year 1999.

In order to take variance heterogeneity into account, we estimate the parameters by using weighted least squares. We note y_{ij}^{dm}, y_{ij}^{nu}, y_{ij}^{l} are the measurements of dry matter, nitrogen uptake, and *LAI* averaged over replicates for year i, $i = 1, \ldots, 6$, and time t_j,

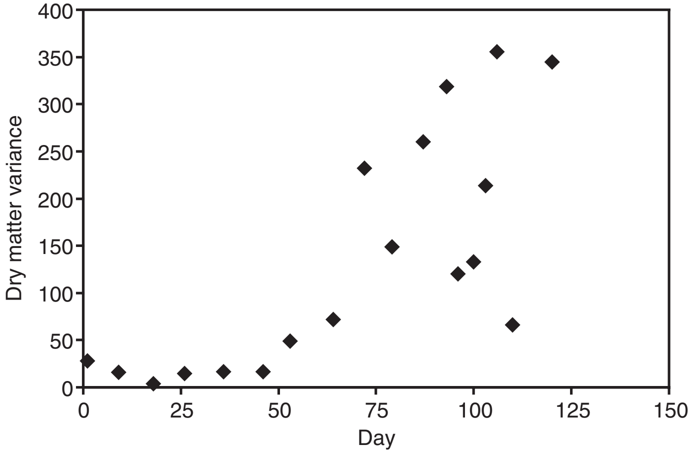


Figure 6. Variances of winter wheat dry matter measurements (kg^2 ha^{-2}) obtained in Grignon in 1995 at different dates between end-of-winter and flowering.

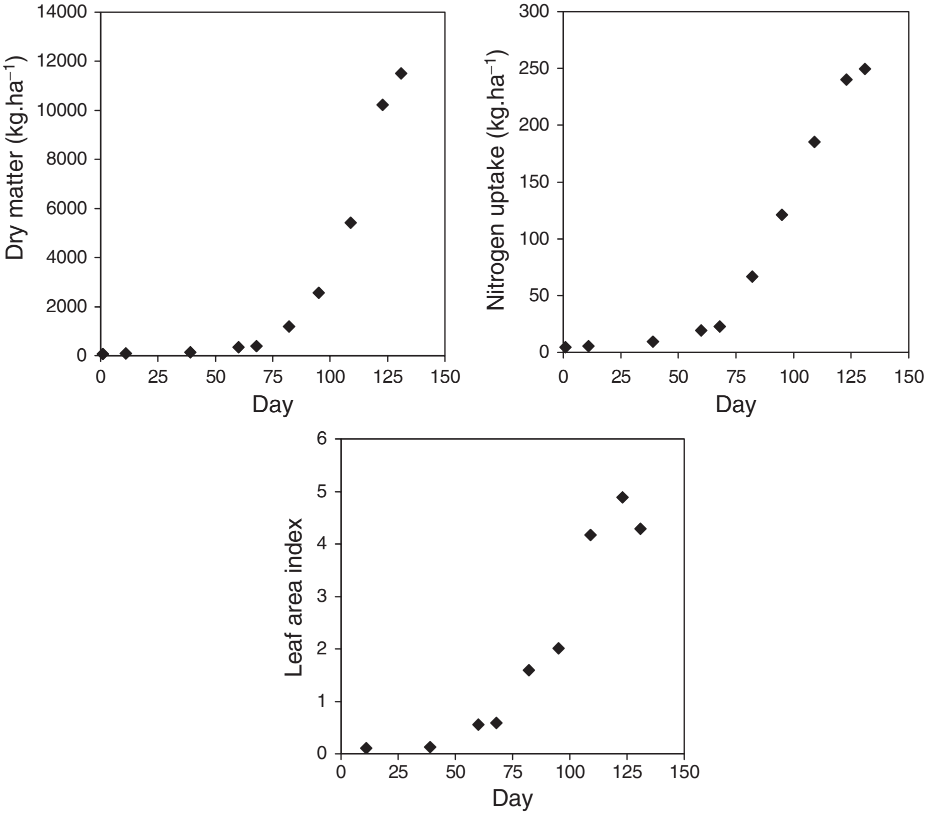


Figure 7. Measurements of dry matter, nitrogen uptake, and LAI obtained in Grignon in 1999. Each point represent an average of three replicates.

$j = 1, \ldots, N_i$. The four parameters are estimated by minimizing:

$$Z_{\mathrm{WLS}}(\theta) = \sum_{i=1}^{6}\sum_{j=1}^{N_i}\frac{[y_{ij}^{DM} - f^{DM}(x_{ij};\theta)]^2}{\hat{var}(y_{ij}^{DM})} + \sum_{i=1}^{6}\sum_{j=1}^{N_i}\frac{[y_{ij}^{NU} - f^{NU}(x_{ij};\theta)]^2}{\hat{var}(y_{ij}^{NU})}$$

$$+\sum_{i=1}^{6}\sum_{j=1}^{N_i}\frac{[y_{ij}^{l} - f^{l}(x_{ij};\theta)]^2}{\hat{var}(y_{ij}^{l})}$$

where $f^{DM}(x_{ij};\theta)$, $f^{NU}(x_{ij};\theta)$, $f^{l}(x_{ij};\theta)$ are the simulated values of dry matter, nitrogen uptake, and *LAI* for year *i*, $i = 1, \ldots, 6$ and time t_j, $j = 1, \ldots, N_i$. $\mathrm{v\hat{a}r}(y_{ij}^{DM})$, $\mathrm{v\hat{a}r}(y_{ij}^{NU})$, $\mathrm{v\hat{a}r}(y_{ij}^{l})$ are the empirical variances (calculated from the replicates) for each of the three types of measurements at time t_j, $j = 1, \ldots, N_i$. The variance $\mathrm{v\hat{a}r}(y_{ij}^{s})$ is calculated as

$$\mathrm{v\hat{a}r}(y_{ij}^{s}) = \frac{1}{R(R-1)}\sum_{k=1}^{R}[y_{ijk}^{s} - y_{ij}^{s}]^2$$

where y_{ijk}^{s} is the *k*th replicate of measurement *s* at year *i* and date t_j and $R = 3$. θ is the vector including the four parameters: $\theta = [EBMAX, K, D, VMAX]^T$. Note that this method adequately accounts for differences in magnitude of the different variables.

$Z_{\mathrm{WLS}}(\theta)$ is minimized by using the S-PLUS function *NLS*. Before applying *NLS*, observations and model predictions were divided by the empirical variances calculated from replicates. This is a way to apply weighted least squares using an ordinary least squares algorithm. Initial and estimated parameter values are reported in Table 3.

Except for *VMAX*, the results show that the estimated parameter values do not differ much from their initial values. Table 4 shows the root mean squared error (*RMSE*) values obtained with initial and estimated parameter values for dry matter, nitrogen uptake and *LAI*. The RMSE values for dry matter and nitrogen uptake are lower when the parameters are estimated by weighted least squares. The *RMSE* values obtained for *LAI* with initial and estimated parameters are similar.

Table 3. Initial parameter values and values estimated by weighted least squares. The standard errors associated with the parameter estimates are presented between brackets.

Parameter	Initial value	Estimated value
EBMAX (g MJ^{-1})	3.3	3.29 (0.11)
K	0.72	0.74 (0.06)
D	0.028	0.028 (0.001)
VMAX (kg ha^{-1}°Cd^{-1})	0.5	0.38 (0.02)

Table 4. RMSE values obtained with initial parameter values and with values estimated by weighted least squares.

Parameter values	RMSE		
	Dry matter (kg ha^{-1})	Nitrogen uptake (kg ha^{-1})	LAI
Initial values	614.9	34.7	0.41
Estimated values	607.2	27.7	0.41

4.2.2. Constrained parameter estimation

As crop model parameters are physically or biologically interpretable, it is sometimes useful to constrain parameter values between lower and upper limits, $\theta_\text{l} \leq \theta \leq \theta_\text{u}$. These limits are set after considering the ranges of values commonly reported in the literature. Algorithms in commercial softwares often let the user set limits. Modelers can also implement constrained parameter estimation themselves by using a transformed model noted $f(x; \theta_{NEW})$. $f(x; \theta_{NEW})$ is determined by expressing the initial parameters θ as

$$\theta = \frac{\theta_\text{u} \exp(\theta_{NEW}) + \theta_\text{l}}{1 + \exp(\theta_{NEW})}.$$

Then, an unconstrained estimation of θ_{NEW} is performed. As exp(.) > 0, the transformation ensures that $\theta_l \leq \theta \leq \theta_u$ during the unconstrained estimation of θ_{NEW}.

Constrained estimation was not necessary with AZODYN because the estimated values reported in Table 3 fall within the limits defined in Table 2. An application of constrained estimation is presented by Vold et al. (1999).

4.2.3. Problem related to sequential estimation of groups of parameters

A widespread approach for estimating parameters of crop models consists in estimating sequentially groups of parameters with different types of measurements. This approach is applied by Mavromatis et al. (2001) for estimating the parameters of the CROPGRO-Soybean model. In the study, a first group of two parameters is estimated using flowering date measurements. Next, a second group of three parameters is estimated using maturity date measurement. Finally, a third group of two parameters is estimated from yield measurements.

The sequential approach is a simple way to take into account several types of measurements. Another interest is that, as only a small number of parameters are estimated at each step, the implementation of this method is generally numerically feasible. However, this method has several important drawbacks. In most of the crop models, a given parameter has an influence on several output variables. In such cases, it is not natural to use only one type of measurement for estimating the parameter value. It seems more logical to use all types of measurements available. Another problem is that the results of the sequential method generally depend on the sequence of the estimation of the different groups of parameters. This point is illustrated below with the AZODYN crop model.

We estimate sequentially the parameters *EBMAX*, *D*, *K*, and *VMAX* with the data described in Section 4.2.1. Some of these parameters influence several output variables. For example, *EBMAX* and *K* influence the three output variables of the model, namely, dry matter, nitrogen uptake, and *LAI*. We define here three groups of parameters and estimate sequentially each group with a particular type of data. The three groups are: $\theta_1 = [EBMAX, K]^{\mathrm{T}}$, $\theta_2 = [VMAX]$, and $\theta_3 = [D]$. First, θ_1 is estimated with dry matter measurements by minimizing:

$$Z_{\mathrm{OLS}}(\theta_1) = \sum_{i=1}^{6} \sum_{j=1}^{N_i} [y_{ij}^{DM} - f^{DM}(x_{ij}; \theta_1, \theta_2, \theta_3)]^2$$

with θ_2 and θ_3 fixed to their initial values (Table 2). Second, θ_2 is estimated with nitrogen uptake measurements by minimizing:

$$Z_{\mathrm{OLS}}(\theta_2) = \sum_{i=1}^{6} \sum_{j=1}^{N_i} [y_{ij}^{NU} - f^{NU}(x_{ij}; \hat{\theta}_1, \theta_2, \theta_3)]^2$$

where $\hat{\theta}_1$ is the parameter vector estimated at Step 1 and θ_3 is fixed to its initial value. Finally, θ_3 is estimated with measurements of *LAI* by minimizing:

$$Z_{\mathrm{OLS}}(\theta_3) = \sum_{i=1}^{6} \sum_{j=1}^{N_i} [y_{ij}^{l} - f^{l}(x_{ij}; \hat{\theta}_1, \hat{\theta}_2, \theta_3)]^2$$

where $\hat{\theta}_2$ is the value of θ_2 estimated at Step 2. The following parameter values are obtained: $\widehat{EBMAX} = 3.05$, $\hat{K} = 0.96$, $\hat{D} = 0.028$, and $\widehat{VMAX} = 0.38$. The same procedure is applied a second time with three other groups of parameters: $\theta_1 = [EBMAX]$, $\theta_2 = [VMAX]$, $\theta_3 = [K, D]^{\mathrm{T}}$. With these new groups, the parameter *K* is estimated from *LAI* measurements and not from dry matter measurements as before. We now obtain the following estimated parameter values: $\widehat{EBMAX} = 3.34$, $\hat{K} = 0.39$, $\hat{D} = 0.035$, and $\widehat{VMAX} = 0.39$. These new values differ strongly from the previous estimates. This result shows that the result of a sequential procedure depends on the sequence of the different estimations and on the type of data used for each estimation. It is much more desirable to estimate all parameters simultaneously by using the weighted least squares method as shown in Section 4.2.1.

4.2.4. How to take into account the correlation of residual errors?

In the previous section, we estimated some of the parameters of AZODYN by using the weighted least squares method with weights calculated from the variances of the measurement replicates. This method gives unbiased parameter estimators with minimum variances if the variances of the model errors are proportional to the variances of the measurement replicates and if the model errors are independent. It is important to check these assumptions. We show below how to check graphically the hypothesis of "independently distributed model errors" for AZODYN.

In Section 4.2.1, four parameters of AZODYN were estimated from six years of data. In order to study the structure of the model residuals, the differences between observations and predictions are now calculated for 1992 and 1995. Model predictions are computed by using the estimated parameter values given in Table 3. The residuals for dry matter predictions for years 1992 and 1995 are displayed in function of time in Figure 8. Different symbols are used for the two years. Obviously, the residuals are not homogeneously distributed. Almost all the residuals obtained in 1995 are positive whereas all the residuals obtained in 1992 are negative. In other words, dry matter is overestimated in 1992 and underestimated in 1995. The hypothesis of "independently distributed model errors" is not realistic here.

This problem is quite general. Correlations between model residuals often arise when several measurements are performed at different dates in a given site-year. Site-year characteristics have a strong influence on observations and, as only a part of the between site-years variability can be predicted by crop models, model residuals obtained in a given site-year are often correlated. For example, when a model overestimate dry matter at a given date and a certain site-year, because of some local factor not accounted for, it is likely that predictions obtained for the same site-year at different dates will also result in overestimation.

Correlation between residuals is an important problem with no simple solution. When model errors are correlated, the application of ordinary or weighted least squares leads to estimators that are unbiased but not of minimum variance. Several methods were developed by statisticians for taking into account correlations but the application of these methods to dynamic crop models is impractical in most cases.

Generalized least squares was specifically developed to take into account correlated model errors. Its implementation supposes the estimation of a matrix including *N* variances

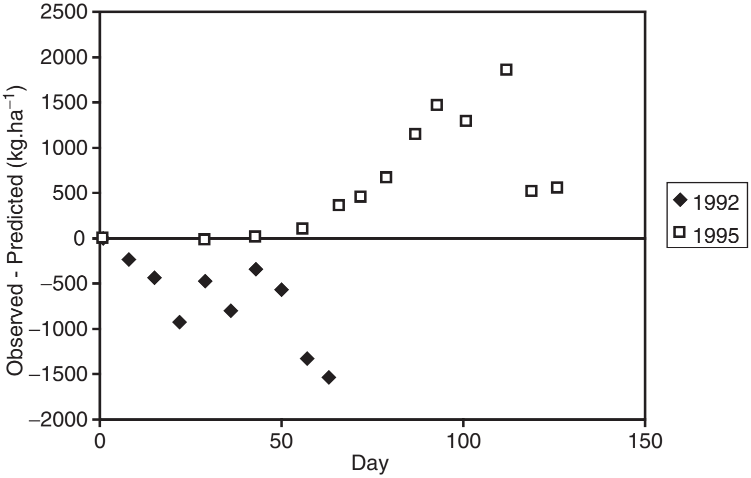


Figure 8. Dry matter residuals (observed–predicted) obtained with AZODYN for two years.

and $N(N-1)/2$ covariances, where N is the number of measurements. With crop models, the dimension of this matrix is usually very high. If we consider our example based on the AZODYN crop model, the number of measurements (averaged over replicates) available for estimating the parameters is equal to 201. Consequently, the variance–covariance matrix of the model errors is 201 by 201 matrix including 201 variances and 20 100 covariances. The estimation of these elements directly from the model residues will lead to inaccurate estimates and may cause numerical problems. A solution is to try to simplify this matrix by assuming many covariances equal to zero. Figure 8 shows that the residuals obtained in the same year are strongly correlated. However, we did not find any evidence of correlation between residuals obtained in different years, or between residuals obtained in the same year but for different types of measurements. Thus, it seems reasonable to estimate only the covariances between residuals obtained at the same year and for the same type of measurement, and to set the other covariances equal to zero. The problem is now reduced to the estimation of eighteen (3 types of measurement × 6 years of experiment) N_i by N_i variance–covariance matrix, where N_i is the number of dates of measurements obtained in year i (N_i is in the range 8–18). The total number of nonzero covariance terms is equal to 1188. This is much lower than 20 100 but still quite large. If all measurements had been performed at the same dates every year, it would have been possible to reduce the number of elements further by assuming for each type of measurement that $\mathrm{cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \mathrm{cov}(\varepsilon_{i'j}, \varepsilon_{i'j'})\ \forall i, i' \in [1, \ldots, 6], \forall j, j' \in [1, \ldots, N_i]$ where i and j are the indices of years and dates of measurement respectively. In our example, the dates of measurement were not the same every year and, as a consequence, this approach cannot be applied here.

4.2.5. Mixed-effect model

The use of a mixed-effect model is a parsimonious way to take into account correlation between residuals (e.g. Davidian and Giltinan, 1995). The general principle is to define some of or all the model parameters as random variables. The probability distribution of these parameters describes the between site-year variability of the parameter values. One important advantage of this method is that the number of estimates to be made from data is relatively low. With the generalized least squares method, the covariances between pairs of error terms have to be estimated. The number of estimates becomes very high as soon as more than few measurements are included in the data set. In a mixed-effect model, only the expected values, variances and covariances of the model parameters have to be estimated.

We give a simple example here. Suppose an observation y is related to time t through the following linear model:

$$y_{ij} = \alpha_i + \beta_i t_{ij} + e_{ij}$$

where y_{ij} is the jth measurement obtained on the ith site-year at time t_{ij}. α_i and β_i are the two parameters of the regression for site-year i. We assume that the linear model holds for all site-years, but that the values of the two parameters vary between site-years. We also assume that α_i and β_i are independent and are normally distributed, $\alpha_i \sim N(\mu_\alpha, \sigma_\alpha^2)$ and $\beta_i \sim N(\mu_\beta, \sigma_\beta^2)$. The expected values, μ_α and μ_β, represent the average values of the parameters in the population of site-years considered. The variances, σ_α^2 and σ_β^2, provide information on the variability of parameters values across site-years. Finally, we assume that the within site-year model error e_{ij} is distributed as $e_{ij} \sim N(0, \sigma_e^2)$ and is independent from α_i and β_i.

In this model, $\mu_\alpha + \mu_\beta t$ represents the "average response" of y for all site-years. Consequently, the difference between an observation y_{ij} and the average response is defined by:

$$\varepsilon_{ij} = y_{ij} - (\mu_\alpha + \mu_\beta t_{ij}) = \alpha_i + \beta_i t_{ij} + e_{ij} - (\mu_\alpha + \mu_\beta t_{ij}).$$

Now, consider another measurement, $y_{ij'}$, obtained in the same site-year i, but at a different date $t_{ij'}$. The covariance of ε_{ij} and $\varepsilon_{ij'}$ is defined by

$$\operatorname{cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \frac{\operatorname{var}(\varepsilon_{ij} + \varepsilon_{ij'}) - \operatorname{var}(\varepsilon_{ij}) - \operatorname{var}(\varepsilon_{ij'})}{2}.$$

As $\operatorname{var}(\varepsilon_{ij}) = \sigma_\alpha^2 + \sigma_\beta^2 t_{ij}^2 + \sigma_e^2$ and $\operatorname{var}(\varepsilon_{ij} + \varepsilon_{ij'}) = 4\sigma_\alpha^2 + \sigma_\beta^2 (t_{ij} + t_{ij'})^2 + 2\sigma_e^2$, we obtain $\operatorname{cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma_\alpha^2 + t_{ij} t_{ij'} \sigma_\beta^2$.

This result shows that the covariance of two error terms ε_{ij} and $\varepsilon_{ij'}$ depends only on two parameters, namely, σ_α and σ_β. In other words, the covariances of all pairs of error terms can be directly calculated from σ_α and σ_β, and the problem of the estimation of the variance–covariance matrix is reduced to the estimation of only these two parameters.

It was shown that, in some cases, the use of a mixed-effect model improves the accuracy of the estimation compared to least squares (Davidian and Giltinan, 1995; Pinheiro and Bates, 2000). Another argument in favor of this type of model is given by Makowski and Wallach (2002). These authors showed that mixed-effects models lead to better nitrogen fertilizer recommendations than fixed parameter models. The quality of model-based decision rules appears to be improved by using random parameters. But some important numerical problems may arise when implementing this method with complex nonlinear models. This is illustrated in Makowski et al. (2001). The authors considered a simple static model including only 10 parameters and showed that it was not possible to estimate a full 10 × 10 variance–covariance matrix of parameters. It was necessary to simplify the matrix. As far as we know, this approach has never been applied to a complex dynamic crop model.

4.3. Bayesian methods

4.3.1. Introduction

Bayesian methods are becoming increasingly popular for estimating parameters of complex mathematical models (e.g. Campbell et al., 1999). This is because the Bayesian approach provides a coherent framework for dealing with uncertainty. This is also due to the increase in the speed of computer calculation and the recent development of new algorithms (Malakoff, 1999).

The principle is to start with a prior probability distribution of the model parameters whose density is noted $P(\theta)$. This prior distribution describes our belief about the parameter values before we observe the set of measurements Y. In practice, $P(\theta)$ is based on past studies, expert knowledge, and literature. The Bayesian methods then tell us how to update this belief about θ using the measurements Y to give the posterior parameter density $P(\theta|Y)$ (density of θ conditional on the data Y). What we now believe about θ is captured in $P(\theta|Y)$.

All the estimation methods described in Sections 3 and 4.2 are called frequentist methods. With these methods, the parameters θ are fixed, but the parameter estimators are random because they depend on observations. The variances of these estimators can also be computed (for example, the variance of an ordinary least squares estimator) and reflect the variability of the data we might have observed in other samples (see Section 3.5 for an illustration).

In the Bayesian approach, the parameters are defined as random variables and the prior and posterior parameter distributions represent our belief about parameter values before and after data observation. This approach has several advantages:

- parameters can be estimated from different types of information (data, literature, expert knowledge);
- the posterior probability distribution can be used to implement uncertainty analysis methods (see Chapter 3);
- the posterior probability distribution can be used for optimizing decisions in face of uncertainty (see Chapter 6).

The purpose of the Bayesian methods presented in this section is to describe $P(\theta|Y)$. We can see that, with simple models, it is possible to determine the analytical expression of the posterior density but that, in most cases, $P(\theta|Y)$ can only be approximated.

4.3.2. Example

The practical interest of a Bayesian approach for estimating parameters is illustrated here in a simple example. Suppose we want to estimate yield for a given field. The unknown true yield value is noted θ. Two types of information are available for estimating θ. The first type of information comes from an expert. According to this expert, the approximate yield value must be $\mu = 5$ t ha^{-1} and the uncertainty about yield value is $\tau = 2$ t ha^{-1}. μ and τ are used to define the *prior* yield distribution, $\theta \sim N(\mu, \tau^2)$. The second type of information is an imperfect yield measurement performed in a small plot within the field. We assume that this measurement is normally distributed and is unbiased i.e. the expected value of the measurement is equal to θ. Under this assumption, the distribution of the measurement is defined by $Y|\theta \sim N(\theta, \sigma^2)$ where Y is the measurement and σ^2 is the variance of the measurement. We assume that σ is known with $\sigma = 1$ t ha^{-1}. Note that $P(Y|\theta)$ represents the likelihood of θ.

Our objective is to determine the analytical expression of the *posterior* distribution of θ. For this, we first derive the joint probability distribution of $\begin{pmatrix}\theta \\ Y\end{pmatrix}$. As $\theta \sim N(\mu, \tau^2)$ and $Y|\theta \sim N(\theta, \sigma^2)$, we have

$$Y \sim N(\mu, \tau^2 + \sigma^2)$$

and

$$\begin{pmatrix}\theta\\ Y\end{pmatrix} \sim N\left[\begin{pmatrix}\mu\\ \mu\end{pmatrix}, \begin{pmatrix}\tau^2 & \tau^2\\ \tau^2 & \tau^2+\sigma^2\end{pmatrix}\right].$$

The posterior distribution is then derived using the following property (e.g. Saporta, 1990):

$$\text{if } \begin{pmatrix}A\\ B\end{pmatrix} \sim N\left\{\begin{bmatrix}E\,(A)\\ E\,(B)\end{bmatrix}, \begin{bmatrix}\operatorname{var}(A) & \operatorname{cov}(A,B)\\ \operatorname{cov}(A,B) & \operatorname{var}(B)\end{bmatrix}\right\}$$

then $A|B$ is normally distributed and

$$E(A|B) = E(A) + \frac{\operatorname{cov}(A,B)}{\operatorname{var}(B)}[B - E(B)]$$

$$\operatorname{var}(A|B) = \operatorname{var}(A) - \frac{\operatorname{cov}(A,B)^2}{\operatorname{var}(B)}.$$

We obtain

$$E(\theta|Y) = \mu + \frac{\tau^2}{\tau^2+\sigma^2}(Y-\mu)$$

and

$$\operatorname{var}(\theta|Y) = \tau^2 - \frac{\tau^4}{\tau^2+\sigma^2}.$$

Finally, the posterior distribution can be expressed as:

$$\theta|Y \sim N\left\lfloor(1-B)\mu + BY, (1-B)\tau^2\right\rfloor \tag{11}$$

where $B = \tau^2/(\tau^2+\sigma^2)$. According to Eq. (11), $E(\theta|Y)$ is a weighted sum of the prior mean and of the measurement. The weight B depends on the prior variance τ^2 and on the variance of the error of measurement σ^2. As $\tau = 2$ t ha^{-1} and $\sigma = 1$ t ha^{-1}, we have $B = 4/5$, $E(\theta|Y) = 1/5\mu + 4/5Y = 1 + 4/5Y$, and $\operatorname{var}(\theta|Y) = 4/5$.

Let us compare two different estimates for θ according to their properties. The first one is $\hat{\theta}_{\mathrm{ML}} = Y$. This is in fact the maximum likelihood estimate of θ. The second estimate is the expected value of the posterior distribution, $\hat{\theta}_B = 1 + 4/5Y$, calculated from Eq. (11). The first estimate depends only on the observation whereas the second one depends on both the data and the prior parameter distribution. It is interesting to note that, contrary to $\hat{\theta}_{\mathrm{ML}}$, $\hat{\theta}_B$ is biased because $E_{Y|\theta}(\hat{\theta}_B) = 1 + \frac{4}{5}\theta$. So $E_{Y|\theta}(\hat{\theta}_B) \neq \theta$ when $\theta \neq 5$. The advantage of using $\hat{\theta}_B$ is that its variance – $\operatorname{var}_{Y|\theta}(\hat{\theta}_B) = \frac{16}{25}\sigma^2 = \frac{16}{25}$ – is lower than the variance of $\hat{\theta}_{\mathrm{ML}}$, $\operatorname{var}_{Y|\theta}(\hat{\theta}_{\mathrm{ML}}) = 1$.

4.3.3. Computation of the posterior mode

Because crop models are very complex, it is impossible to derive an analytical expression of $P(\theta|Y)$ but, under some assumptions, it is possible to calculate its mode. This method returns only a single value for each parameter, the value maximizing $P(\theta|Y)$.

Suppose that p parameters, $\theta = (\theta_1, \dots, \theta_p)^T$, have to be estimated. Define the Normal prior parameter density as

$$P(\theta) = (2\pi)^{-p/2}|\Omega|^{-1/2} \exp\left\{-\frac{1}{2}[\theta - \mu]^T \Omega^{-1}[\theta - \mu]\right\},$$

where $\mu = (\mu_1, \dots, \mu_p)^T$ is the $(p \times 1)$ vector of prior means and Ω is the $(p \times p)$ variance–covariance matrix.

Suppose that N observations, $Y = (y_1, \dots, y_N)^T$, are available for estimating parameters and that these observations are normally distributed. The likelihood is then defined as

$$P(Y|\theta) = (2\pi)^{-N/2}|V|^{-1/2} \exp\left\{-\frac{1}{2}[Y - F(\theta)]^T V^{-1}[Y - F(\theta)]\right\}$$

where $F(\theta)$ is a vector including the N model predictions, $F(\theta) = [f(x_1;\theta), \dots, f(x_N;\theta)]^T$, and V is $(N \times N)$ variance–covariance matrix of the model errors.

According to the Bayes theorem, the posterior distribution $P(\theta|Y)$ is related with $P(Y|\theta)$ and $P(\theta)$ as follows:

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \tag{12}$$

where $P(Y)$ is the distribution of the observations and is independent of the parameters.

If the matrix V is known, we get from Eq. (12)

$$P(\theta|Y) = K_1 \exp\left\{-\frac{1}{2}[Y - F(\theta)]^T V^{-1}[Y - F(\theta)]\right\} \exp\left\{-\frac{1}{2}[\theta - \mu]^T \Omega^{-1}[\theta - \mu]\right\}$$

where K_1 is a constant independent of θ. The posterior mode is the value of θ maximizing $P(\theta|Y)$ or maximizing $\log P(\theta|Y)$ where $\log P(\theta|Y)$ is expressed as

$$\log P(\theta|Y) = K_2 - [Y - F(\theta)]^T V^{-1}[Y - F(\theta)] - [\theta - \mu]^T \Omega^{-1}[\theta - \mu]$$

where K_2 is a constant independent of θ. Consequently, the posterior mode is the value of θ that minimizes

$$[Y - F(\theta)]^T V^{-1}[Y - F(\theta)] + [\theta - \mu]^T \Omega^{-1}[\theta - \mu] \tag{13}$$

Equation (13) includes two terms. The first term, $[Y - F(\theta)]^T V^{-1}[Y - F(\theta)]$, is equal to the function minimized by the generalized least squares estimate ($Z_{\text{GLS}}(\theta)$).

The second term, $[\theta - \mu]^T \Omega^{-1} [\theta - \mu]$, is a penalty term that penalizes the parameter values that differ strongly from the prior mean μ. When the observations are mutually independent and so are the parameters, the matrices V and Ω are diagonal and Eq. (13) is equal to

$$\sum_{i=1}^{N} \frac{[y_i - f(x_i;\theta)]^2}{\sigma_i^2} + \sum_{j=1}^{p} \frac{[\theta_j - \mu_j]^2}{\omega_i^2} \tag{14}$$

where σ_i^2, $i = 1, \ldots, N$, and ω_j^2, $j = 1, \ldots, p$ are the diagonal elements of V and Ω. If ω_j^2, $j = 1, \ldots, p$, take very small values, the parameter values minimizing Eq. (14) will not differ much from the prior mean μ.

The minimization of Eq. (13) or Eq. (14) can easily be performed with the same algorithms as those used to apply generalized least squares. The trick is to consider the prior mean μ as p additional data and then to implement the generalized least squares method. The main drawback of this method is that it provides only the posterior mode and not the whole posterior parameter distribution.

In a recent study, Tremblay and Wallach (2004) studied the interest of using the posterior mode as an estimator. The authors considered a model that is part of the STICS model (Brisson et al., 1998), which we refer to as Mini-STICS. Mini-STICS includes 14 parameters and simulates sunflower development over a period of 20 days, starting at the stage Maximal Acceleration of Leaf growth (AMF). Tremblay and Wallach (2004) compared generalized least squares and a Bayesian approach that consists in minimizing Eq. (13). Generalized least squares was applied to estimate a small number of parameter (1–7) selected by using statistical methods of the type presented in Section 4.1.4. The other parameters were fixed at their initial values. With the Bayesian approach, all 14 parameters were estimated simultaneously. The authors applied the two types of estimation method to several training data sets each with 14 observations and calculated MSEP values for different model output variables (*LAI* and soil water content, each at two dates). The results showed that the MSEP values were lower with the Bayesian approach than with generalized least squares.

4.3.4. Prior distribution for the variance–covariance matrix of the errors

In practice, it is often difficult to give a value to the variance–covariance matrix of the model errors V. Then, it is useful to estimate the elements of V at the same time as the model parameters θ.

Different types of prior distribution can be used for V but, when no information about V is available, it is convenient to define a noninformative prior density function for V, for example, the Jeffreys distribution $P(V) = K|V|^{-(N+1)/2}$, where $|V|$ is the determinant of V and K a constant.

The posterior mode is then calculated by maximizing

$$P(\theta, V|Y) = \frac{P(Y|\theta, V)P(\theta, V)}{P(Y)} = \frac{P(Y|\theta, V)P(\theta)P(V)}{P(Y)}$$

or, equivalently, by minimizing

$$
\begin{aligned}
-\log P(\theta, V|Y) = R + \left(\frac{N}{2}+1\right)\log|V| + [Y - F(\theta)]^T V^{-1}[Y - F(\theta)] \\
+ [\theta - \mu]^T \Omega^{-1}[\theta - \mu]
\end{aligned}
\tag{15}
$$

where R is independent from V and θ.

As already explained in Section 4.2.4, the number of nonzero elements in V can be very large when the model errors are correlated. In such cases, the minimization of Eq. (15) is often difficult and the parameter estimates may be inaccurate.

When the observations are mutually independent and so are the parameters, the matrices V and Ω are diagonal and the Jeffrey's prior density function is

$$P(V) = K\frac{1}{\sigma_1^2 \times \cdots \times \sigma_i^2 \times \cdots \times \sigma_N^2}.$$

The posterior mode is then obtained by minimizing $-\log P(\theta, V|Y)$ with

$$-\log P(\theta, V|Y) = R + \frac{3}{2}\sum_{i=1}^{N}\log(\sigma_i^2) + \sum_{i=1}^{N}\frac{[y_i - f(x_i;\theta)]^2}{\sigma_i^2} + \sum_{j=1}^{p}\frac{[\theta_j - \mu_j]^2}{\omega_j^2}. \tag{16}$$

4.3.5. Monte Carlo methods

The principle here is to generate a random sample of parameter values from which one derives an approximate of the posterior distribution. The interest of this approach is that it can be applied to complex nonlinear models including a large number of parameters. For example, Harmon and Challenor (1997) use a Monte Carlo method for estimating 10 parameters of a complex ecological model. Monte Carlo methods are probably the most promising methods for estimating parameters of complex nonlinear models. According to an article in *Science* (Malakoff, 1999), these methods explain the new popularity of the Bayesian methods. A detailed description of Monte Carlo methods can be found, for example, in Geyer (1992), in Gilks et al. (1995), in Carlin and Louis (2000), and in Theobald and Talbot (2002). Here, we only briefly present two of these methods, importance sampling and the Metropolis–Hastings algorithm.

The importance sampling method requires the definition of a density function $g(\theta)$, called the importance function, from which realizations of θ can be sampled. This function is used to generate a sample of Q parameter vectors denoted θ_q, $q = 1, \ldots, Q$. A weight ω_q is then calculated for each generated vector as

$$\omega_q = \frac{P(Y|\theta_q)P(\theta_q)}{g(\theta_q)}.$$

The weight values are then used to describe the posterior parameter distribution. For example, the expected value of the posterior distribution of θ is approximated by:

$$\hat{E}(\theta|Y) = \frac{\sum_{q=1}^{Q} \omega_q \theta_q}{\sum_{q=1}^{Q} \omega_q}.$$

Importance sampling gives good results if the importance function $g(\theta)$ is not too different than $K\ P(Y|\theta)P(\theta)$ where K is a normalization constant. A particular importance function is the prior density function. In this case, we have $g(\theta) = P(\theta)$ and $\omega_q = P(Y|\theta_q)$. This importance function was used by Makowski et al. (2002) for estimating the parameters of a complex nonlinear model. The estimation method GLUE (generalized likelihood uncertainty estimation, e.g. Shulz et al., 1999) can be seen as a version of importance sampling with prior density as the importance function and with a particular type of likelihood function.

The Metropolis–Hastings algorithm is a Markov chain Monte Carlo algorithm (MCMC) (see, for example, Geyer 1992 and Gilks et al. 1995, for more details). The objective of this method is to randomly generate a sample of parameter values from the posterior parameter distribution. The algorithm is iterative and starts with an initial parameter vector θ_0. A series of Q vectors θ_q, $q = 1, \ldots, Q$, is then generated as follows:

i. Generate a candidate vector θ^* from a proposal distribution denoted as $P(\theta^*|\theta_{q-1})$, for instance a normal distribution with mean equal to θ_{q-1}.

ii. Calculate

$$T = \frac{P(Y|\theta^*)P(\theta^*)P(\theta_{q-1}|\theta^*)}{P(Y|\theta_{q-1})P(\theta_{q-1})P(\theta^*|\theta_{q-1})}.$$

iii. If min $(1,T) > u$, where u is drawn from a uniform distribution on the interval (0,1) then $\theta_q = \theta^*$ otherwise $\theta_q = \theta_{q-1}$.

After a phase of say M iterations, the chain of values $\theta_1, \theta_2, \ldots$ thus constructed will converge to a chain with elements drawn from the posterior parameter distribution. The first M iterations should be discarded. Before using the Metropolis–Hastings algorithm, it is necessary to choose the starting value θ_0, the proposal distribution $P(\theta^*|\theta_{q-1})$, the total number of iterations Q and the number of discarded iterations M. The definition of precise rules for choosing these elements is currently an area of active research. According to Gilks et al. (1995), the choice of θ_0 is not very critical. On the other hand, the choice of the proposal distribution $P(\theta^*|\theta_{q-1})$ is an important issue. A common practice is to use a normal distribution with mean θ_{q-1} and constant covariance matrix Σ. This is equivalent to assuming that $\theta^*|\theta_{q-1} \sim N(\theta_{q-1}, \Sigma)$. Several authors (Campbell et al., 1999; Harmon and Challenor, 1997) suggest choosing Σ such that the acceptance rate of the test performed in Step iii of the algorithm is in the range 20–70%. Several methods for determining M (number of iterations to be discarded) and Q (total number of iterations) are presented in Geyer (1992), Gilks et al. (1995) and Carlin and Louis (2000). Makowski et al. (2002) compared the GLUE method and the Metropolis–Hastings algorithm in a simulation study and showed that the latter gives slightly better results in terms of MSEP values.

4.3.6. Application 1: Yield estimation

We consider again the problem of estimating crop yield as described in Section 4.3.2. The Metropolis–Hastings algorithm is used to approximate the posterior distribution Eq. (11). Note that the use of this algorithm is not of practical interest in this context because the analytical expression of the posterior distribution is known here. This example shows a simple application of the algorithm.

We assume here that the prior distribution for the yield value is $\theta \sim N(\mu, \tau^2)$ where $\mu = 6$ t ha^{-1} and $\tau = 3$ t ha^{-1}, that the yield value measured in the field is equal to $Y = 10$ t ha^{-1} and that its standard error is equal to $\sigma = 0.5$ t ha^{-1}. According to Eq. (11), the posterior distribution is then $\theta | Y \sim N(9.892, 0.243^2)$.

We now apply the Metropolis–Hastings algorithm and demonstrate that it gives a good approximation of the posterior distribution. The first value of the chain is set equal to the prior mean μ. The algorithm is then implemented as follows:

i. Generate a candidate vector θ^* from a proposal distribution denoted as $\theta^* | \theta_{q-1} \sim N(\theta_{q-1}, \eta^2)$

ii. Calculate

$$T = \frac{P(Y|\theta^*)P(\theta^*)P(\theta_{q-1}|\theta^*)}{P(Y|\theta_{q-1})P(\theta_{q-1})P(\theta^*|\theta_{q-1})}.$$

Here, we have $P(\theta^*|\theta_{q-1}) = P(\theta_{q-1}|\theta^*)$ because the proposal distribution is normal. Consequently, T can be expressed as

$$T = \frac{P(Y|\theta^*)P(\theta^*)}{P(Y|\theta_{q-1})P(\theta_{q-1})}$$

with

$$P(\theta^*) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left[-\frac{1}{2}\left(\frac{\theta^* - \mu}{\tau^2}\right)\right],$$

$$P(\theta_{q-1}) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left[-\frac{1}{2}\left(\frac{\theta_{q-1} - \mu}{\tau^2}\right)\right],$$

$$P(Y|\theta^*) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{Y - \theta^*}{\sigma^2}\right)\right],$$

$$P(Y|\theta_{q-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{Y - \theta_{q-1}}{\sigma^2}\right)\right],$$

iii. If min (1, T) $> u$, where u is drawn from a uniform distribution on the interval (0, 1) then $\theta_q = \theta^*$ otherwise $\theta_q = \theta_{q-1}$.

The total number of iterations is set equal to $Q = 1000$. The performance of the algorithm depends on the variance of the proposal distribution η^2 (see Exercises). Here, η is set equal to 1.5 t ha^{-1}. With this value, the acceptance rate of the test is equal to 39%.

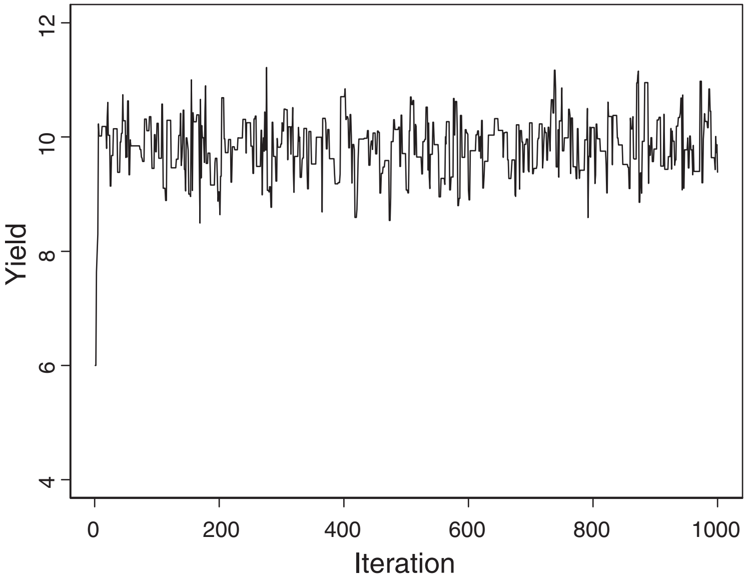


Figure 9. Yield values generated by the Metropolis–Hastings algorithm.

The generated yield values are displayed in Figure 9. The first value is equal to 6 t ha^{-1}. After few iterations, the generated yield values vary in the range 8.5–11.5 t ha^{-1}. The average value and standard deviation computed from the 950 last iterations are equal to 9.842 and 0.22 t ha^{-1} respectively. These values are close to the true posterior mean and standard deviation (9.892 and 0.243, respectively).

4.3.7. Application 2: Estimation of the four parameters of AZODYN

The Metropolis–Hastings algorithm is used here to estimate the four parameters *EBMAX*, *D*, *K* and *VMAX* of AZODYN. The prior distributions of these parameters are defined as uniform distributions, with lower and upper bounds as shown in Table 2. Three datasets are used successively for estimating the parameters. The first one includes all the measurements of dry matter, *LAI*, and nitrogen uptake performed in Grignon during six years. The total number of measurements included in this dataset is equal to 201. The second dataset includes only the measurements obtained in 1999. The number of measurements in this dataset is equal to 28. Finally, the last dataset includes only the three observations obtained in 1999 at the first date of measurements: one measurement of dry matter, *LAI*, and nitrogen uptake. Note that the number of measurements included in the third dataset is smaller than the number of parameters to estimate.

The Metropolis–Hasting algorithm is applied successively to the three datasets. The initial parameter vector θ_0 is set equal to the initial values defined in Table 2. The proposal distribution is defined by $\theta^*|\theta_{q-1} \sim N(\theta_{q-1}, \Sigma)$ where Σ is a diagonal matrix. The diagonal elements are set proportional to the initial parameter values θ_0. The proportionality coefficient is chosen by trial and error in order to obtain an acceptance rate in the range 20–70% for the test performed at Step iii. The total number of iterations Q is set equal to 1000 and the number of discarded iterations M is fixed to 50. Thus, a series of 950 parameter values is generated for each dataset by the algorithm. Sample means and variances of the four parameters are shown in Tables 5 and 6. The values approximate the expected values and variances of the posterior parameter distribution.

Table 5 shows that the expected values of *EBMAX* and *VMAX* differ from the initial values by more than 10%. Values are more similar for the other parameters. Table 6 shows that the posterior variances are lower than the prior variances and depend strongly on the number of data. This is a logical result. Experimental data give information on the parameter values. As a consequence, the uncertainty about parameter values is strongly decreased when numerous data are used for estimating the four parameters. On the other hand, the posterior variances are almost equal to the prior variance when only three data are used for estimating the parameters.

The parameter vectors generated by the Metropolis–Hastings algorithm can be used to derive probability distributions of model output variables. Such distributions are useful for studying the uncertainty about output variables resulting from uncertainty about parameter values. Figure 10 shows two distributions of dry matter predicted by AZODYN at flowering. One of these distributions is derived from the series of parameters generated by using the complete dataset. The second distribution is obtained from the series of parameters generated by using only the data of 1999. The x and y-axis represent dry matter and cumulative probability respectively. Figure 10 shows that the dry matter variance is lower when all the data are used to estimate the parameters. Probability distributions of model outputs are useful because they give information about the uncertainty associated with each prediction (see Chapter 3).

Table 5. Expected values of the posterior distributions obtained with 3, 28, and 201 measurements.

		Posterior mean		
Parameter	Prior mean	3 data	28 data	201 data
EBMAX	2.9	3.2	3.1	3.3
K	0.70	0.71	0.78	0.73
D	0.032	0.032	0.029	0.028
VMAX	0.45	0.39	0.37	0.39

Table 6. Variances of the posterior distributions obtained with 3, 28, and 201 measurements.

		Posterior variance		
Parameter	Prior variance	3 data	28 data	201 data
EBMAX	0.403	0.38	0.005	0.0012
K	0.003	0.003	0.0001	0.0001
D	5.2×10^{-5}	1.1×10^{-5}	5.4×10^{-7}	1.1×10^{-7}
VMAX	0.021	0.012	0.0002	0.00009

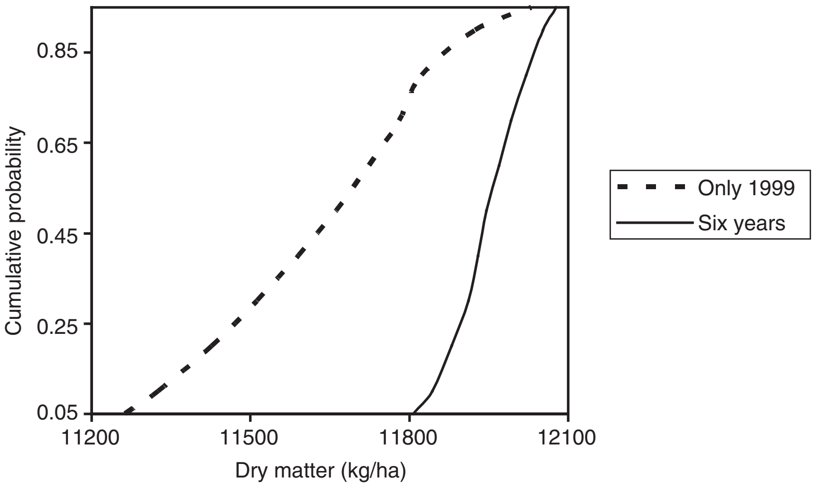


Figure 10. Probability distributions of dry matter at flowering.

5. Discussion

Estimation of crop model parameters requires carefully selected methods because of crop model complexity and because of the structure of the agronomic datasets. Here, we define some practical rules for estimating crop model parameters.

The first step is to determine the different types of information available for estimating parameters. In general, two types of information are available: information provided by literature and expert knowledge, and experimental data. It is important to analyze the structure of the dataset. How many different types of measurements? Are the measurement variances homogeneous? Are the data correlated?

The second step is to select the parameters to estimate from data. It is generally not sensible to estimate all the parameters from data when the number of parameters is larger than 10–15 or when the ratio of number of parameters to number of data is higher than 1/20. The estimation of a large number of parameters leads usually to numerical problems and to inaccurate parameter estimates. Several methods may be used to select parameters to estimate: selection based on the literature, equation analysis, sensitivity analysis, and statistical choice of parameters to estimate. In theory, parameter selection is less important when parameters are estimated by using a Bayesian method because, in this case, parameters are estimated from both data and expert knowledge. Consequently, Bayesian methods can be used to estimate a high number of parameters even if the data are scarce. However, even with this kind of method, it is often useful to select parameters in order to avoid very long calculation time.

After information analysis and parameter selection, the final step is to choose a method for estimating parameter values. When parameter estimation is only based on experimental data, least squares and maximum likelihood methods are appropriate. The method of ordinary least squares gives good results if error variances are homogeneous and if residues are independent. If not, generalized least squares or maximum likelihood should be applied. An alternative is to use a Bayesian method. With this kind of method, parameters are estimated from both prior information and data. In a recent study, Tremblay and Wallach (2004) showed that Bayesian methods can perform better than least squares methods when the ratio of the number of data to the number of parameters is low.

Some important problems related to parameter estimation remain open. One of these problems concerns correlation of model errors. In particular, correlations between errors for observations at different dates in the same site-year are often nonnegligible. As a result, the variance–covariance matrix of errors includes numerous nonzero elements. Estimating this matrix is a difficult problem. Other problems are related to the implementation of Bayesian methods. First, the calculation time required to obtain accurate results is very long when the number of parameters is high. More experience is required before implementing Bayesian methods with models including hundreds of parameters. Second, Bayesian methods use prior parameter distributions and no clear methodology has been established yet to define such distributions.

References

Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M-H., Ruget, F., Nicoullaud, B., Gate, P., Devienne-Barret, F., Antonioletto, R., Durr, C., Richard, G., Beaudoin, N., Recous, S., Tayot, X., Plenet, D., Cellier, P., Machet, J-M., Meynard, J-M., Delécolle, R., 1998. STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn. Agronomie 18, 311–346.

Bonesmo, H., Bélanger, G., 2002. Timothy yield and nutritive value by the CATIMO model I. Growth and Nitrogen. Agronomy Journal 94, 337–345.

Campbell, E.P., Fox, D.R., Bates B.C., 1999. A Bayesian approach to parameter estimation and pooling in nonlinear flood event models. Water Resource Research 35, 211–220.

Carlin, B.P., Louis, T.A., 2000. Bayes and empirical Bayes methods for data analysis. Chapman & Hall, London.

Davidian, M., Giltinan, D.M., 1995. Nonlinear models for repeated measurement data. Chapman & Hall, London.

Efron, B., Gong, G., 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. American Statistician 37, 36–48.

Efron, B., Tibshirani, R.J., 1998. An introduction to the Bootstrap. Chapman & Hall, London.

Fournier, C., Andrieu, B., 2000. Dynamics of the elongation of internodes in Maize. Effects of shade treatment on elongation patterns. Annals of Botany 86, 1127–1134.

Gallant, R., 1987. Nonlinear statistical models. John Wiley & Sons, New York.

Geyer, C.J., 1992. Practical Markov chain Monte Carlo. Statistical Science 7, 473–511.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1995. Markov Chain Monte Carlo in practice. Chapman & Hall, London.

Goffe, W.L., Ferrier, G.D., Rogers, J., 1994. Global optimisation of statistical functions with simulated annealing. Journal of Econometrics 60, 65–99.

Grimm, S.S., Jones, J.W., Boote, K.J., Hesketh, J.D., 1993. Parameter estimation for predicting flowering date of soybean cultivars. Crop Science 33, 137–144.

Harmon, R., Challenor, P., 1997. A Markov chain Monte Carlo method for estimation and assimilation into models. Ecological Modelling 101, 41–59.

Harrel, F.E., Jr., 2001. Regression modeling strategies. Springer, New York.

Huet, S., Bouvier, A., Gruet, M-A., Jolivet, E., 1996. Statistical tools for nonlinear regression. Springer, New York.

Huet, S., Jolivet, E., Messean, A., 1992. La régression non-linéaire. INRA editions, Paris.

Jeuffroy, M-H., Recous, S., 1999. Azodyn: a simple model for simulating the date of nitrogen deficiency for decision support in wheat fertilization. European Journal of Agronomy 10, 129–144.

Makowski, D., Wallach, D., 2002. It pays to base parameter estimation on a realistic description of model errors. Agronomie 22, 179–789.

Makowski, D., Wallach, D., Meynard, J-M., 2001. Statistical methods for predicting responses to applied nitrogen and for calculating optimal nitrogen rates. Agronomy Journal 93, 531–539.

Makowski, D., Wallach, D., Tremblay, M., 2002. Using a Bayesian approach to parameter estimation; comparison of the GLUE and MCMC methods. Agronomie 22, 191–203.

Malakoff, D., 1999. Bayes offers a 'New' way to make sense of numbers. Science 286, 1460–1464.

Mavromatis, T., Boote, K.J., Jones, J.W., Irmak, A., Shinde, D., Hoogenboom, G., 2001. Developing genetic coefficients for crop simulation models with data from crop performance trials. Crop Science 41, 40–51.

Metselaar, K., 1999. Auditing predictive models: a case study in crop growth. Thesis, Wageningen Agricultural University.

Miller, A.J., 1990. Subset selection in regression. Chapman & Hall, London.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-effects models in S and S-PLUS. Springer, New York.

Ripley, B.D., 2001. The R project in Statistical computing, MSOR Connections. The Newsletter of the LTSN Maths, Stats & OR network 1(1):23–25.

Sakamoto, Y., Ishiguro, M., Kitagawa, G., 1986. Akaike Information Criterion Statistics. Reidel, Dordrecht.

Saporta, G., 1990. Probabilités, analyse des données et statistique. Editions Technip, Paris.

SAS/STAT User's Guide, Version 6 (1990). SAS Institute Inc., Cary, USA.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Seber, G.A.F., Wild, C.J., 2003. Nonlinear Regression. Wiley-Interscience, Hoboken, NJ.

Shulz, K., Beven, K., Huwe, B., 1999. Equifinality and the problem of robust calibration in nitrogen budget simulations. Soil Science Society of America Journal 63, 1934–1941.

SPLUS 6, Guide to Statistics, 2001. Insightful Corporation, Seattle, Washington.

Theobald, C.M., Talbot, M., 2002. The Bayesian choice of crop variety and fertilizer dose. Applied Statistics 51, 23–56.

Tremblay, M., Wallach, D., 2004. Comparison of parameter estimation methods for crop models. Agronomie 24, 351–365.

Tremblay, M., 2004. Estimation des paramètres des modèles de culture. Ph.D. Thesis, Université Paul Sabatier, Toulouse, France.

Vold, A., Breland, T.A, Søreng, J.S., 1999. Multiresponse estimation of parameter values in models of soil carbon and nitrogen dynamics. Journal of agricultural, biological, and environmental statistics 4, 290–309.

Wallach, D., Goffinet, B., Bergez, J-E., Debaeke, Ph., Leenhardt, D., Aubertot, J-N., 2001. Parameter estimation for crop models: a new approach and application to a corn model. Agronomy Journal 93, 757–766.

Exercises

Estimating the parameters of STICS for LAI prediction

Some of these exercises are based on results obtained by Tremblay (2004).

An agronomist and a statistician want to estimate the parameters of a simple crop model. The model is a part of the STICS model (Brisson et al., 1998) and is referred to as Mini-STICS. Mini-STICS predicts sunflower leaf area index (LAI) and soil water content during 20 days from the stage Maximal Acceleration of Leaf growth. The model includes 14 parameters (Table 7). The agronomist has defined an initial value and a range of variation for each parameter from expert knowledge (Table 7).

Some of the model equations are given below. LAI on day t is calculated in function of LAI on day $t-1$ as follows:

$$LAI(t) = LAI(t-1) + DELTAI(t)$$

Table 7. Parameters of Mini-STICS.

Parameter	Meaning	Initial value	Range of variation
ADENS	Parameter of compensation between stem number and plant density.	−0.696	−0.974, −0.417
BDENS	Maximum density above which there is competition between plants.	1.1029 plants m^{-2}	0.662, 1.544
CROIRAC	Growth rate of the root front.	0.2913 cm °C Day^{-1}	0.175, 0.407
DLAIMAX	Maximum rate of the setting up of LAI.	0.0061 m² leaves plant^{-1}°C Day^{-1}	0.00366, 0.0085
EXTIN	Extinction coefficient of photosynthetic active radiation in the canopy.	0.6396	0.384, 0.895
KMAX	Maximum crop coefficient for water requirements.	1.4101	0.846, 1.974
LVOPT	Optimum root density.	0.5672 cm root/cm^{-3} soil	0.34, 0.794
PSISTO	Absolute value of the potential of stomatal closing.	12.29 bar	7.37, 17.21
PSISTURG	Absolute value of the potential of the beginning of decrease of the cellular extension.	3.79 bar	2.27, 5.31
RAYON	Average radius of roots.	0.0167 cm	0.010, 0.023
TCMIN	Minimum temperature for growth.	7.1°C	4.26, 9.94
TCOPT	Optimum temperature for growth.	32.1°C	19.26, 44.94
ZPENTE	Depth where the root density is ½ of the surface root density for the reference profile.	113.1 cm	67.86, 158.3
ZPRLIM	Maximum where the root profile for the reference profile.	154.9 cm	92.94, 216.9

$$DELTAI(t) = \left\{ \frac{DLAIMAX}{1 + \exp[5.5(2.2 - \mathrm{ULAI}(t))]} \right\} \times [TCULT(t-1) - TCMIN]$$
$$\times\, TURFAC(t) \times DENSITE \times \left[\frac{DENSITE}{BDENS} \right]^{ADENS}$$

where *DLAIMAX*, *ADENS*, *BDENS*, *TCMIN* are four parameters, *TCULT(t)* is the average temperature on day *t*, *TURFAC(t)* is the turgescent stress index on day *t*, *ULAI(t)* is the leaf development unit on day *t*.

A dataset including 14 LAI measurements obtained on different sites-years is used for estimating parameters:

Site-Year	LAI
1	3.89
2	3.67
3	4.54
4	4.25
5	4.84
6	3.66
7	3.02
8	3.58
9	2.79
10	4.59
11	3.76
12	3.30
13	4.90
14	3.87

Each measurement was obtained on day 20 (i.e. on the last day simulated by the model). The input variables of the model are known for all site-years.

The agronomist and the statistician proceed in four steps to solve their problem.

1. Because of the small size of the dataset, the two colleagues decide to estimate only a subset of the 14 parameters.
   (a) The agronomist first suggests to estimate only the parameter ADENS and to fix the other 13 parameters to their initial values. Which method would you recommend to estimate ADENS using only the dataset, and without taking into account the initial parameter value?
   (b) The agronomist still wants to estimate only one parameter from the dataset but not necessarily ADENS. He asks the statistician which parameter should be estimated from data. Define a method to select the parameter to estimate from data.
   (c) Each one of the 14 parameters was estimated in turn using the dataset. Table 8 shows the sums of squared differences between measured and predicted LAI values (SSD). Each value of SSD is obtained by estimating only one parameter from data. Choose the parameter to estimate from the results.

Table 8. Values of sum of squared differences between observed and predicted LAI (SSD). Each value is obtained by estimating one parameter from data. '–' indicates that the corresponding parameter cannot be estimated.

Estimated parameter	Estimated value	SSD
ADENS	−0.5460	6.0937
BDENS	1.6002	5.8386
CROIRAC	–	–
DLAIMAX	0.0079	5.8386
EXTIN	−0.3102	9.0716
KMAX	−0.6112	9.0716
LVOPT	6.4951	9.0716
PSISTO	–	–
PSISTURG	55.69	9.0716
RAYON	–	–
TCMIN	3.6045	5.7219
TCOPT	32.01	9.6983
ZPENTE	19 591.54	9.4679
ZPRLIM	−387.57	9.4637

(d) The question now is how many parameters should be estimated from data? Define a method based on *MSEP* and *SSD* to determine the optimal number of parameters. This method must select the subset of parameters giving the most accurate *LAI* predictions.

(e) The statistician presents two other methods to select the subset of parameters to estimate from data. In the first method, the *MSEP* is replaced by another criterion, namely BIC_c (corrected Schwarz criterion) defined by

$$BIC_c = -2 \ln L(\hat{\theta}) + p \log(n)/(n - p - 2)$$

where p is the number of estimated parameters, n is the number of data, $L(\hat{\theta})$ is the likelihood value obtained with the estimated parameter values $\hat{\theta}$. $L(\hat{\theta})$ can be calculated from *SSD*. The lower *SSD*, the higher likelihood. The second method is based on sensitivity analysis. The principle is to increase and decrease by $x\%$ the initial parameter values reported in Table 7. Each parameter is considered in turn. The difference between the LAI values predicted with the high and low parameter value is then calculated for each parameter. The LAI difference is noted ΔLAI. The statistician suggests to take $x = 30\%$ and to estimate from data only the parameters for which $\Delta\text{LAI}/\text{LAI}_0 > 10\%$ where LAI_0 is the LAI predicted with initial parameter values.

What is the usefulness of the two methods proposed by the statistician compared to the method based on MSEP?

(f) Table 9 shows *MSEP* values estimated by cross validation and values of BIC_c for different subsets of parameters. How many parameters must be estimated from data?

Table 9. *MSEP* values estimated by cross validation and BIC_c values obtained for different subsets of parameters minimizing SSD.

Number of estimated parameters	Estimated parameters	*MSEP*	BIC_c
1	*TCMIN*	0.47	−0.65
2	*TCMIN, ADENS*	0.46	−0.68
3	*TCMIN, ADENS, KMAX*	0.49	−0.34
4	*TCMIN, ADENS, KMAX, EXTIN*	0.54	9.43
5	*TCMIN, ADENS, KMAX, EXTIN, LVOPT*	0.55	0.66

2. According to the agronomist, the parameter values obtained with the methods described in Exercise 1 are not realistic. For example, when TCMIN and ADENS are estimated from the 14 LAI measurements, the estimated values are equal to −180.15 and −2.24 respectively. These two values are not within the ranges defined in Table 7.

   The statistician proposes two other estimation methods to solve this problem. The first method is ordinary least squares with constraints on parameter values (lower and upper bounds). The second method is to minimize the following criterion:

$$\sum_{i=1}^{14}\left[Y_i^{\mathrm{LAI}} - f(X_i;\theta)\right]^2 + \sigma^2\sum_{k=1}^{K}\frac{(\theta_k-\mu_k)^2}{\sigma_k^2}$$

   where Y_i^{LAI} is the ith measurement of LAI, σ^2 is the variance of the model residues, K is the number of parameters estimated from data, μ_k is the initial value of the parameter θ_k, σ_k^2 is the initial level of uncertainty about parameter θ_k. The second method is called further "penalized least squares".

   (a) Which parameter values are penalized by the second estimation method?
   (b) Which additional information must be supplied by the agronomist to apply the second method? What influence does this information have?
   (c) Table 10 shows the parameters selected by sensitivity analysis and the parameter values estimated by ordinary least squares.

Table 10. Parameters selected by sensitivity analysis and their estimated values.

Parameter	Value estimated without constraints	Value estimated with constraints
ADENS	−2.2082	−0.974
BDENS	0.8755	1.4816
DLAIMAX	0.0019	0.0076
TCMIN	−1019.61	4.2693

Table 11. Results obtained by penalized least squares.

Parameter	Estimated value
ADENS	−0.6468
BDENS	1.1707
CROIRAC	0.2853
DLAIMAX	0.0067
EXTIN	0.6385
KMAX	1.4383
LVOPT	0.5672
PSISTO	12.31
PSISTURG	3.7789
RAYON	0.0167
TCMIN	6.6961
TCOPT	32.01
ZPENTE	113.09 881
ZPRLIM	154.9017

Two series of estimated values are displayed in this table, one obtained with constraints on parameter values and one without. Table 11 shows the results obtained by penalized least squares. With this method, all the parameters are estimated from data. What effects have the constraints on parameter values ? Compare the results obtained by ordinary least square and those obtained by penalized least squares.

Posterior parameter distribution

The agronomist is satisfied with the last results but would like to have information about parameter uncertainty. The statistician suggests calculating a posterior parameter distribution using a Bayesian method. For illustration, the statistician presents a simple example, as reported below.

Suppose we want to estimate yield value for a given field. The unknown true yield value is denoted as θ. Two types of information are available for estimating θ. First, according to an expert, the yield value must be near from $\mu = 5$ t ha^{-1} and the uncertainty about yield value is $\tau = 2$ t ha^{-1}. μ and τ are used to define a *prior* yield distribution as follows:

$$\theta \sim N(\mu, \tau^2).$$

Second, an imperfect yield measurement was performed in the field. We suppose that this measurement is distributed as:

$$Y \sim N(\theta, \sigma^2)$$

where Y is the measurement and σ^2 is the variance of the measurement. σ^2 is calculated from replicates.

The prior distribution and the measurement are both used to derive a *posterior* yield distribution. The analytical expression of the posterior yield distribution is:

$$\theta|Y \sim N[B\mu + (1 - B)Y, (1 - B)\sigma^2]$$

where $B = \sigma^2/(\sigma^2 + \tau^2)$. According to this formula, the expected value of the posterior distribution is

$$E(\theta|Y) = B\mu + (1 - B)Y = \frac{\sigma^2\mu + \tau^2 Y}{\sigma^2 + \tau^2}$$

and the variance of the posterior distribution is

$$\mathrm{var}(\theta|Y) = (1 - B)\sigma^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

The expected value $E(\theta|Y)$ can be seen as an estimator of θ taking into account both the expert knowledge and the measurement. The posterior variance $\mathrm{var}(\theta|Y)$ gives information on the final level of uncertainty about θ.

(d) Now, suppose that the measured yield value is $Y = 9$ t ha^{-1} and $\sigma = 1$ t ha^{-1}. Calculate the expected value and variance of the posterior distribution of θ when $\mu = 5$ t ha^{-1} and $\tau = 2$ t ha^{-1}. What is the effect of using a less accurate measurement (e.g. $\sigma = 2$ t ha^{-1})?

(e) The agronomist is fascinated by the Bayesian method and asks the statistician to apply it to Mini-STICS. The statistician says that the application of this method to Mini-STICS is not straightforward. The model is nonlinear and includes a large number of parameters. As a consequence, it is not possible to determine the analytic expression of the posterior distribution. To solve this problem, the statistician suggests to apply the Metropolis–Hastings algorithm. The principle is to generate a sample of parameters from the posterior parameter distribution. This sample can be used to derive various quantities of interest such as expected values and variances. The application of the Metropolis–Hastings algorithm requires knowledge of the prior parameter distribution and of the relationship between data and parameters.

For illustration, the statistician considers again the problem of yield estimation. He uses the Metropolis–Hastings algorithm to sample 1000 values of θ in the posterior distribution

$$P(\theta|Y) = N\left[B\mu + (1 - B)Y, (1 - B)\sigma^2\right]$$

where $B = \sigma^2/(\sigma^2 + \tau^2)$, $\mu = 5$, $\tau = 2$, $Y = 9$, and $\sigma = 1$. The idea of the statistician is to show that the generated sample can be used to estimate the expected value and variance of the posterior distribution. The performance of the algorithm depends on some tuning coefficients, such as the initial parameter values and the variance of the probability distribution used to generate parameter values. The results obtained with two series of tuning coefficients are displayed

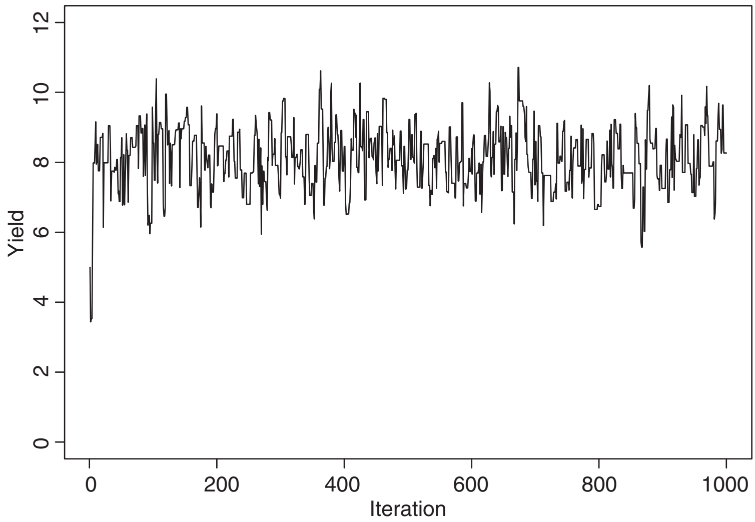


Figure 11. Sample of 1000 values of θ generated by the Metropolis–Hastings algorithm. First series of tuning coefficients.

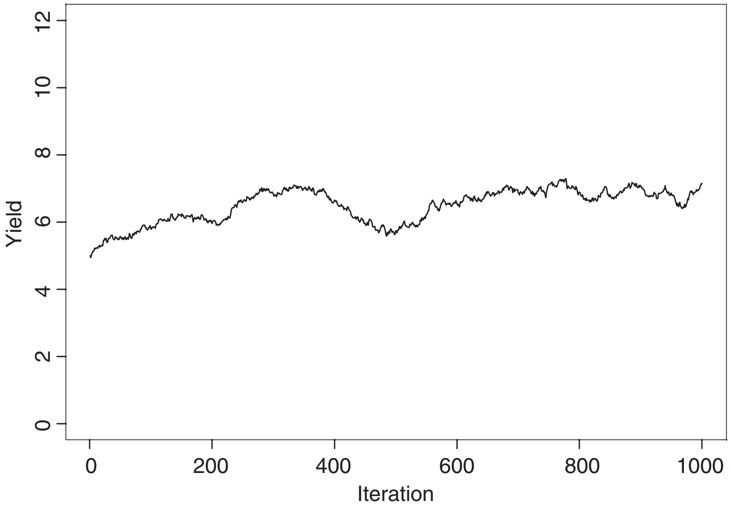


Figure 12. Sample of 1000 values of θ generated by the Metropolis-Hastings algorithm. Second series of tuning coefficients.

in Figures 11 and 12. Each figure shows 1000 generated values of θ. In both cases, the first value of θ was fixed to 5 but two different probability distributions were used to generate parameter values. Based on the true posterior distribution calculated in Q2 (d), which sample of parameter values gives the best result? How do you use the 1000 generated values of θ to calculate the expected value and variance of the posterior distribution?

Table 12. Estimation methods.

Method	Name	Bounds	Method to select parameters
1	Ordinary least squares	No	BIC_c
2	Ordinary least squares	Yes	BIC_c
3	Ordinary least squares	No	Sensitivity analysis
4	Ordinary least squares	Yes	Sensitivity analysis
5	Penalized least squares	No	Sensitivity analysis
6	Metropolis–Hastings	–	–

(f) Bayesian methods require a prior parameter distribution. Suggest such a distribution for Mini-STICS.

3. The two colleagues decide to compare all the methods described in Exercises 1 and 2. The statistician suggests making a simulation study. To do that, he defines the following "true" model:

$$LAI = f(X;\theta) + \varepsilon$$

where *LAI* is the LAI observation at day 20, *f* is the crop model, *X* is the vector including input variables, θ is the vector of true parameter values and ε is the model residue such as $\varepsilon \sim N(0, 0.36)$. True parameter values are different from the initial values shown in Table 7. For example, the true value of TCMIN is equal to 6. The "true" model is used to generate 80 samples of data. Each sample includes 14 LAI measurements. Model parameters are estimated from each sample by using six estimation methods successively (Table 12). Only a subset of parameters is estimated with methods 1–5. On the contrary, all the parameters are estimated with the Metropolis–Hastings algorithm. For this method, the prior parameter distribution is a uniform distribution with lower and upper bounds given in Table 7.

(a) Define a criterion to compare the methods. How do you calculate this criterion?
(b) What is the smallest *MSEP* possible value for the model?
(c) Table 13 shows the *MSEP* values obtained with the different methods. Table 14 shows the average values of 80 TCMIN values estimated from 80 generated

Table 13. MSEP values obtained with different estimation methods.

Estimation method	Number of estimated parameters	*MSEP*
1	1–2	0.41
2	1–2	0.39
3	4	0.45
4	4	0.42
5	4	0.39
6	14	0.39
Initial parameter values	0	0.63

Table 14. Average values of 80 *TCMIN* values estimated from 80 generated samples of data.

Estimation method	Average value of *TCMIN*
1	−8.07
2	6.32
3	−65.14
4	5.18
5	6.79
6	7.2
Initial parameter value	7.1
True value	6

samples of data. Use the results displayed in Tables 13 and 14 to choose an estimation method.

4. The agronomist wonders if it would be useful to estimate parameters from a dataset including several types of measurements. Two new datasets are proposed:

   - A dataset including 14 measurements of LAI and 14 measurements of soil water content obtained on 14 different site-years at one date (day 20).
   - A dataset including 28 measurements of LAI and 28 measurements of soil water content obtained on 14 site-years at two dates (day 10 and day 20).

   (a) Which method(s) may be used to estimate parameters from these datasets?

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 5

Data assimilation with crop models

D. Makowski, M. Guérif, J.W. Jones and W. Graham

1. Introduction **151**
2. The Kalman filter **152**
3. Data assimilation for non-linear models **161**
4. Conclusion **169**
References **170**
Exercises **171**

1. Introduction

Up to now, we have assumed that the objective is to develop a model that is a good predictor on the average over some target distribution of situations. In particular, we have assumed that parameter estimation is based on a random sample of data from the target population. This is in keeping with the idea that we want the parameters to represent on the average the target population. Furthermore, we have emphasized the mean squared error of prediction for model evaluation. This criterion explicitly measures how well the model predicts on the average over the entire target population.

In this chapter, however, we consider modifying the population model to make it a better predictor for a specific situation (e.g. for a specific agricultural field). We assume that we have some information that allows us to modify the model specifically for the situation of interest. In general, this information takes the form of one or several measurements of model state variables during the crop growth period. The model can then be modified based on the measurement, and the modified model used to make predictions for the future growth of the crop. If for example, the measurement shows that the value of leaf area index (*LAI*) is smaller than the model predicts, then we can change *LAI* in the model to a smaller value, and use the corrected model for predicting the future evolution of the system.

Model modification based on measurements is called data assimilation because the data are incorporated into or assimilated into the model. Various methods can be used for data assimilation. Methods of parameter estimation can be used to adjust the values of the model parameters to the data obtained for the situation of interest. This approach was described in detail in Chapter 4 and the use of a parameter estimation method for data assimilation is illustrated in a case study presented in Chapter 17 of this book (Guérif et al.). In this chapter, we consider another family of methods often referred to as filtering. A *filter* is an algorithm that is applied to a time series to improve it in some way (like filtering out noise). Here, the time series is the successive values for the model state variables, and the improvement comes from using measured values to update the model state variables. The state variables are updated sequentially, i.e. each time an observation is available. The best-known algorithm for doing this is the *Kalman filter*, which applies if the model is linear and the errors have a normal distribution (Kalman, 1961). This chapter is mostly devoted to that algorithm and its extensions (Welch and Bishop, 1992; Burgers et al., 1998; Anderson and Anderson, 1999; Pastres et al., 2003).

Measurements of output variables are in fact increasingly commonly available, with increases in detection and transmission capability. Satellite systems give information about plant biomass, *LAI*, or leaf chlorophyll content. Tensiometers can be used to give information about soil moisture. Several methods are available for giving information about plant nitrogen status. In each case, the results of the measurements can be compared to model predictions, and the model can be adjusted in the light of those measurements.

Potentially, such measurements could lead to a large improvement in model predictions. In the chapter on model evaluation we showed that MSEP, the mean squared error of prediction, cannot be smaller than a lower bound, which is a measure of how much variability that remains is not accounted for by the explanatory variables in the model. This lower limit to prediction error no longer applies, however, if measurements are used to correct the model, for now we are injecting extra information in addition to the explanatory variables. How good will predictions be after correction with measurements? This depends on a number of factors, in particular how closely the measured variables and the predicted variables are related. A measurement of *LAI* early in the season may not improve yield prediction much, while a late measurement of biomass may lead to substantial improvement in yield prediction. We emphasize again that this discussion only concerns the situation where the measurements were made. The correction will not be used when applying the model in other situations.

Section 2 treats the case of models whose dynamic equations are linear in the state variables and parameters to be modified. It is further assumed that all errors have normal distributions. We first treat simple special cases, which help introduce the methods and the consequences of data assimilation. Then we treat a fairly general case that covers all the special cases. Of course essentially all crop models are non-linear. Nonetheless it is important to treat the linear case. First of all, it is easier to get a basic understanding of how assimilation works from the linear case, where we can derive analytical equations. Secondly, the methods for non-linear models draw more or less on the theory developed for linear models. In Section 3, we consider the problem of data assimilation for non-linear models. We discuss two different approaches to data assimilation for such models.

2. The Kalman filter

2.1. Filter to update one state variable

2.1.1. Method

We consider a linear dynamic model including a single state variable and defined by

$$Z_t = GZ_{t-1} + B_{t-1} + \varepsilon_{t-1} \tag{1}$$

where Z_t is the model state variable at time t, G is a parameter, B_{t-1} is an input variable, and ε_{t-1} is the model error. This is the equation that describes the evolution of the state variable over periods where there are no measurements. At time t_1, Z_{t_1} can be expressed as a function of Z_1, the initial value of the state variable, as

$$Z_{t_1} = G^{t_1-1} Z_1 + \sum_{k=1}^{t_1-1} G^{t_1-1-k} B_k + \sum_{k=1}^{t_1-1} G^{t_1-1-k} \varepsilon_k .$$

We assume that the error terms ε_t obtained at different dates are independent and have normal distributions with zero expectation. We further assume that Z_1 has a normal distribution and is independent of ε_t. Under these assumptions and if no measurement is available before time t_1, the expectation and variance of Z_{t_1} are defined by

$$E(Z_{t_1}) = G^{t_1-1} E(Z_1) + \sum_{k=1}^{t_1-1} G^{t_1-1-k} B_k$$

and

$$\mathrm{var}(Z_{t_1}) = G^{2(t_1-1)} \mathrm{var}(Z_1) + \sum_{k=1}^{t_1-1} G^{2(t_1-1-k)} \mathrm{var}(\varepsilon_k).$$

We now consider that a measurement M_{t_1} is available at time t_1. We assume that one measures directly the state variable. The measurement equation is then

$$M_{t_1} = Z_{t_1} + \tau_{t_1} \tag{2}$$

where τ_{t_1} is the measurement error, normally distributed with zero mean and independent of Z_{t_1}. Then, M_{t_1} has a normal distribution and $E(M_{t_1}) = E(Z_{t_1})$ and $\mathrm{var}(M_{t_1}) = \mathrm{var}(Z_{t_1}) + \mathrm{var}(\tau_{t_1})$.

The joint distribution of M_{t_1} and Z_{t_1} is

$$\begin{pmatrix} Z_{t_1} \\ M_{t_1} \end{pmatrix} \sim N \left\{ \begin{bmatrix} E(Z_{t_1}) \\ E(M_{t_1}) \end{bmatrix}, \begin{bmatrix} \mathrm{var}(Z_{t_1}) & \mathrm{cov}(Z_{t_1}, M_{t_1}) \\ \mathrm{cov}(Z_{t_1}, M_{t_1}) & \mathrm{var}(M_{t_1}) \end{bmatrix} \right\}$$

with $\mathrm{cov}(Z_{t_1}, M_{t_1}) = \mathrm{cov}(Z_{t_1}, Z_{t_1} + \tau_{t_1}) = \mathrm{var}(Z_{t_1})$. The distribution of Z_{t_1} conditionally to M_{t_1} is derived from the joint distribution as follows (see the appendix, Statistical notions):

$$Z_{t_1} | M_{t_1} \sim N \left[E(Z_{t_1} | M_{t_1}), \mathrm{var}(Z_{t_1} | M_{t_1}) \right]$$

with

$$E(Z_{t_1} | M_{t_1}) = E(Z_{t_1}) + K_{t_1} \left[M_{t_1} - E(Z_{t_1}) \right] \tag{3}$$

$$\mathrm{var}(Z_{t_1} | M_{t_1}) = (1 - K_{t_1}) \mathrm{var}(Z_{t_1}) \tag{4}$$

and

$$K_{t_1} = \frac{\mathrm{cov}(Z_{t_1}, M_{t_1})}{\mathrm{var}(M_{t_1})} = \frac{\mathrm{var}(Z_{t_1})}{\mathrm{var}(Z_{t_1}) + \mathrm{var}(\tau_{t_1})} \tag{5}$$

K_{t_1} is often referred to as the Kalman gain.

Consider first the conditional expectation Eq. (3). It is a weighted sum of $E(Z_{t_1})$, the model prediction in the absence of any measurement, and M_{t_1}, the measured value. Each term is weighted by the variance of the other and the sum of the weights is one. If the measurement has a large variance compared to the model variance, then one gives most weight to the model prediction. If on the other hand, the measurement has a small variance compared to the model variance, then most weight is given to the measurement.

In some studies, the authors simply replace the model prediction by the measured value. It is easily seen from Eq. (3) that this corresponds to assuming that there is no measurement error ($\mathrm{var}(\tau_{t_1}) = 0$). In fact this assumption is rarely if ever true, and often measurement error is substantial. Equation (3) on the other hand shows how to take into account both sources of information, namely the model and the measurement, weighting each according to its level of uncertainty.

Consider now the variance of the posterior distribution defined by Eq. (4). Since the Kalman gain is in the range 0–1, the variance after measurement, $\mathrm{var}(Z_{t_1}|M_{t_1})$, is smaller than the variance before measurement $\mathrm{var}(Z_{t_1})$. The larger the gain K_{t_1}, the smaller the $\mathrm{var}(Z_{t_1}|M_{t_1})$. This is a gain in the sense that it measures how much knowledge we have gained by using the measurement.

Note that the two essential properties of this case are linearity and the normal distribution of the random variables. As a result of linearity the state variable at any time just involves linear combinations of the errors. If the errors have a normal distribution, then the state variable is also normally distributed. In the rest of this chapter, unless stated otherwise, we assume that all the random variables are independent and are normally distributed.

We now introduce the general equations for the model Eq. (1). We note $M_{1:t}$ the vector including the measurements obtained up to time t. Then, the distribution of the state variable Z_t conditionally to $M_{1:t}$ is defined by

$$Z_t|M_{1:t} \sim N\left[E(Z_t|M_{1:t}), \mathrm{var}(Z_t|M_{1:t})\right]$$

with

$$E(Z_t|M_{1:t}) = E(Z_t|M_{1:t-1}) + K_t\left[M_t - E(Z_t|M_{1:t-1})\right]$$

$$\mathrm{var}(Z_t|M_{1:t}) = (1 - K_t)\mathrm{var}(Z_t|M_{1:t-1})$$

$$K_t = \frac{\mathrm{var}(Z_t|M_{1:t-1})}{\mathrm{var}(Z_t|M_{1:t-1}) + \mathrm{var}(\tau_t)}$$

2.1.2. Application

We describe here a simple dynamic crop model with a single state variable representing above-ground winter wheat biomass per unit ground area ($g \cdot m^{-2}$). This state variable is simulated on a daily basis as a function of daily temperature and daily incoming radiation according to the classical efficiency approach (Monteith, 1977; Varlet-Grancher et al., 1982). The biomass at time $t + 1$ is linearly related to the biomass at time t as follows:

$$Z_{t+1} = Z_t + E_b E_{i\max}[1 - e^{-KLAI_t}]PAR_t + \varepsilon_t \tag{6}$$

where t is the number of days since sowing, Z_t is the true above-ground plant biomass on day t ($g \cdot m^{-2}$), PAR_t is the incoming photosynthetically active radiation on day t ($MJ \cdot m^{-2} \cdot j^{-1}$), LAI_t is the green leaf area index on day t (m^{-2}leaf $\cdot$ m^{-2}soil), ε_t is a random term representing the model error. ε_t is assumed normally distributed with zero mean and constant variance Q. The error terms are assumed independent.

The crop biomass at sowing is set equal to zero, $Z_1 = 0$. LAI_t is calculated as a function of the cumulative degree-days (above 0°C) from sowing until day t, noted T_t, as follows (Baret, 1986):

$$LAI_t = L_{\max}\left\{\frac{1}{1 + e^{-A[T_t - T_1]}} - e^{B[T_t - T_2]}\right\}.$$

Parameter T_2 is set equal to $\frac{1}{B}\log[1 + \exp(A \times T_1)]$ in order to have $LAI_1 = 0$ (Déjean et al., 2002). The model includes two input variables $X_t = (T_t, PAR_t)^T$ and seven parameters $\theta = (E_b, E_{i\max}, K, L_{\max}, A, B, T_1)^T$. E_b is the radiation use efficiency which expresses the biomass produced per unit of intercepted radiation ($g \cdot MJ^{-1}$), $E_{i\max}$ is the maximal value of the ratio of intercepted to incident radiation, K is the coefficient of extinction of radiation, $L_{\max}$ is the maximal value of LAI, T_1 defines a temperature threshold, and A and B are two additional parameters describing the rates of growth and senescence of the green LAI. The parameter values were estimated for durum wheat crops in previous studies (Guérif, personal communication). Nominal values and ranges of variation are displayed in Table 1.

Table 1. Values of parameters.

Parameter	Unit	Nominal value	Range of values
E_b	$g \cdot MJ^{-1}$	1.85	0.9–2.8
$E_{i\max}$	–	0.945	0.9–0.99
K	–	0.7	0.6–0.8
$L_{\max}$	–	7.5	3–12
T_1	°C	900	700–1100
A		0.0065	0.003–0.01
B		0.00205	0.0011–0.003

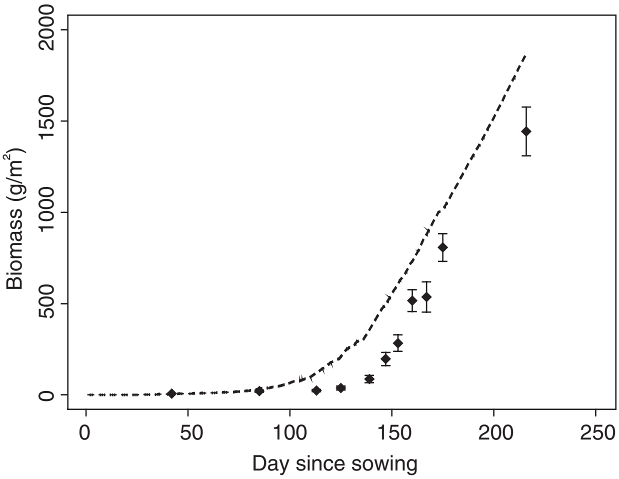


Figure 1. Wheat biomass predicted for the field "Carmague-1987." Black points represent biomass measurements. The 11th measurement is the value measured at harvest ($t = 216$ days). Error bars indicate ±2 standard errors.

The model Eq. (6) is a particular case of model Eq. (1) with $G = 1$ and $B_t = E_b E_{i\,\max}[1 - e^{-K\,LAI_t}]PAR_t$. Here, we use this model to predict the crop biomass for a field of durum wheat (*Triticum durum*, cultivar Creso) located in southern France (Camargue). The field was sown on October 26, 1986 and harvested on June 29, 1987, i.e. 216 days after sowing.

We first run the model to predict biomass without using any measurement. The parameters were fixed to their nominal values (Table 1). Between sowing and the time t_D = date of harvest – 40 days, the biomass was predicted each day from Eq. (6) with $\varepsilon_t = 0$. After day t_D, 14 biomass predictions were derived with 14 years of climate variables and then averaged. Thus, the values $X_t = (T_t,\ PAR_t)^T$ obtained for 1986–1987 were used to predict the biomass up to time t_D and the climate was supposed to be unknown after this date. The predicted biomass values are reported in Figure 1.

The Kalman filter was then applied to update the model state variable using ten biomass measurements $M_1, \ldots, M_{10}$ obtained at different dates before harvest from day 41 to 176 since sowing (Fig. 1). The parameters were not updated and were fixed to their nominal values. Each measurement was obtained by averaging 10 replicates. We further assume that the measurements were independent and were related to the true biomass Z_t according to Eq. (2). The values of $\mathrm{var}(\tau_t)$ were set equal to the empirical variances calculated from replicates (Fig. 1).

The Kalman filter was implemented to the model, as described in Section 2.1.1. Thus, at the date of the first measurement, $E(Z_{t_1})$, $\mathrm{var}(Z_{t_1})$, and K_{t_1} were calculated as follows:

$$E(Z_{t_1}) = \sum_{t=1}^{t_1-1} E_b E_{\mathrm{imax}}[1 - e^{-KLAI_t}]PAR_t,\ \mathrm{var}(Z_{t_1}) = (t_1 - 1)Q,$$

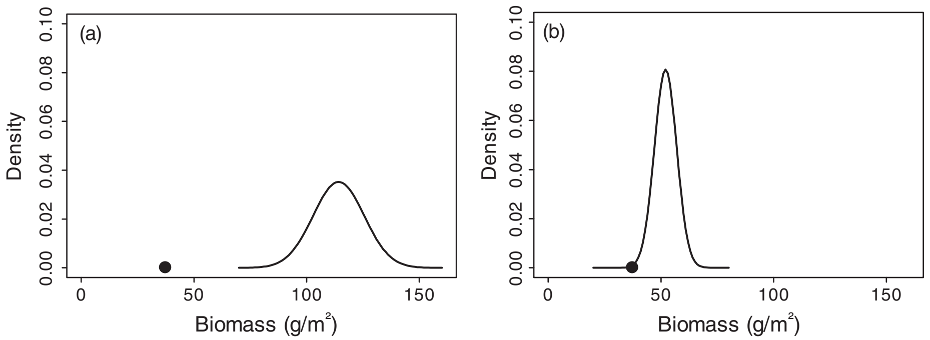


Figure 2. Probability distribution of $Z_{t_4}|M_{1:t_3}$ (a) and $Z_{t_4}|M_{1:t_4}$ (b) obtained with $Q = 10\ \mathrm{g}^2 \cdot \mathrm{m}^{-4}$. The black point indicates the value of biomass measured at time t_4.

and

$$K_{t_1} = \frac{(t_1 - 1)Q}{(t_1 - 1)Q + \operatorname{var}(\tau_{t_1})}.$$

Two values of Q, 0.1 and 10 $\mathrm{g}^2 \cdot \mathrm{m}^{-4}$, were tested successively in order to study the influence of the model error variance on the model output.

Figure 2 shows the two probability distributions obtained at time $t_4 = 125$ (date of the fourth measurement) obtained with $Q = 10\ \mathrm{g}^2 \cdot \mathrm{m}^{-4}$. The first distribution is defined by $Z_{t_4}|M_{1:t_3} \sim N[E(Z_{t_4}|M_{1:t_3}), \operatorname{var}(Z_{t_4}|M_{1:t_3})]$. It represents the biomass probability distribution at $t = t_4$ when the fourth measurement is not taken into account. The second distribution is defined by $Z_{t_4}|M_{1:t_4} \sim N[E(Z_{t_4}|M_{1:t_4}), \operatorname{var}(Z_{t_4}|M_{1:t_4})]$. Both distributions are normal but the two distributions are characterized by very different expected values and variances. $E(Z_{t_4}|M_{1:t_3})$ is equal to 114.19 $\mathrm{g} \cdot \mathrm{m}^{-2}$ and is much higher than the value of $E(Z_{t_4}|M_{1:t_4})$ (52.03 $\mathrm{g} \cdot \mathrm{m}^{-2}$). It is important to note that $E(Z_{t_4}|M_{1:t_4})$ is not strictly equal to the biomass measured at time t_4. $E(Z_{t_4}|M_{1:t_4})$ is a weighted sum of $E(Z_{t_4}|M_{1:t_3})$ and of the measurement. The weight depends both on the variance of the model error and on the variance of the measurement error as shown in Eq. (5).

Another interesting result is that $\operatorname{var}(Z_{t_4}|M_{1:t_3})$ is equal to 128.33 $\mathrm{g}^2 \cdot \mathrm{m}^{-4}$ and is much higher than $\operatorname{var}(Z_{t_4}|M_{1:t_4})$ (24.39 $\mathrm{g}^2 \cdot \mathrm{m}^{-4}$). This result shows that the use of the measurement at time t_4 has reduced the uncertainty in the crop model prediction.

Figure 3 presents the initial and updated crop model predictions obtained between sowing and harvest for two different values of Q. The updated predictions reported in Figure 3 (continuous lines) correspond to the expected values of the biomass distributions. For example, the updated prediction at time t_4 computed with $Q = 10\ \mathrm{g}^2 \cdot \mathrm{m}^{-4}$ is equal to $E(Z_{t_4}|M_{1:t_4})$. Figure 3 shows that the errors of predictions are large when the crop model is not updated with measurements (dashed curve). The value predicted by the model at harvest is equal to 1867.4 $\mathrm{g} \cdot \mathrm{m}^{-2}$ and, so, is much higher than the measured value (1443.4 $\mathrm{g} \cdot \mathrm{m}^{-2}$). The biomass at harvest is more accurately predicted when the crop model is adjusted to the first 10 measurements by using the Kalman filter (continuous curves) with $Q = 10\ \mathrm{g}^2 \cdot \mathrm{m}^{-4}$. It is not the case when the coefficient Q (describing the

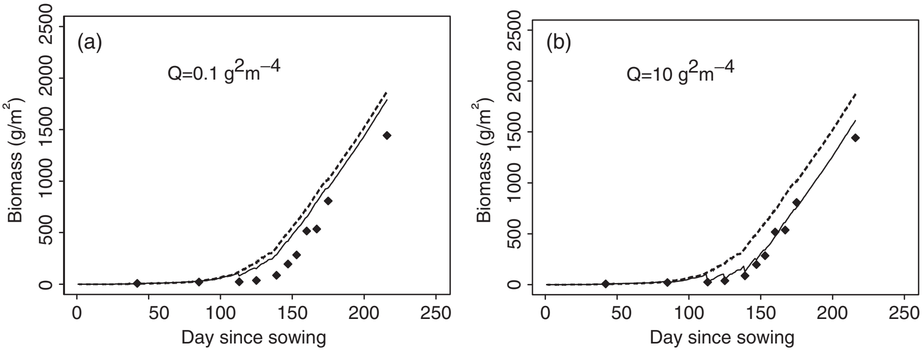


Figure 3. Initial model predictions (dashed line) and updated model predictions (continuous line) obtained with $Q = 0.1\ \mathrm{g}^2 \cdot \mathrm{m}^{-4}$ (a) and $Q = 10\ \mathrm{g}^2 \cdot \mathrm{m}^{-4}$ (b). The black points represent biomass measurements. Model predictions were updated with the Kalman filter method using the first 10 measurements. The 11th measurement is the value measured at harvest ($t = 216$ days).

size of the model error) is fixed to a lower value, $0.1\ \mathrm{g}^2 \cdot \mathrm{m}^{-4}$; the model is not strongly adjusted to the first 10 measurements and the error of prediction is large at harvest. This is logical because high values of Q tend to increase the Kalman gain K_t. The correction is important only if Q is fixed to a high value. In practice, the parameters describing the model errors must be chosen carefully, for instance by using a training data set.

2.2. A more general linear model

2.2.1. Method

In this section, we consider a more general linear model including several state variables. We also consider the possibility of updating both state variables and parameters from measurements. Denote $\varphi_t = [Z_t^{(1)}, \ldots, Z_t^{(n)}, \theta^{(1)}, \ldots, \theta^{(p)}]^{\mathrm{T}}$ the vector of the n state variables and p parameters of the model. The model is defined by

$$\varphi_t = S_{t-1}\varphi_{t-1} + B_{t-1} + \varepsilon_{t-1} \tag{7}$$

where S_{t-1} is a $(n + p) \times (n + p)$ matrix, B_{t-1} is a $(n + p)$-vector of input variables defined by $B_{t-1} = [B_{t-1}^{(1)}, \ldots, B_{t-1}^{(n)}, 0, \ldots, 0]^{\mathrm{T}}$, and ε_{t-1} is a $(n + p)$-vector of error terms defined by $\varepsilon_{t-1} = [\varepsilon_{t-1}^{(1)}, \ldots, \varepsilon_{t-1}^{(n)}, 0, \ldots, 0]^{\mathrm{T}}$. The vector ε_t is assumed normally distributed and the error vectors obtained at different dates are assumed independent. The matrix S_{t-1} is defined as follows

$$S_{t-1} = \begin{pmatrix} A_{t-1} & C_{t-1} \\ 0_n & I_p \end{pmatrix}$$

where A_{t-1} is a $(n \times n)$ matrix of coefficients relating the n state variables at time t to the state variables at time $t - 1$, C_{t-1} is a $(n \times p)$ matrix of coefficients relating the n state variables at time t to the p parameters, 0_n is a $(p \times n)$ matrix of zero, I_p is a $(p \times p)$ identity matrix. With this matrix structure, the state variables at time t are related to the state variables at time $t-1$ by

$$\begin{bmatrix} Z_t^{(1)} \\ \dots \\ Z_t^{(n)} \end{bmatrix} = A_{t-1} \begin{bmatrix} Z_{t-1}^{(1)} \\ \dots \\ Z_{t-1}^{(n)} \end{bmatrix} + C_{t-1} \begin{bmatrix} \theta^{(1)} \\ \dots \\ \theta^{(p)} \end{bmatrix} + \begin{bmatrix} B_{t-1}^{(1)} \\ \dots \\ B_{t-1}^{(n)} \end{bmatrix} + \begin{bmatrix} \varepsilon_{t-1}^{(1)} \\ \dots \\ \varepsilon_{t-1}^{(n)} \end{bmatrix}.$$

The measurement equation is

$$M_t = R_t \varphi_t + \tau_t \tag{8}$$

where M_t is a vector of q measurements, R_t is a $q \times (n+p)$ matrix relating the measurements to the state variables and parameters, τ_t is a vector of q error terms. τ_t is assumed independent of ε_t, and normally distributed with zero expectation. We have $E(M_t) = R_t E(\varphi_t)$ and $\text{var}(M_t) = R_t \text{var}(\varphi_t) R_t^{\mathrm{T}} + \text{var}(\tau_t)$.

Under these assumptions, the distribution of φ_t conditionally to $M_{1:t-1}$ and the distribution of φ_t conditionally to $M_{1:t}$ are both normal with expectations and variances defined by (Sullivan, 1992):

$$\begin{aligned}
E(\varphi_t|M_{1:t-1}) &= S_{t-1} E(\varphi_{t-1}|M_{1:t-1}) + B_{t-1} \\
\text{var}(\varphi_t|M_{1:t-1}) &= S_{t-1} \text{var}(\varphi_{t-1}|M_{1:t-1}) S_{t-1}^{\mathrm{T}} + \text{var}(\varepsilon_{t-1}) \\
E(\varphi_t|M_{1:t}) &= E(\varphi_t|M_{1:t-1}) + \text{var}(\varphi_t|M_{1:t-1}) \\
&\quad \times R_t^{\mathrm{T}} [R_t \text{var}(\varphi_t|M_{1:t-1}) R_t^{\mathrm{T}} + \text{var}(\tau_t)]^{-1} [M_t - R_t E(\varphi_t|M_{1:t-1})] \\
\text{var}(\varphi_t|M_{1:t}) &= \text{var}(\varphi_t|M_{1:t-1}) - \text{var}(\varphi_t|M_{1:t-1}) \\
&\quad \times R_t^{\mathrm{T}} [R_t \text{var}(\varphi_t|M_{1:t-1}) R_t^{\mathrm{T}} + \text{var}(\tau_t)]^{-1} R_t \text{var}(\varphi_t|M_{1:t-1})
\end{aligned}$$

To apply the above equations in any particular case we need to specify the vector φ_t, the distribution of φ_t at $t=1$, the matrix S_{t-1}, the matrix R_t, and the variance–covariance matrix $\text{var}(\varepsilon_t)$ and $\text{var}(\tau_t)$.

2.2.2. Application

We consider the practical problem described in Section 2.1.2. We treat here the problem of estimating simultaneously the state variable Z_t and one parameter, namely E_{b}. The vector φ_t is thus defined by $\varphi_t = \begin{pmatrix} Z_t \\ E_{\mathrm{b}} \end{pmatrix}$. We assume that $Z_1 = 0$ and that $E_{\mathrm{b}} \sim N(\mu_{E_{\mathrm{b}}}, \sigma_{E_{\mathrm{b}}}^2)$. The transition matrix S_{t-1} is defined by $S_{t-1} = \begin{bmatrix} 1 & C_{t-1} \\ 0 & 1 \end{bmatrix}$ with

$C_{t-1} = E_{\mathrm{i\,max}}[1 - \mathrm{e}^{-K\,LAI_{t-1}}]PAR_{t-1}$ and B_{t-1} is set equal to zero. We assume that $\varepsilon_t = \begin{pmatrix}\varepsilon_t^{(1)}\\ 0\end{pmatrix}$ with $\varepsilon_t^{(1)} \sim N(0, Q)$. The biomass measured at time t, M_t, is related to φ_t by $M_t = R\varphi_t + \tau_t$ where $R = (1, 0)$ and τ_t is the measurement error distributed as $\tau_t \sim N[0, \mathrm{var}(\tau_t)]$.

A numerical application is presented below. In this application, the expected value of $E_\mathrm{b}(\mu_{E_\mathrm{b}})$ is set equal to its nominal value 1.85 $\mathrm{g} \cdot \mathrm{MJ}^{-1}$ and Q is fixed to 10 $\mathrm{g}^2 \cdot \mathrm{m}^{-4}$. Three different values are considered successively for the variance of $E_\mathrm{b}(\sigma^2_{E_\mathrm{b}})$: 2.3×10^{-1}, 2.3×10^{-3}, 2.3×10^{-5} $\mathrm{g}^2 \cdot \mathrm{MJ}^{-2}$. The results are displayed in Figures 4–6. Figure 4 shows the distribution of φ_{t_4} conditionally to $M_{1:t_3}$ (the first three measurements) and the distribution of φ_{t_4} conditionally to $M_{1:t_4}$ (the first four measurements) when $\sigma^2_{E_\mathrm{b}}$ is set equal to 2.3×10^{-1} $\mathrm{g}^2 \cdot \mathrm{MJ}^{-2}$.

The distribution of φ_{t_4} conditionally to $M_{1:t_3}$ (Fig. 4a) is defined by $\begin{pmatrix} Z_{t_4} \\ E_\mathrm{b} \end{pmatrix} | M_{1:t_3} \Big) \sim N\left[\begin{pmatrix} 58.9 \\ 0.72 \end{pmatrix}, \begin{pmatrix} 314.8 & 3.84 \\ 3.84 & 0.079 \end{pmatrix}\right]$. The expected value of the biomass distribution, 58.9 $\mathrm{g} \cdot \mathrm{m}^{-2}$, is very different from the biomass measured at time t_4, $M_{t_4} = 37.4$ $\mathrm{g} \cdot \mathrm{m}^{-2}$. Note also that Z_{t_4} and E_b are positively correlated.

The distribution of φ_{t_4} conditionally to $M_{1:t_4}$ (Fig. 4b) is defined by $\Big(\begin{matrix} Z_{t_4} \\ E_\mathrm{b} \end{matrix} | M_{1:t_4} \Big) \sim N\left[\begin{pmatrix} 39.31 \\ 0.48 \end{pmatrix}, \begin{pmatrix} 27.49 & 0.34 \\ 0.34 & 0.036 \end{pmatrix}\right]$. After the fourth measurement, the expected values of Z_{t_4} and E_b are equal to 39.31 and 0.48, respectively. These values are lower than the expected values obtained before the fourth measurements (58.9 and 0.72, respectively) and the expected biomass value is near the measured biomass value. This is due to the positive correlation between Z_{t_4} and E_b and, also, to the low value of the measurement M_{t_4}. Another result is that the variances are much lower after the fourth measurement than before.

Figures 5 and 6 show the expected values of Z_t and E_b between sowing and harvest for $Q = 10$ $\mathrm{g}^2 \cdot \mathrm{m}^{-4}$ and $\sigma^2_{E_\mathrm{b}} = 2.3 \times 10^{-1}$, $\sigma^2_{E_\mathrm{b}} = 2.3 \times 10^{-3}$, $\sigma^2_{E_\mathrm{b}} = 2.3 \times 10^{-5}$ $\mathrm{g}^2 \cdot \mathrm{MJ}^{-2}$. The expected values are computed by using the first ten measurements. The results depend highly on the value of $\sigma^2_{E_\mathrm{b}}$. When $\sigma^2_{E_\mathrm{b}}$ is fixed to a low value, the expected value of E_b

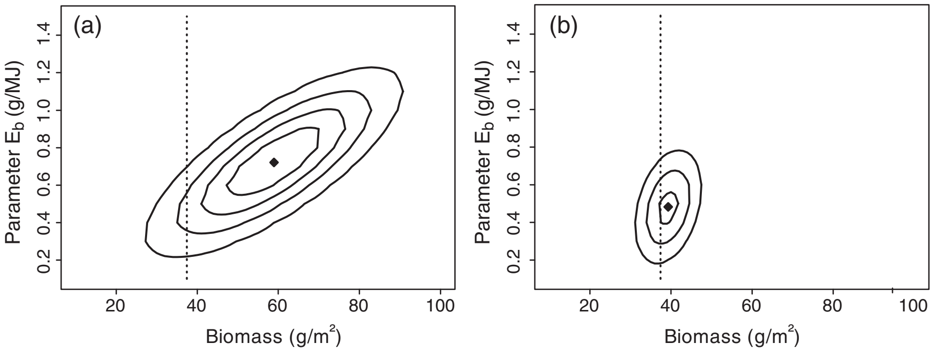


Figure 4. Distribution of $\varphi_{t_4}|M_{1:t_3}$ (a) and $\varphi_{t_4}|M_{1:t_4}$ (b) at the date of the fourth measurement ($t_4 = 125$ days) for $\sigma^2_{E_\mathrm{b}} = 0.2304$ $\mathrm{g}^2 \cdot \mathrm{MJ}^{-2}$. The dotted line indicates the biomass value measured at time t_4. The black points indicate the expected values of the distributions.

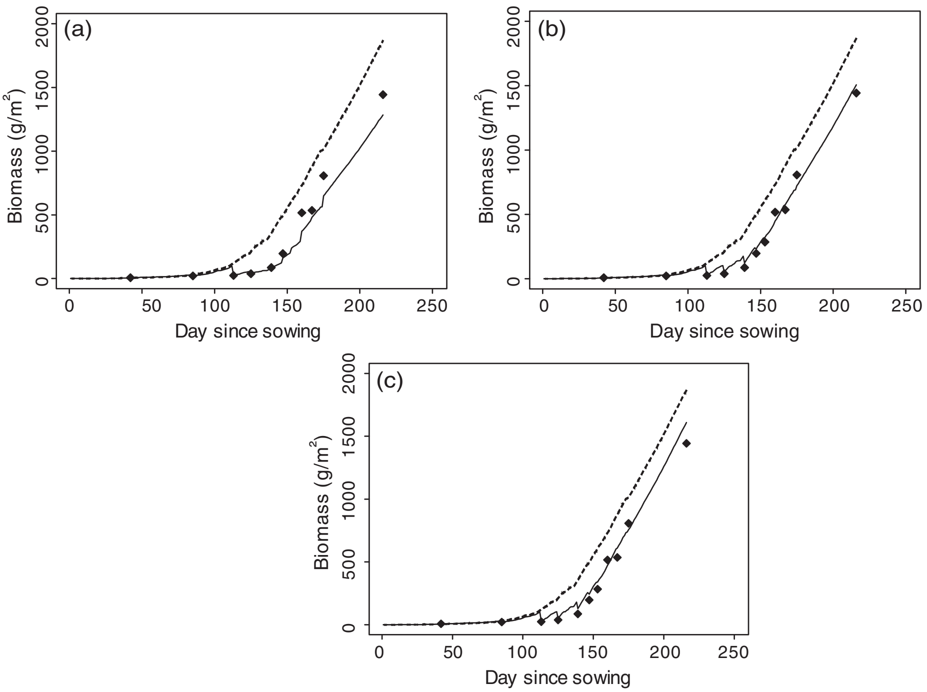


Figure 5. Expected biomass values (continuous line) obtained with three different values of $\sigma^2_{E_\mathrm{b}}$: 2.3×10^{-1} g$^2 \cdot$ MJ^{-2} (a), 2.3×10^{-3} g$^2 \cdot$ MJ^{-2} (b), and 2.3×10^{-5} g$^2 \cdot$ MJ^{-2} (c). The black points represent biomass measurements. Biomass values were updated using the first 10 measurements. The 11th measurement is the value measured at harvest ($t = 216$ days). The dashed lines represent the initial predictions.

remains always very close to the nominal value of the parameter (1.85 g $\cdot$ MJ^{-1}) (Fig. 6c) and the expected values of Z_t (Fig. 5c) are very close to the values obtained when only Z_t is updated (Fig. 3b). This result is logical because a low value for $\sigma^2_{E_\mathrm{b}}$ means that we put a lot of confidence in the nominal value of the parameter. In this case, E_b is almost not modified by the measurements and only the state variable is updated. On the contrary, if $\sigma^2_{E_\mathrm{b}}$ is set equal to a high value, the parameter is strongly corrected at each date of measurement (Fig. 6a) and the expected biomass values differ from the values reported in Figure 3b.

3. Data assimilation for non-linear models

3.1. Methods

The dynamic equations of crop models are almost never linear in the state variables nor in the parameters. As a result the state variables do not have a normal distribution, even if

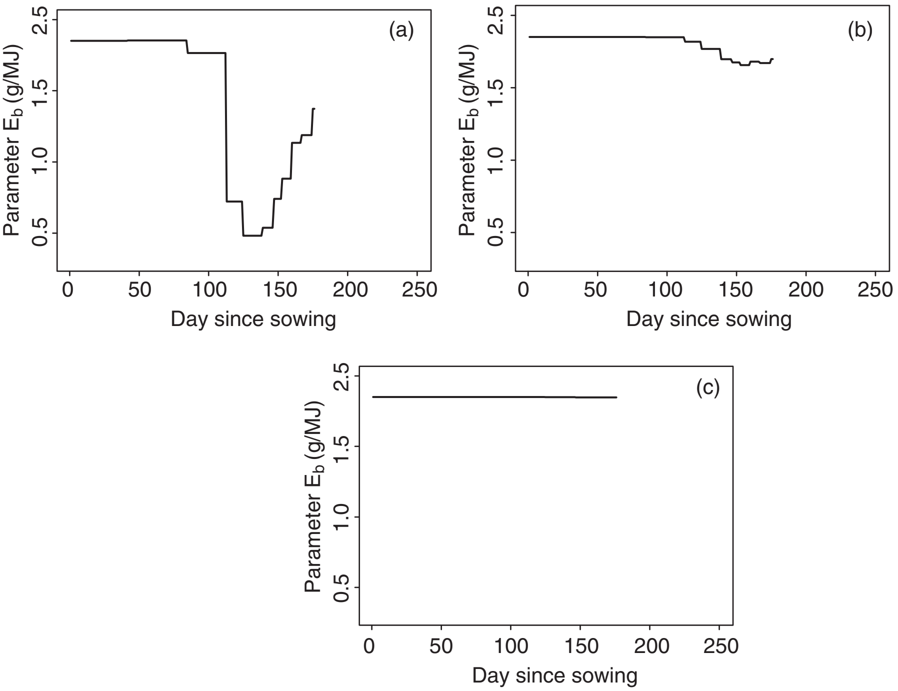


Figure 6. Expected values of E_b obtained with three different values of $\sigma^2_{E_b}$ equal to 2.3×10^{-1} (a), 2.3×10^{-3} (b), and 2.3×10^{-5} $g^2 \cdot MJ^{-2}$ (c).

all the errors are normally distributed. Here, we present two methods to handle non-linear models, namely the extended Kalman filter and the Ensemble Kalman filter.

3.1.1. Extended Kalman filter

The dynamic model is now defined by:

$$\varphi_t = F(\varphi_{t-1}) + \varepsilon_{t-1} \tag{9}$$

where φ_t is the $(n + p)$ vector including the state variable and parameter values at time t, F is a series of arbitrary functions, one for each component of φ_t, and ε_t is a vector of errors. In the previous section, we considered a particular case with $F(\varphi_t) = S_t\varphi_t$. Here, we consider a more general case where $F(\varphi_t)$ can be non-linear. As before, we assume that the measurement M_t is related to φ_t by $M_t = R_t\varphi_t + \tau_t$ where M_t is a vector of q measurements, R_t is a $q \times (n + p)$ matrix relating the measurements to the state variables and parameters, τ_t is a vector of q error terms. τ_t is assumed independent of ε_t, and normally distributed with zero expectation.

When F is non-linear, it is generally impossible to determine the analytic expression of the distribution φ_t conditionally to M_t, but several methods have been developed to approximate this distribution. A first method is called Extended Kalman filter (e.g. Welch and Bishop, 2002; Pastres et al., 2003). The principle is to linearize Eq. (9) and to apply the standard Kalman filter method to the following model:

$$\varphi_t \approx F(\hat{\varphi}_{t-1}) + H_{t-1}(\varphi_{t-1} - \hat{\varphi}_{t-1}) + w_{t-1}$$

where H_t is a $(n + p) \times (n + p)$ matrix of partial derivatives of F with respect to the $n + p$ elements of φ_t, $\hat{\varphi}_t$ is the predicted state variable at time t, $\hat{\varphi}_t = \hat{E}(\varphi_t|M_{1:t})$, w_t is a $(n + p)$ error term vector assumed to be normally distributed. The main drawback of this method is that the linearization has been shown to be a poor approximation in a number of applications. The linear approximation may not give a good description of how the model errors evolve over time. This method is illustrated in a case study presented in Chapter 18 of this book (Jones and Graham).

3.1.2. Ensemble Kalman filter

The Ensemble Kalman filter is another popular method described by Burgers et al. (1998). The principle is to approximate the probability distributions using random samples of state variable and parameter values. First, an ensemble of N values of φ_t, $\varphi_t^1, \ldots, \varphi_t^j, \ldots, \varphi_t^N$ and an ensemble of N values of $M_t + \tau_t$, $M_t^1, \ldots, M_t^j, \ldots, M_t^N$, are randomly generated. Second, the Kalman filter equation is applied to each ensemble element as follows:

$$\varphi_{t,K}^j = \varphi_t^j + K_t^{\mathrm{e}}(M_t^j - R_t\varphi_t^j) \tag{10}$$

where K_t^{e} is a $(n+p) \times q$ matrix defined by $K_t^{\mathrm{e}} = \Sigma_t^{\mathrm{e}} R_t^{\mathrm{T}}[R_t\Sigma_t^{\mathrm{e}}R_t^{\mathrm{T}} + \mathrm{var}(\tau_t)]^{-1}$, $\mathrm{var}(\tau_t)$ is the variance–covariance matrix of the measurement error, and Σ_t^{e} is the $(n + p) \times (n+p)$ variance–covariance matrix of N vectors φ_t^j, $j = 1, \ldots, N$. The ensemble of state variables φ_t^j, $j = 1, \ldots, N$, describes the uncertainty in the state variable and parameter values before using the measurement M_t. In this approach, the updated model prediction is set equal to the average value of $\varphi_{t,K}^j$, $j = 1, \ldots, N$, noted further $\bar{\varphi}_{t,K}$. Note that $\bar{\varphi}_{t,K}$ is related to the average value, $\bar{\varphi}_t$, of the initial ensemble φ_t^j, $j = 1, \ldots, N$, by

$$\bar{\varphi}_{t,K} = \bar{\varphi}_t + K_t^{\mathrm{e}}(\bar{M}_t - R_t\bar{\varphi}_t) \tag{11}$$

where $\bar{M}_t$ is the average value of M_t^j, $j = 1, \ldots, N$. The attractive feature of this method is that its implementation does not require a linear approximation of the crop model. However, it is necessary to choose the value of N, to define a procedure for generating φ_t^j, $j = 1, \ldots, N$, and to define another procedure for generating M_t^j, $j = 1, \ldots, N$. The values of M_t^j, $j = 1, \ldots, N$, can be simply generated by adding random terms to M_t: $M_t^j = M_t + \tau_t^j$ with $\tau_t^j \sim N[0, \mathrm{var}(\tau_t)]$ (Burgers et al., 1998). This is straightforward if $\mathrm{var}(\tau_t)$ is known. On the contrary, there is no systematic method for choosing N and for generating the ensemble of vectors φ_t^j, $j = 1, \ldots, N$. The value of N must be chosen carefully. Too small ensemble can give very poor approximation. Moreover, according to Burgers et al. (1998), the matrix Σ_t^{e} tends to underestimate the true error variance–covariance matrix when N is too small. For generating φ_t^j, $j = 1, \ldots, N$, a common approach consists in calculating N vectors of state variables and parameters at each time step as follows (e.g. Allen et al., 2002):

$$\varphi_t^j = F(\varphi_{t-1}^j) + \varepsilon_{t-1}^j$$

where $\varepsilon_{t-1}^j \sim N[0, \operatorname{var}(\varepsilon_{t-1})]$. The procedure requires the knowledge of $\operatorname{var}(\varepsilon_{t-1})$. As before, different values of $\operatorname{var}(\varepsilon_{t-1})$ can be tested by using a training data set. Another approach is to generate randomly N values for all the uncertain elements of the crop models (parameters and input variables) (Margulis et al., 2002). Several other filters have been developed for non-linear models like, for instance, the particle filter. See Anderson and Anderson (1999) and Doucet et al. (2000) for more details.

3.2. Application

We present two numerical applications of the Ensemble Kalman filter based on the model simulating winter wheat biomass described in Sections 2.1.2 and 2.2.2. In the first application, the Ensemble Kalman filter is used to approximate the distribution of $\varphi_t | M_t$ when $\varphi_t = (Z_t, E_{\mathrm{b}})^{\mathrm{T}}$. The objective of this first application is to evaluate the capabilities of the method for approximating the distribution. As $\varphi_t = (Z_t, E_{\mathrm{b}})^{\mathrm{T}}$ depends on φ_{t-1} through the linear function Eq. (7), it is possible in this case to calculate the analytical expression of the distribution of $\varphi_t | M_t$ by using the standard Kalman filter as shown in Section 2.2.2. This distribution is compared to the approximated distributions obtained with the Ensemble Kalman filter for different N values. The second application shows how to use the Ensemble Kalman filter to estimate simultaneously Z_t and more than one parameter.

3.2.1. Application 1: analysis of the performance of the Ensemble Kalman filter

We update the state variable and the single parameter E_{b}. The model is linear with respect to both of these, so that we can calculate the exact results and compare with the Ensemble Kalman filter. In particular, we explore the effect of different choices for N, the size of the random samples.

Consider the $\varphi_{t_1} = (Z_{t_1}, E_{\mathrm{b}})^{\mathrm{T}}$ at the date of the first measurement M_{t_1}. The Ensemble Kalman filter method is used to approximate the distribution of $\varphi_{t_1} = (Z_{t_1}, E_{\mathrm{b}})^{\mathrm{T}}$ conditionally to M_{t_1} as follows:

(1) Generate an ensemble of N values of parameter E_{b} from $N(\mu_{E_{\mathrm{b}}}, \sigma_{E_{\mathrm{b}}}^2)$. The values are noted $\{E_{\mathrm{b}}^1, \ldots, E_{\mathrm{b}}^j, \ldots, E_{\mathrm{b}}^N\}$.

(2) Generate an ensemble of N values of biomass at the date of the first measurement t_1, $\{Z_{t_1}^1, \ldots, Z_{t_1}^j, \ldots, Z_{t_1}^N\}$. Each $Z_{t_1}^j$, $j = 1, \ldots, N$, is calculated as

$Z^j_{t_1} = \sum_{t=1}^{t_1-1} E^j_{\mathrm{b}} E_{\mathrm{i\,max}}[1 - \mathrm{e}^{-KLAI_t}]PAR_t + \sum_{t=1}^{t_1-1} \varepsilon^j_t$ where E^j_{b} is one of the N values of E_{b} generated at step 1 and ε^j_t is an error term randomly generated from $\varepsilon^j_t \sim N(0, Q)$. Define $\varphi^j_{t_1} = (Z^j_{t_1}, E^j_{\mathrm{b}})^{\mathrm{T}}$, $j = 1, \ldots, N$.

(3) Compute the (2×2) variance–covariance matrix $\Sigma^{\mathrm{e}}_{t_1}$ of $\varphi^j_{t_1} = (Z^j_{t_1}, E^j_{\mathrm{b}})^{\mathrm{T}}$ from the ensembles, $\Sigma^{\mathrm{e}}_{t_1} = \begin{bmatrix} \mathrm{var}(Z^j_{t_1}) & \mathrm{cov}(Z^j_{t_1}, E^j_{\mathrm{b}}) \\ \mathrm{cov}(Z^j_{t_1}, E^j_{\mathrm{b}}) & \mathrm{var}(E^j_{\mathrm{b}}) \end{bmatrix}$.

(4) Generate an ensemble of N observations $M^1_{t_1}, \ldots, M^j_{t_1}, \ldots, M^N_{t_1}$. Each $M^j_{t_1}$, $j = 1, \ldots, N$, is calculated as $M^j_{t_1} = M_{t_1} + \tau^j_{t_1}$ where $\tau^j_{t_1}$ is drawn from $N[0, \mathrm{var}(\tau_{t_1})]$.

(5) Update $\varphi^j_{t_1}$, $j = 1, \ldots, N$, as follows: $\varphi^j_{t_1,K} = \varphi^j_{t_1} + K^{\mathrm{e}}_{t_1}(M^j_{t_1} - R\varphi^j_{t_1})$ where $\varphi^j_{t_1} = (Z^j_{t_1}, E^j_{\mathrm{b}})^{\mathrm{T}}$, $K^{\mathrm{e}}_{t_1} = \Sigma^{\mathrm{e}}_{t_1} R^{\mathrm{T}}[R\Sigma^{\mathrm{e}}_{t_1} R^{\mathrm{T}} + \mathrm{var}(\tau_{t_1})]^{-1}$, $R = (1, 0)$, and $\varphi^j_{t_1,K}$ is an updated value of $\varphi^j_{t_1}$.

(6) Estimate the expected value and variance–covariance matrix of $\varphi_{t_1} | M_{t_1}$ by $\frac{1}{N}\sum_{j=1}^{N} \varphi^j_{t_1,K}$ and $\mathrm{var}(\varphi^j_{t_1,K})$, respectively.

(7) Replace $\varphi^j_{t_1}$ by $\varphi^j_{t_1,K}$, $j = 1, \ldots, N$.

This procedure is applied successively to the ten measurements. The calculations give values of biomass and of the parameter E_{b} for each date between sowing and harvest and for each of the N elements of the ensemble. For illustration, Figure 7 shows three elements of the ensemble of biomass and E_{b} generated by the Ensemble Kalman filter method with $N = 100$, $Q = 10\,\mathrm{g}^2 \cdot \mathrm{m}^{-4}$, $\mu_{E_{\mathrm{b}}} = 1.85\,\mathrm{g} \cdot \mathrm{MJ}^{-1}$, and $\sigma^2_{E_{\mathrm{b}}} = 2.3 \times 10^{-1}\,\mathrm{g}^2 \cdot \mathrm{MJ}^{-2}$. One hundred values of biomass and of E_{b} are generated every day. If a measurement is available at time t, each one of the 100 values is updated by using the Kalman filter equation as explained above. Figure 8 shows the 100 biomass and E_{b} values obtained with the Ensemble Kalman filter at the date of the fourth measurement $t = t_4$. These values can be used to approximate the distribution of φ_{t_4} conditionally to $M_{1:t_4}$. The expected value of the distribution can be estimated by averaging the 100 values of biomass and E_{b}.

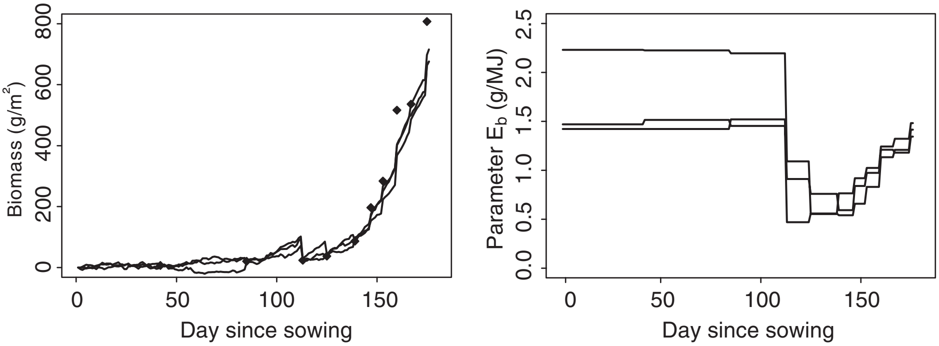


Figure 7. Three elements of the ensemble of biomass and E_{b} values generated by the Ensemble Kalman filter method with $N = 100$, $Q = 10\ \mathrm{g}^2 \cdot \mathrm{m}^{-4}$, $\mu_{E_{\mathrm{b}}} = 1.85\ \mathrm{g} \cdot \mathrm{MJ}^{-1}$, and $\sigma^2_{E_{\mathrm{b}}} = 2.3 \times 10^{-1}\ \mathrm{g}^2 \cdot \mathrm{MJ}^{-2}$.

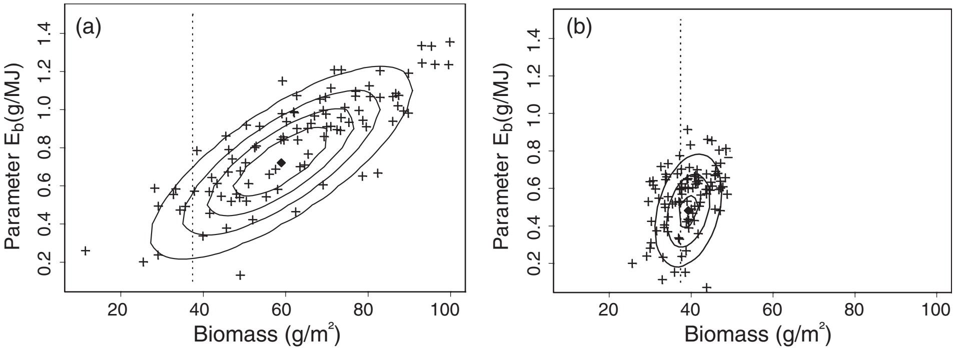


Figure 8. Distribution of φ_{t_4} conditionally to $M_{1:t_3}$ (a) and conditionally to $M_{1:t_4}$ (b) at the date of the fourth measurement ($t_4 = 125$ days). The ellipses represent the true distributions. The black points indicate the true expected values of the distributions. The crosses are the values generated with the Ensemble Kalman filter. The dotted line indicates the biomass value measured at time t_4.

The performance of the Ensemble Kalman filter method depends on the ensemble size N. To study the effect of N on the accuracy of the results, the Ensemble Kalman filter is now applied with 11 different N values in the range 5–100. For each size N, the N values of biomass and E_b obtained with the Ensemble Kalman filter at the date of the fourth measurement are averaged. The average values are then compared to the true expected value $E(\varphi_{t_4}|M_{1:t_4})$. These true values were calculated with the standard Kalman filter, leading to $E(Z_{t_4}|M_{t_4}) = 39.31 \text{ g} \cdot \text{m}^{-2}$ and $E(E_b|M_{t_4}) = 0.482 \text{ g} \cdot \text{MJ}^{-1}$. The Ensemble Kalman filter is run 100 times for each N value, giving 100 different estimates of $E(\varphi_{t_4}|M_{1:t_4})$ for each size N. The estimates are noted $\hat{\mu}_1, \ldots, \hat{\mu}_k, \ldots, \hat{\mu}_{100}$. The quality of the approximation is then evaluated by calculating a root mean squared error (*RMSE*) for each N value from the differences between the true expected values and the 100 different approximations obtained with the Ensemble Kalman filter as $RMSE = \sqrt{\frac{1}{100}\sum_{k=1}^{100}[E(\varphi_{t_4}|M_{1:t_4}) - \hat{\mu}_k]^2}$. The results are displayed in Figure 9. This

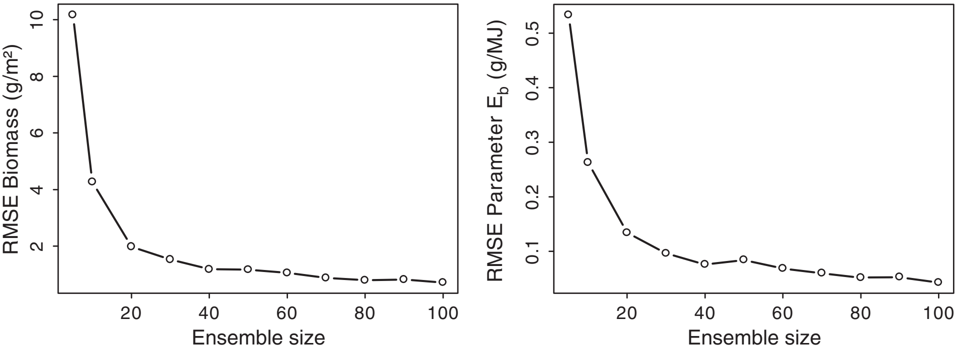


Figure 9. RMSE values obtained from the differences between the true expected value $E(\varphi_{t_4}|M_{1:t_4})$ and the average of N biomass and E_b values generated by the Ensemble Kalman filter. Eleven different values of N were considered successively. For each N value, the Ensemble Kalman filter was run 100 times.

figure shows that the estimation is quite inaccurate when the ensemble size is equal to 5. Increasing the ensemble size strongly decreases the *RMSE* and, so, improves the accuracy of the estimation of the expected values of the distribution. There is essentially no further improvement beyond about $N = 70$ (Fig. 9). These results show that the accuracy of the estimation obtained with the Ensemble Kalman filter depends highly on the size of the ensemble. This parameter must be chosen carefully.

3.2.2. Application 2: application of the Ensemble Kalman filter to estimate Z_t and three parameters

In this second application, the Ensemble Kalman filter is used to estimate Z_t and three parameters, namely E_b, A, and B. These parameters were selected on the basis of the sensitivity indices calculated in Chapter 3. It was shown that E_b, A, and B explain more than 80% of the total biomass variability. We present below an algorithm to approximate the distribution of $\varphi_{t_1}|M_{t_1}$ with $\varphi_{t_1} = (Z_{t_1}, E_b, A, B)^T$:

(1) Generate an ensemble of N values of parameters E_b, A, and B from $N(\mu_{E_b}, \sigma^2_{E_b})$, $N(\mu_A, \sigma^2_A)$, and $N(\mu_B, \sigma^2_B)$. The values noted are: $\{E_b^1, \ldots, E_b^j, \ldots, E_b^N\}$, $\{A^1, \ldots, A^j, \ldots, A^N\}$, and $\{B^1, \ldots, B^j, \ldots, B^N\}$.

(2) Generate an ensemble of N values of biomass at the date of the first measurement t_1, $\{Z^1_{t_1}, \ldots, Z^j_{t_1}, \ldots, Z^N_{t_1}\}$. Each $Z^j_{t_1}$, $j = 1, \ldots, N$, is calculated as $Z^j_{t_1} = \sum_{t=1}^{t_1-1} E_b^j E_{i\max}[1 - e^{-K\,LAI_t^j}]PAR_t + \sum_{t=1}^{t_1-1} \varepsilon_t^j$ where ε_t^j is an error term randomly generated from $\varepsilon_t^j \sim N(0, Q)$, $LAI_t^j = L_{\max}\{1/(1 + e^{-A^j[T_t - T_1]}) - e^{B^j[T_t - T_2]}\}$, E_b^j, A^j, B^j are the *j*th parameter values generated at Step 1. Define $\varphi^j_{t_1} = (Z^j_{t_1}, E_b^j, A^j, B^j)^T$, $j = 1, \ldots, N$.

(3) Compute the (4×4) variance–covariance matrix $\Sigma^e_{t_1}$ of $\varphi^j_{t_1} = (Z^j_{t_1}, E^j_b, A^j, B^j)^T$ from the four ensembles:

$$\Sigma^e_{t_1} = \begin{bmatrix} \mathrm{var}(Z^j_{t_1}) & \mathrm{cov}(Z^j_{t_1}, E^j_b) & \mathrm{cov}(Z^j_{t_1}, A^j) & \mathrm{cov}(Z^j_{t_1}, B^j) \\ \mathrm{cov}(Z^j_{t_1}, E^j_b) & \mathrm{var}(E^j_b) & \mathrm{cov}(A^j, E^j_b) & \mathrm{cov}(B^j, E^j_b) \\ \mathrm{cov}(Z^j_{t_1}, A^j) & \mathrm{cov}(A^j, E^j_b) & \mathrm{var}(A^j) & \mathrm{cov}(A^j, B^j) \\ \mathrm{cov}(Z^j_{t_1}, B^j) & \mathrm{cov}(B^j, E^j_b) & \mathrm{cov}(A^j, B^j) & \mathrm{var}(B^j) \end{bmatrix}.$$

(4) Generate an ensemble of N observations $M^1_{t_1}, \ldots, M^j_{t_1}, \ldots, M^N_{t_1}$. Each $M^j_{t_1}$, $j = 1, \ldots, N$, is calculated as $M^j_{t_1} = M_{t_1} + \tau^j_{t_1}$ where $\tau^j_{t_1}$ is drawn from $N[0, \mathrm{var}(\tau_{t_1})]$.

(5) Update $\varphi^j_{t_1}$, $j = 1, \ldots, N$, as follows: $\varphi^j_{t_1,K} = \varphi^j_{t_1} + K^e_{t_1}(M^j_{t_1} - R\varphi^j_{t_1})$ where $\varphi^j_{t_1} = (Z^j_{t_1}, E^j_b, A^j, B^j)^T$, $K^e_{t_1} = \Sigma^e_{t_1} R^T [R \Sigma^e_{t_1} R^T + \mathrm{var}(\tau_{t_1})]^{-1}$, $R = (1, 0, 0, 0)$ and $\varphi^j_{t_1,K}$ is an updated value of $\varphi^j_{t_1}$.

(6) Estimate the expected value and variance–covariance matrix of $\varphi_{t_1}|M_{t_1}$ by $\frac{1}{N}\sum_{j=1}^{N} \varphi^j_{t_1,K}$ and $\mathrm{var}(\varphi^j_{t_1,K})$, respectively.

(7) Replace $\varphi^j_{t_1}$ by $\varphi^j_{t_1,K}$, $j = 1, \ldots, N$.

This algorithm can be applied to update simultaneously the biomass and the three parameters at each date of measurement. Its implementation requires the knowledge of Q, μ_{E_b}, $\sigma^2_{E_\mathrm{b}}$, μ_A, σ^2_A, μ_B, σ^2_B, and N. We further assume that $Q = 10\,\mathrm{g}^2\cdot\mathrm{m}^{-4}$, $\mu_{E_\mathrm{b}} = 1.85\,\mathrm{g}\cdot\mathrm{MJ}^{-1}$, $\sigma^2_{E_\mathrm{b}} = 2.3\times10^{-1}\,\mathrm{g}^2\cdot\mathrm{MJ}^{-2}$, $\mu_A = 6.5\times10^{-3}\,\mathrm{g}\cdot\mathrm{MJ}^{-1}$, $\sigma^2_A = 3.17\times10^{-3}\,\mathrm{g}^2\cdot\mathrm{MJ}^{-2}$, $\mu_B = 2.05\times10^{-3}\,\mathrm{g}\cdot\mathrm{MJ}^{-1}$, $\sigma^2_B = 2.352\times10^{-3}\,\mathrm{g}^2\cdot\mathrm{MJ}^{-2}$. The value of N cannot be determined with the procedure described in the first application because, here, the true expected values are unknown. In such cases, a common approach consists in running the Ensemble Kalman filter several times for different N values and in calculating the variance (or standard error) of the estimated expected values for each N value. In this application, the Ensemble Kalman filter is run 100 times for $N = 10$, 50, 100, 150, and 200. The standard errors of the estimated expected values are then calculated at harvest for each N value. The results are shown in Figure 10. The standard errors are very high when $N = 5$. Increasing the ensemble size strongly decreases the standard errors and, so, improves the accuracy of the estimation of the expected values. There is no further improvement beyond about $N = 100$ for the biomass, E_b, and B. But this size is not sufficient for parameter A. It is necessary to use $N = 200$ to avoid inaccurate values of A. Figure 11 shows the results obtained in one of the run of the Ensemble Kalman filter when $N = 200$.

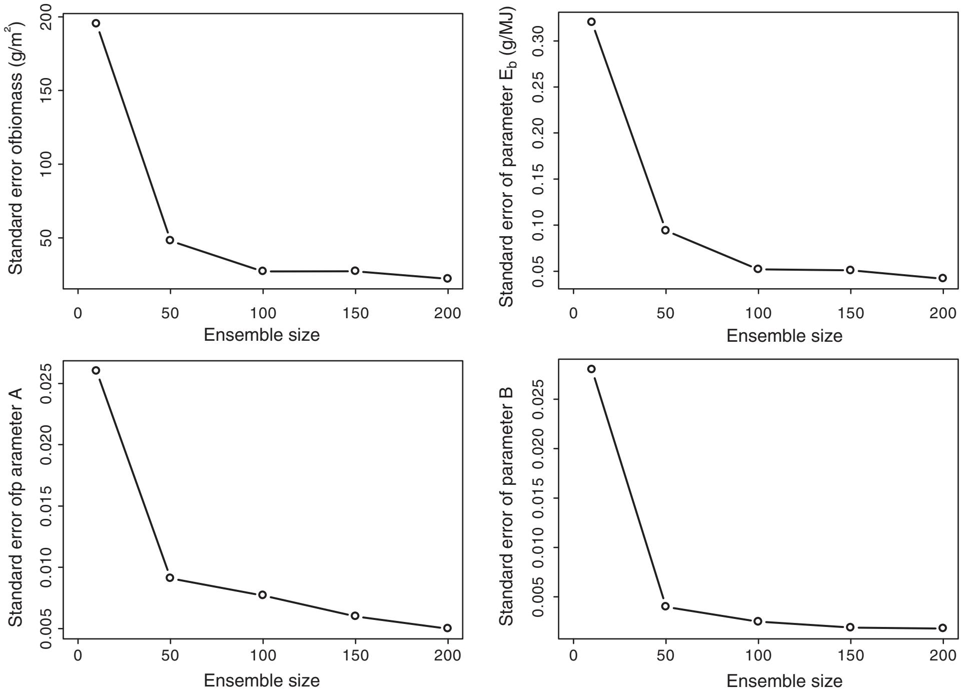


Figure 10. Standard error of the estimated expected values of biomass, E_b, A, and B obtained at harvest with the Ensemble Kalman filter. Five ensemble sizes in the range 5–200 were considered successively. For each size, the Ensemble Kalman filter was run 100 times.

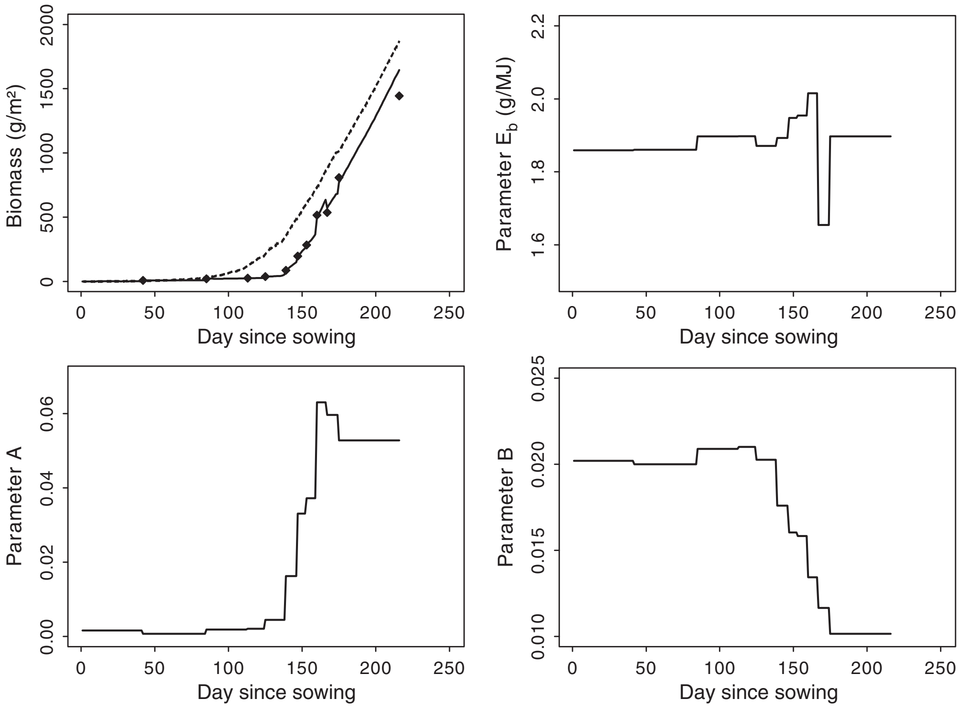


Figure 11. Expected values of biomass, E_b, A, and B (continuous line) obtained with the Ensemble Kalman filter ($N = 200$, $Q = 10\ \mathrm{g}^2 \cdot \mathrm{m}^{-4}$, $\mu_{E_b} = 1.85\,\mathrm{g} \cdot \mathrm{MJ}^{-1}$, $\sigma^2_{E_b} = 2.3 \times 10^{-1}\,\mathrm{g}^2 \cdot \mathrm{MJ}^{-2}$, $\mu_A = 6.5 \times 10^{-3}\,\mathrm{g} \cdot \mathrm{MJ}^{-1}$, $\sigma^2_A = 3.17 \times 10^{-3}\,\mathrm{g}^2 \cdot \mathrm{MJ}^{-2}$, $\mu_B = 2.05 \times 10^{-3}\,\mathrm{g} \cdot \mathrm{MJ}^{-2}$, $\sigma^2_B = 2.352 \times 10^{-3}\,\mathrm{g}^2 \cdot \mathrm{MJ}^{-2}$). The black points represent biomass measurements. Biomass values and parameters were updated using the first 10 measurements. The 11th measurement is the value measured at harvest ($t = 216$ days). The dashed lines represent the initial predictions. The continuous lines represent the average of the 200 values generated by the Ensemble Kalman filter between sowing and harvest.

4. Conclusion

This chapter has concentrated on the mechanics of how to use measurements to update state variables or parameters of a crop model. The use of in-season measurements to improve predictions is potentially very powerful, but a number of difficult decisions lie with the user.

We have seen that the updating equations involve the variances of the errors in the model dynamic equations. However, these variances are seldom if ever explicitly given for a crop model. Determining reasonable values may be a difficult problem.

We have seen that one can update one or several state variables using the same measurement. Which and how many variables should one choose to update? This again can be a difficult decision.

We have seen that one can choose to update parameters in place of or in addition to updating state variables. The choice between updating parameters and state variables is important. Updating a state variable but not parameters implies that even though the past evolution of the field being studied was different than the evolution of an average field, the future evolution will obey the same equations as for the average field. If on the other hand parameters are updated, then one is assuming that the future will obey different equations than the average field. Furthermore, one is assuming that the past gives information about how the future evolution will differ from that of the average field.

Agronomic understanding of why the field being studied differs from an average field will obviously be a very valuable aid to making these decisions.

References

Allen, J.I., Eknes, M., Evensen, G., 2002. An ensemble Kalman filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan sea. Annales Geophysicae 20, 1–13.

Anderson, J.L., Anderson, S.L., 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Monthly Weather Review 127, 2741–2758.

Baret, F., 1986. Contribution au suivi radiométrique de cultures de céréales. Ph.D. Dissertation, Université d'Orsay, Orsay, France.

Burgers, G., van Leeuwen, P.J., Evensen, G., 1998. Analysis scheme in the ensemble Kalman filter. Monthly Weather Review 126, 1719–1724.

Déjean, S., Faivre, R., Goulard, M., 2002. Modèle nonlinéaire à paramètres aléatoires de la dynamique de cultures observées par télédétection: comparaison de deux procédures d'estimation. Journal de la société française de statistique 143, 205–213.

Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and Computing 10, 197–208.

Kalman, R.E., 1961. A new approach to linear filtering and prediction theory. Journal of Basic Engineering 83, 95–108.

Margulis, S.A., McLaughlin, D., Entekhabi, D., Dunne, S., 2002. Land data assimilation and estimation of soil moisture using measurements from the Southern Great Plains 1997 field experiment. Water Resources Research 38, 1299–1318.

Monteith, J.L., 1977. Climate and the efficiency of crop production in Britain. Philosophical Transactions of the Royal Society, London, B 281, 277–294.

Pastres, R., Ciavatta, S., Solidoro, C., 2003. The extended Kalman filter as a toll for the assimilation of high frequency water quality data. Ecological Modelling 170, 227–235.

Sullivan, P.J., 1992. A Kalman filter approach to catch-at-length analysis. Biometrics 48, 237–257.

Varlet-Grancher, C., Bonhomme, R., Chartier, M., Artis, P., 1982. Efficience de la conversion de l'énergie solaire par un couvert végétal, Acta Oecologica 3, 3–26.

Welch, G., Bishop, G., 2002. An introduction to the Kalman filter, Report TR-95-041, Department of Computer Science, University of North Carolina.

Exercises

1. The objective of this exercise is to estimate the nitrogen content of a crop, noted Z, by combining expert knowledge and a measurement noted M. Suppose that, according to an expert, the possible values of the nitrogen content can be described by a normal distribution, $Z \sim N(\mu_Z, \sigma_Z^2)$. We assume that a measurement M is performed in the field and that this measurement represents an index in the range 0–1 based on reflectance values measured for different wavelengths. The measurement M is related to the nitrogen content as follows:

   $$M = aZ + b + \tau$$

   where a and b are two parameters and $\tau \sim N(0, \sigma_M^2)$ is a measurement error independent from Z. We assume that μ_Z, σ_Z^2, a, b, and σ_M^2 are known.

   (a) What is the joint distribution of the vector of random variables $\theta = \binom{Z}{M}$? Express the expected value of θ and its variance–covariance matrix in terms of μ_Z, σ_Z^2, and σ_M^2.
   (b) Determine the distribution for Z conditionally to the measurement M. Express $E(Z|M)$ and $\text{var}(Z|M)$ in terms of M, μ_Z, σ_Z^2, and σ_M^2.
   (c) What is the effect of σ_Z^2 and σ_M^2 on $E(Z|M)$ and $\text{var}(Z|M)$?
   (d) Numerical application. Calculate $E(Z|M)$ and $\text{var}(Z|M)$ with $M = 0.9$, $\mu_Z = 0.25$, $\sigma_Z^2 = 0.01$, $a = 1.2$, $b = 0$, and $\sigma_M^2 = 0.001$.
   (e) Suppose that the expert is not very confident in the value of μ_Z. Perform a sensitivity analysis of $E(Z|M)$ to the value of μ_Z when μ_Z varies in the range 0.15–0.35.

2. In this second exercise, the expert knowledge is replaced by the prediction of a dynamic model defined as:

   $$Z_t = Z_{t-1} + B_{t-1} + \varepsilon_{t-1}$$

   where Z_t is the crop nitrogen content at time t, B_{t-1} is a known input variable, and ε_{t-1} is the model error. ε_{t-1} is normally distributed and has zero mean and constant variance, $\varepsilon_{t-1} \sim N(0, Q)$. We also assume that $Z_1 = 0$ and $\text{var}(Z_1) = 0$.

   Suppose that a measurement M is performed at $t = 10$ and that M is related to the nitrogen content as $M = aZ_{10} + b + \tau$ where a and b are two parameters and $\tau \sim N(0, \sigma_M^2)$ is a measurement error independent from Z.

   (a) What is the distribution for Z_{10} before using the measurement? Express $E(Z_{10})$ and $\text{var}(Z_{10})$ in terms of the input variable values and Q.
   (b) Determine the joint distribution for the vector of random variables $\theta_{10} = \begin{pmatrix} Z_{10} \\ M \end{pmatrix}$. Express the expected value of θ and its variance–covariance matrix in terms of $B_1, \ldots, B_9$, Q, and σ_M^2.
   (c) Give the distribution for Z_{10} conditionally to the measurement M. Express $E(Z_{10}|M)$ and $\text{var}(Z_{10}|M)$ in terms of M, $B_1, \ldots, B_9$, Q, and σ_M^2.
   (d) Compute the Kalman gain.

(e) What are the effects of Q and σ_M^2 on $E(Z_{10}|M)$ and $\mathrm{var}(Z_{10}|M)$?

(f) Numerical application. Calculate $E(Z_{10}|M)$, $\mathrm{var}(Z_{10}|M)$, and the Kalman gain with $B_1 = 0.01$, $B_2 = 0.015$, $B_3 = 0.02$, $B_4 = 0.035$, $B_5 = 0.02$, $B_6 = 0.025$, $B_7 = 0.025$, $B_8 = 0.018$, $B_9 = 0.02$, $Q = 0.001$, $a = 1.2$, $b = 0$, $M = 0.9$, and $\sigma_M^2 = 0.001$.

(g) Is the uncertainty higher with the expert knowledge or with the model?

3. In this exercise, we consider a more sophisticated dynamic model simulating two state variables noted $Z_t^{(1)}$ and $Z_t^{(2)}$. $Z_t^{(1)}$ is the crop nitrogen content at time t and $Z_t^{(2)}$ is the crop biomass at time t. The model is defined by:

$$\begin{pmatrix} Z_t^{(1)} \\ Z_t^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & c \\ 0 & 1 \end{pmatrix} \begin{pmatrix} Z_{t-1}^{(1)} \\ Z_{t-1}^{(2)} \end{pmatrix} + \begin{pmatrix} B_{t-1}^{(1)} \\ B_{t-1}^{(2)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{t-1}^{(1)} \\ \varepsilon_{t-1}^{(2)} \end{pmatrix}$$

where c is a known parameter, $B_{t-1}^{(1)}$ and $B_{t-1}^{(2)}$ are two known input variables, and $\varepsilon_{t-1}^{(1)}$ and $\varepsilon_{t-1}^{(2)}$ are two errors. We assume that $\varepsilon_{t-1}^{(1)}$ and $\varepsilon_{t-1}^{(2)}$ are normally distributed, independent, have zero means, and that their variances are equal to Q_1 and Q_2. We also assume that $Z_1^{(1)} = Z_1^{(2)} = 0$ and that their variances are equal to zero.

As before, we suppose that a measurement M is performed at $t = 10$ and that M is related to the nitrogen content as $M = aZ_{10}^{(1)} + b + \tau$ where a and b are two parameters and $\tau \sim N(0, \sigma_M^2)$ is a measurement error independent from $\varepsilon_{t-1}^{(1)}$ and $\varepsilon_{t-1}^{(2)}$.

(a) Determine the distribution for $\begin{pmatrix} Z_{10}^{(1)} \\ Z_{10}^{(2)} \end{pmatrix}$ before using the measurement. Give the expected value and the variance–covariance matrix for this random vector.

(b) What is the correlation between $Z_{10}^{(1)}$ and $Z_{10}^{(2)}$?

(c) What is the sensitivity of this correlation to the values of c, Q_1 and Q_2?

(d) What is the joint distribution of $\begin{pmatrix} Z_{10}^{(1)} \\ Z_{10}^{(2)} \\ M \end{pmatrix}$?

(e) Determine the distribution for $\begin{pmatrix} Z_{10}^{(1)} \\ Z_{10}^{(2)} \end{pmatrix}$ conditionally to M.

(f) What are the effects of Q_1, Q_2, and σ_M^2 on $E\begin{pmatrix} Z_{10}^{(1)} \\ Z_{10}^{(2)} \end{pmatrix}|M$ and $\mathrm{var}\begin{pmatrix} Z_{10}^{(1)} \\ Z_{10}^{(2)} \end{pmatrix}|M$?

(g) Compute the Kalman gain for the two state variables.

(h) Numerical application. Calculate $E\begin{pmatrix} Z_{10}^{(1)} \\ Z_{10}^{(2)} \end{pmatrix}|M$, $\mathrm{var}\begin{pmatrix} Z_{10}^{(1)} \\ Z_{10}^{(2)} \end{pmatrix}|M$, and the Kalman gains with $B_1^{(1)} = 0.001$, $B_2^{(1)} = 0.0015$, $B_3^{(1)} = 0.002$, $B_4^{(1)} = 0.0035$, $B_5^{(1)} = 0.002$, $B_6^{(1)} = 0.0025$, $B_7^{(1)} = 0.0025$, $B_8^{(1)} = 0.0018$, $B_9^{(1)} = 0.002$, $B_1^{(2)} = 10$, $B_2^{(2)} = 50$, $B_3^{(2)} = 65$, $B_4^{(2)} = 64$, $B_5^{(2)} = 70$, $B_6^{(2)} = 35$, $B_7^{(2)} = 38$, $B_8^{(2)} = 51$, $B_9^{(2)} = 25$, $Q_1 = 0.0005$, $Q_2 = 0.09$, $a = 1.2$, $b = 0$, $c = 0.001$, and $\sigma_M^2 = 0.001$.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 6

Representing and optimizing management decisions with crop models

J.E. Bergez, F. Garcia and D. Wallach

1. Introduction **173**
2. Which decisions, which variables? **174**
3. Representations of decisions and decision rules **174**
4. The uses of decision rules **175**
5. The optimization problem **178**
6. Simulation-based optimization **185**
7. Simulation-based control **195**
8. A comparison between optimization and control **201**
References **203**
Exercises **206**

1. Introduction

Among the explanatory variables in a crop model, the variables that represent management decisions have a special status because they are under the control of the farmer. In this chapter we consider two specific questions related to these variables. The first is how to characterize them. The simplest solution is to ascribe a fixed value to each management variable. However, it is also possible to express management variables as functions of other explanatory variables or of model state variables. Such functions are called decision rules and are discussed in Section 3. In Section 4, we consider different uses of crop models and discuss what representation of management decisions is adapted to each use.

The rest of the chapter is devoted to the second question, which is how to calculate optimal decisions. The elements of the optimization problem are presented in Section 5. The discussion concentrates on the objective function and on the optimization domain, i.e. the range of values within which an optimal solution will be sought. Sections 6 and 7 present methods for calculating optimal decisions. A major difficulty with crop models is the fact that future climate is unknown except as a probability distribution. The problem is thus a stochastic optimization problem. Section 6 presents simulation-based optimization methods and Section 7 control-based optimization methods. The two different approaches are compared in the final section.

It should be kept in mind that calculated optimal decisions suppose that the model correctly describes the effect of management variables on outputs. Obviously, the results need to be evaluated in the field.

2. Which decisions, which variables?

Decisions that are commonly taken into account in crop models include choice of variety, sowing date, sowing density, nitrogen fertilization dates and amounts, irrigation dates and amounts and harvest. This is the case with APSIM (Keating et al., 2003), CROPSYST (Stockle et al., 2003), EPIC (Cabelguenne et al., 1996), OZCOT (Hearn, 1994), PCYIELD (Welch et al., 2002), the CERES models (Pang, 1997) or the STICS crop model (Brisson et al., 1998, 2003). Management actions are often collected in a section called "management options" (Swaney et al., 1983; Hearn, 1994; Ghaffari et al., 2001).

Other decisions are rarely considered. The choice of crop species is seldom optimized directly using a dynamic crop model. More often optimization is based on linear programming which takes into account farm constraints as well as objectives. Decisions that are only rarely treated include dates of herbicide or pesticide applications, tillage type and residue management. Authors who have taken these decisions into account include van Evert and Campbell (1994), who propose a model for aphid population and aphid immigration management, Stockle et al. (2003a) who propose using the CROPSYST model for the specific problem of management of pesticide spraying and Batchelor et al. (1993), who treat the impact of pests on crops with the SOYGRO and PNUTGRO models.

If multiple cropping seasons are considered, then decisions include the decisions for the crop each year as well as decisions concerning catch crops. In general, optimization only concerns a subset of all the decisions that crop management entails.

The decision variables can be of several types. Crop species, variety and type of tillage are categorical variables. For example, species can take the values "corn", "wheat", "sunflower," etc. Dates are integer variables, and are usually given as the day number of the year. Amounts are real variables.

3. Representations of decisions and decision rules

"Management decisions" is a very general term and covers a large spectrum of representations. The simplest way to represent decisions is by fixed values. For example, the value of crop species might be "corn", the value of sowing day of year 123, the value of amount of nitrogen 50 kg ha^{-1}, etc.

A second approach is to represent decisions using a decision rule. A decision rule is a function which relates a decision variable to other explanatory variables or to state variables. We will refer to the variables in a decision rule as the indicator variables of the rule. A decision rule for sowing would be to sow on the first day after day of year 121 when soil moisture in the upper 10 cm of soil is below 70% of maximum water holding capacity. Here the indicator variables are the day of year and the relative water content. For the decision rule to be compatible with a model, indicator variables that are state variables (this would normally be the case for relative water content) must be calculated by the model. The remaining indicator variables must also be available. They could be explanatory variables for the model or variables that are supplied specifically for the decision rule.Finally, a decision rule normally involves some parameters. The threshold day 121 and the value of 70% are the parameters of the above decision rule.

Decision rules are also common in agronomy outside dynamic models. A very well known decision rule is the one for the amount of nitrogen fertilizer given by the "balance" method. The recommended dose of nitrogen is $d = (P_\mathrm{f} - P_\mathrm{i}) - (M_\mathrm{n} + R_\mathrm{i} - L - R_\mathrm{f})$, where $(P_\mathrm{f} - P_\mathrm{i})$ is the difference between the total nitrogen requirement of the crop and the amount of nitrogen absorbed up to the time of fertilization, M_n is total mineralization of soil nitrogen during the growing period, $R_\mathrm{i} - R_\mathrm{f}$ is the difference between initial and final soil mineral nitrogen and L is the amount of mineral nitrogen lost to deep drainage. The method includes algorithms for calculating each of the quantities in the equation in terms of climate, characteristics of the soil, crop, field and the management variables of the preceding crop. All these represent the indicator variables of the decision rule.

As this example shows, decision rules can be complex functions of the indicator variables. They can also be implicit functions, where the decision rule is simply an algorithm that allows one to calculate the value of the decision variable. At the other extreme, a fixed value is also a decision rule, a particularly simple one. For example, the rule "fertilizer amount = 50 kg ha^{-1}" is a rule with no indicator variables and a single parameter.

For the control methods of optimization presented below, decisions are represented as vectors which have a value for each day of the simulation. For example, the sowing decision is represented as a vector which assigns to each day either the value "no" (don't sow) or "yes" (sow). In this particular case there can be only a single "yes" value, and the corresponding day is the sowing day. If the decision concerns both irrigation dates and amounts, there will be a vector which gives the amount of irrigation each day, with zeroes for days with no irrigation.

4. The uses of decision rules

The number of possible decision rules for a particular decision is unlimited. The rules may involve different indicator variables, different mathematical forms or different parameter values. The choice between different decision rules or more fundamentally the criterion for choosing a particular decision rule will depend on the objectives of the study.

4.1. Reproducing past decisions

A model is often used to reproduce past situations, and of course one requires the management decisions as inputs for those situations. In this case, the management decisions will have the form of fixed values, the values that were actually applied.

One use of simulating past situations is to furnish a diagnosis of how and why yield differed from potential yield. This is termed "Yield gap analysis" (see Meynard and David, 1992; Kropff et al., 2001; Matthews, 2002). The purpose of comparing the model with past data might be to identify limiting factors. As stated by Kropff et al. (2001): "Crop simulation models offer a way of estimating what the potential yield of a crop is and a step-wise analysis of the various inputs can help identify the limiting factors." Yield gap analysis can be applied on larger scales, such as a whole region, not only a field (Doré et al., 1997; Affholder and Scopel, 2001).

Other uses of simulating past situations include model evaluation, where calculated and observed values are compared, and parameter estimation, where the parameters are adjusted to minimize the difference between calculated and observed values.

Another use of past decisions is as a baseline for examining alternative decisions. One can pose questions like: "what would have happened if I had sown 3 days earlier?" (see Ghaffari et al. (2001)). This is an example of virtual experimentation (Matthews, 2002).

4.2. Predicting management decisions

Suppose that the purpose of the modeling exercise is to make predictions, for example, of yield or of water consumption for irrigation in a region. Then one needs to predict what management decisions will actually be taken in each field. The criterion for judging a decision rule in this case is how closely it imitates farmer behavior. One rather simple approach here is to assume that farmers follow recommendations for good practices and to base the decision rules on those recommendations. For example, there may be recommendations for fertilizer amount or for planting date which take the form of decision rules. Then one could use those decision rules for prediction. This approach was used by Leenhardt et al. (2004) to predict agricultural water consumption in a 12 000 km^2 area in southwestern France.

Alternatively, one could try to model farmer behavior more realistically. For example, irrigation decisions might depend on the available irrigation equipment on the farm (Maton et al., 2005). The availability of equipment is then an additional indicator variable.

A particular aspect of imitating farmer behavior is to avoid decisions that would not be taken in practice. This is particularly important if the model does not simulate correctly the results of such decisions. Thus allowing them might lead to very substantial errors. An example would be the decision to enter the field for soil tillage. A farmer would avoid doing so in practice if the soil was wet since that could deteriorate soil structure, hamper germination, increase the risk of water logging and finally reduce yield. However, most models do not simulate soil structure and therefore would not penalize tillage under wet conditions. The problem could be avoided using a decision rule for tillage that has soil moisture as an indicator variable, and that forbids tillage under wet conditions. Examples are given by Swaney et al. (1983) and Hearn (1994).

A simple and common way to restrict sowing to dry conditions is to wait until a threshold day is passed and then to sow as soon as cumulative rainfall in the previous n days is less than σ. The rainfall part of the decision rule would be:

$$\text{IF} \sum_{k=1}^{n} P_{t-k} < \sigma \text{ THEN } \text{sowing\_date} = t \tag{1}$$

where P_{t-k} is rainfall on day $t-k$ and t is the current day in the simulation. This rule would be evaluated each day until the condition were met or until the latest acceptable sowing date were reached.

4.3. Scenario testing

Another common use of models is for scenario testing. The choice of decision rules in this case will depend on the exact objective of the modeling exercise. In some cases it may be reasonable to use decision rules that imitate current farmer behavior. For example, if the objective is to test the effect of warming temperatures on yield, a first study may use decision rules for irrigation based on current practices, since they automatically include adaptation of practices to climate. In other cases the scenario may specifically concern new decision rules, which would then replace the decision rules that imitate current farmer behavior. For example, one might want to test the consequences of reducing the amount of nitrogen applied to wheat at the second application and adding a third application with an amount based on a measurement of plant nitrogen status. Then it would be necessary to change the decision rule for the amount at the second application, compared to current practices, and add a decision rule for the amount of nitrogen at the third application.

4.4. Optimizing management decisions

The use of a model to calculate optimum decisions is the main topic of the following sections. We will treat the question of decision rules for optimization there in. However, optimization studies rarely concern all the management decisions for a crop. The problem then arises as how to treat those decisions that are not optimized. The choice may be crucial for the optimization study. For example, suppose the objective is to optimize the dates and amounts of irrigation of corn. Suppose further that the amount of nitrogen fertilizer is not optimized but is given as a fixed value or a fixed decision rule. If the fertilization rule leads to nitrogen stress, this may favor irrigation strategies which also limit production. It is important to be aware that the optimization results may depend on the way one treats the decisions that are not optimized.

4.5. Decision strategies

We use the term decision strategy, noted π, to refer to the collection of all the decision rules. In many cases, the strategy is simply the collection of the decision rules for each individual decision. In more complex cases, the strategy may include relationships between the different decisions. For example, if nitrogen is applied with the irrigation water, then the time of irrigation and the time of fertilization must coincide. As shown by studies done on the subject of formulating decision processes in agriculture (Aubry et al., 1998; Papy, 2000; Cros et al., 2001, etc.), it is necessary not only to address the adaptive character of the decision-making process using decision rules, but also to include scheduling characteristics. A decision rule with a simple structure of the type "IF (condition involving indicator variables) THEN (conclusion that sets values of decision variables)" does not always suffice to model the scheduling of different management interventions. A more general structure like a "REPEAT … WHILE" loop could be used to represent the repetitive character of some actions, which would be repeated until an event interrupted the sequence. This type of structure is used in the Moderato model (Bergez et al., 2001a).

Sequence or loop control structures are easy to write and are well adapted to simple crop systems (one field, homogeneous, elementary management intervention). However, if the crop system is more complex (several fields, choice of equipment, etc.), it may become necessary to allow and manage actions in parallel. It is then necessary to describe the duration of each action and to identify those that must be synchronous. It is also necessary to describe the management of shared resources (manpower, machinery, etc.). We then need priority rules to handle simultaneous actions that demand the same resources. It may also be necessary to have adaptation rules to model the evolution of the sequence of actions during the growing season. Examples of decision structures are shown in the Otelo (Papy et al., 1988; Attonaty et al., 1994), Conserto (Jeannequin et al., 2003) and Sepatou (Cros et al., 2001) models.

5. The optimization problem

As presented in Chapter 1, the dynamic equations of a crop model are of the form

$$U(t+1) - U(t) = g(U(t), X(t); \theta), \quad t = 1, \dots, T \tag{2}$$

where $U(t)$ represents the state variables of the system, $X(t)$ the vector of explanatory variables and θ the vector of model parameters. In this chapter we partition the vector $X(t)$ into the variables that represent decisions to be optimized, noted as $D(t)$, and the remaining explanatory variables noted as $C(t)$. For example, $D(t)$ could include planting date and density, which are to be optimized, and $C(t)$ could include the climate variables, soil characteristics, initial conditions and all remaining decision variables such as crop species and variety, dates and amounts of nitrogen applications, etc. Note that here we need daily values of the decision variables. It is natural then to use the control representation of the decision variables. That is, the couple $\left[D_{\text{sow\_date}}(t), D_{\text{density}}(t)\right]$ has the value [yes, density] on the sowing day and [no, 0] every other day.

The crop model written as a response model (Chapter 1) is now

$$\hat{Y}(T) = f(D, C; \theta), \quad D = \{D(t),\ t = 1, \dots, T\}, \quad C = \{C(t),\ t = 1, \dots, T\} \tag{3}$$

Given C, one could evaluate decision rules to obtain D. Thus for this formulation, the decision variables can be expressed as decision rules.

We can now introduce the essential elements of the optimization problem. These are:

(a) the amount of information available at decision time about the variables in C. This concerns the climate in particular. The simple case is when one assumes that climate for the entire season is known. This might be of interest in an a *posteriori* study, to determine the best decisions that could have been taken. When the problem is to make recommendations for decisions, it is generally assumed that the decision maker only has access to a probability distribution for future climate.

(b) the criterion that the optimal decisions maximize (or minimize). This criterion, called the objective function (it is a function of the management decisions), defines what exactly is meant by "optimal".

(c) the optimization domain. This defines the range of values within which we seek the optimal decisions.

Now, we consider the objective function and the optimization domain in more detail. The methods and algorithms for optimizing crop management decisions are also presented.

5.1. The objective function and utility

The evaluation of a crop management strategy can be based on different criteria, which may involve economic results, environmental impact, organizational considerations, etc. In order to use a simulation model to evaluate strategies, the model must calculate all the necessary results. Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be the vector of all results of interest simulated by the model. These outputs can have numeric (discrete or continuous) or symbolic values.

If we note π as a management strategy generating the $D(t)$ decisions, the Y_i values are the output of a response model that can be written as

$$Y_i = f(\pi, C; \theta), \quad C = \{C(t), t = 1, \ldots, T\} \tag{4}$$

In the simplest case where the explanatory variables C are assumed to be known, and where there is one unique numeric variable of interest Y, the problem of defining the best strategy is simple; it is the strategy with the largest value of Y, knowing C. In general, however, C is only known as a probability distribution (climate, soil, …) and several Y_i have to be considered. Then Y is a random vector. In this case the notion of expected utility is the traditional criterion for deciding which of the two strategies is to be preferred.

5.1.1. Maximizing the expected utility

In order to manipulate scalar quantities, which are easy to compare and optimize, we define a quantity called the utility and noted as U, which is a scalar function of Y

$$U = U(Y_1, \ldots, Y_n) \tag{5}$$

Thus even if Y is a vector, the utility which is derived from it is a scalar. This utility is a summary of the satisfaction gained by the farmer at the end of the crop, considering all the outputs represented by the Y_i values.

When the different C values are defined by a probability distribution $dP(C)$, the scalar $U(Y)$ is also a random variable with a probability distribution $dP_\pi(U)$, which is a function of the management strategy π. The expected utility criterion then consists in preferring management strategy π_1 to management strategy π_2 if and only if the expectation of the function $U(Y)$ for the dP_{π_1} distribution is greater than its expectation for the dP_{π_2} distribution:

$$\pi_1 > \pi_2 \Leftrightarrow \mathrm{E}\left[U(Y)|dP_{\pi_1}\right] > \mathrm{E}\left[U(Y)|dP_{\pi_2}\right] \tag{6}$$

where

$$\mathrm{E}[U(Y)|dP_\pi] = \int_U U dP_\pi(U) = \int_C U(f(\pi, C; \theta)) dP(C) \tag{7}$$

Example 1

Consider management of a rapeseed crop evaluated using two criteria: Y_1 is yield and Y_2 is amount of applied nitrogen. We assume that for the two strategies π_1 and π_2, $Y = (Y_1, Y_2)$ has a normal distribution $N(m_\pi, \Sigma_\pi)$. We assume that the utility function is $U(Y) = Y_1 - \alpha Y_2$. Then the random variable U also has a normal distribution $N(m, \Sigma)$ and using the expected utility criterion, we choose the strategy with the greatest mean value m (see Fig. 1).

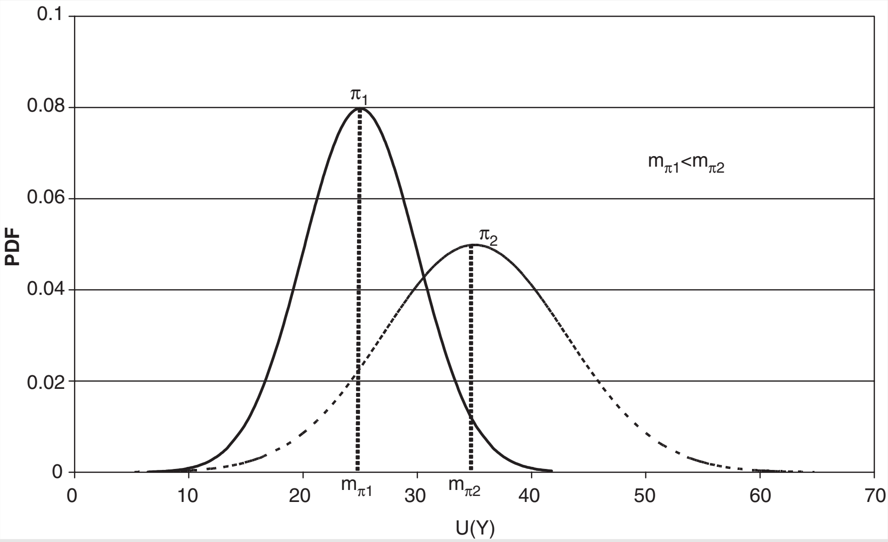


Figure 1. Normal distribution of the utility U for two strategies π_1 and π_2.

5.1.2. Estimating the expected utility criterion

To estimate the expected utility of a management strategy π, a Monte-Carlo approach is traditionally used, where Eq. (6) is replaced by the mean value of the utilities U(Y) calculated by simulation for a large number N of possible values of the uncontrollable variables C, observed or randomly generated. We write this estimation as

$$E[U(Y)|dP_\pi] \approx 1/N \sum_i U(f(\pi, C_i;\theta)). \tag{8}$$

Note that the chosen criterion also depends on the parameters θ of the crop model, which in general are not perfectly known but only estimated. When it is possible to define the distribution of the parameter estimators, we can extend the definition of the expected utility by calculating also the expectation over θ

$$E[U(Y)\,|\,dP_\pi] = \int_\theta \int_C U(f(\pi, C;\theta))dP(C)dP(\theta). \tag{9}$$

This criterion is again estimated using the Monte-Carlo method:

$$\mathrm{E}[U(Y)|dP_{\pi}] \approx 1/M 1/N \sum_{j} \sum_{i} \mathrm{U}\left(f(\pi, C_i; \theta_j)\right) \tag{10}$$

where M values θ_j are drawn randomly using the parameter distribution.

5.1.3. Modeling risks and other criteria

Decision makers are concerned not only with the average value of the objective function but also by its variability. If, for the same average result, one prefers stable values to varying values, one is risk averse. In the other case one is risk-seeking. One important advantage of the notion of utility is that it can be used to take into account risk aversion or inclination toward risk of the manager.

Example 2

Suppose that Y is net profit and that there are just two possible future climates C_1 and C_2. Consider two management strategies π_1 and π_2. Suppose that for the first management strategy $Y(\pi_1, C_1) = 2000€$ and $Y(\pi_1, C_2) = 1000€$, while for the second management strategy $Y(\pi_2, C_1) = 1500€$ and $Y(\pi_2, C_2) = 1500€$. If we simply use the average of Y as our criterion, then the two strategies are equivalent. For both the average net profit is 1500€. Suppose however that we want to imitate a manager who is risk averse. For this decision maker, the advantage of making 2000€ instead of 1500€ in one year is not as important as the disadvantage of making 1000€ instead of 1500€. The utilities might then be $U(\pi_1, C_1) = 1800$ and $U(\pi_1, C_2) = 1000$ for the first management strategy and $U(\pi_2, C_1) = 1500$ and $U(\pi_2, C_2) = 1500$ for the second. Now the expected utilities are 1400 and 1500 respectively, and strategy 2 is preferred.

More generally, risk aversion corresponds to concave utility functions like $U(Y) = \ln Y$, for which we have $E[U(Y)] < U(E[Y])$. Conversely, inclination toward risk corresponds to convex utility functions.

Example 1 (cont'd)

Let us consider again the previous rapeseed crop model with now $\mathrm{U(Y)} = (\mathrm{Y}_1 - \alpha \mathrm{Y}_2)^{1-\beta} / (1-\beta)$, $\beta > 0$. The utility function is concave and the model imitates a risk averse decision maker. The distributions of U(Y) for π_1 and π_2 are modified and now π_2 can be preferred to π_1 depending on the value assigned to β.

Note that other criteria than the utility expectation can be used, for instance, if it is difficult to determine *a priori* distribution for C, or to define a utility function which combines correctly the different outputs $Y_1, \ldots, Y_n$.

A common approach is the maximin approach, which does not require knowledge of the probability distribution. Here

$$\pi_1 > \pi_2 \Leftrightarrow \min_C \mathrm{U}(f(\pi_1, C; \theta)) > \min_C \mathrm{U}(f(\pi_2, C; \theta)) \tag{11}$$

According to this approach, one prefers strategies with greater utility in the worst case. This method is in fact not very discriminating.

When a unique utility function cannot be defined in a satisfying manner, but nonetheless it is possible to define a utility function U_i for each variable of interest $Y_i, i = 1, \ldots, n$, a multiple criteria comparison technique may be used. Then, criteria $E[U(Y_1)]$, $E[U(Y_2)], \ldots, E[U(Y_n)]$ are estimated for each management strategy π, and multi-criteria methods can be employed (Roy, 1996). See Chapter 20 for an example.

5.2. The optimization domain

In this section, we consider the definition of the range of possible decision strategies – the "optimization domain" $\Pi = \{\pi\}$ within which the optimal solution will be sought. We also discuss the criteria for choosing between different possible optimization domains. One of these criteria is how does the optimization domain compare to a domain which imposes no constraints on the optimal solutions. We begin by defining this unconstrained domain.

5.2.1. The unconstrained optimal decisions

For each situation, that is for each decision date and each set of values or distribution functions of C, there is some management strategy which maximizes the chosen utility function. We will call this the unconstrained optimal strategy.

What optimization domain would we use if we wanted to be sure that it includes the unconstrained optimal solution? It would be a domain that, for each C, includes all possible management strategies. We call this the unconstrained domain. The important point here is that in the unconstrained domain, the management strategy could be different for each C.

Consider for example, the simple case of optimizing the amount of nitrogen to apply to a wheat crop at heading. All other decisions are assumed fixed, so the optimization concerns just a single variable. We assume that at the time the decision is made (at heading), climate up to that time is known and future climate is unknown except as a probability distribution. The unconstrained optimization domain is, for each year, an amount of nitrogen in the range 0 to some very large value. The domain specifically includes the possibility that the optimal amount can be different every year.

The unconstrained optimization domain has the important advantage that it includes the unconstrained optimal strategy. Nevertheless, in practice it may be necessary or even preferable to accept a constrained domain and a sub-optimal management strategy.

First of all, there is the difficulty of calculating the unconstrained optimal strategy. We can define the calculations to be carried out (see the section on simulation-based control) but actually doing them may be very difficult. The difficulty is related both to the complexity of the model (number of state variables) and to the complexity of the decision problem (number of decisions). Second, even if the calculations are possible, there is a major practical problem. The optimal management strategy is based on all the information available at decision time (in particular climate up to that day). To take advantage of this information one has two choices. Either one does the calculations for all possible climates, so the decision maker can look up the results for his particular climate variables, or one does the calculations each day using the climate variables up to that day. The first solution involves in general an enormous number of optimization calculations. (For a decision on day 5 with 4 daily climate variables and just 5 possible values for each, the number of possible climate sequences is $5^{20} \approx 10^{14}$). The second possibility, of redoing the calculations each day, is not adapted to an extension agent making recommendations for a large number of fields. The third disadvantage is that the procedure of calculating an unconstrained optimal amount each year is not based on an understanding of the factors which influence the optimal amount. One must simply have confidence in the model. One may prefer a procedure which is more clearly related to our knowledge of agricultural systems even if the resulting amounts are slightly sub-optimal. The final disadvantage, related to the previous one, concerns the case where the optimization calculation is aimed at a better understanding of the determinants of optimal management rather than at making recommendations. An optimization calculation that outputs a different strategy every year may not provide much insight.

5.2.2. Decision rules

The use of fairly simple decision rules provides a different way of specifying the optimization domain. For example, suppose the decision rule is that the optimal amount of nitrogen/ha Q_N is θ_1 times average yield for this field in the absence of nitrogen stress $\bar{Y}$ (we assume that this is an available indicator variable) minus θ_2 times mineral nitrogen in the soil at sowing, N_{soil} (also assumed to be available) with θ_1 in the range [0–1] and θ_2 in the range [0–2]:

$$Q_N = \theta_1 \bar{Y} - \theta_2 N_{soil} \tag{12}$$

For fixed values of the parameters θ_1 and θ_2 this is a decision rule as we have defined it, since the value of the decision variable is a function of two indicator variables, namely average yield and mineral N at sowing. The optimization domain Π is the domain defined by the ranges of θ_1 and θ_2. The use of this decision rule may lead to different amounts in different years, but it is not as flexible as the unconstrained domain. Average yield is a property of the field and so may be considered constant from year to year. The decision rule then automatically implies that two years with identical mineral N at sowing will have the same optimal amount of nitrogen, whereas this is not necessarily true for the unconstrained domain. Thus this decision domain is not as general as the unrestricted domain and therefore may lead to sub-optimal decisions.

On the other hand, an optimization domain based on a decision rule does not suffer from the disadvantages that we listed for the unconstrained domain. First, the optimization calculation is in general much easier. In our example of nitrogen fertilization, the optimization calculation involves finding the optimal values of the two parameters. This can be done using one of the simulation-based optimization techniques described below.

This is in general much easier than using a simulation-based control method. Second, the optimization calculation can now be done at leisure. Climate up to decision time will be taken into account automatically in applying the decision rule. Third, a decision rule is, in general based on agronomic knowledge and is easily understood. It will then probably be more readily accepted than an unconstrained optimal amount. Finally, in general it will be much easier to understand and analyze optimal decision rules than the unconstrained management strategies.

We began by assuming that the optimization concerns a single field. In many cases, however, one seeks optimal decision rules for a range of soil types, climates, initial conditions, etc. One could treat each case separately, but this requires multiple calculations and may provide little insight. A different possibility would be to use decision rules that apply to the full range of conditions considered. Normally the decision rules would then include indicator variables related to soil type, average climate, etc. For example, the decision rule for fertilizer amount at the first application might include soil organic N content as an indicator variable, since mineralization rate depends on organic N content.

5.2.3. The criteria for judging decision rules for optimization

There are two major criteria for judging decision rules and their accompanying optimization domains in optimization studies. The first is how closely does the optimization domain allow one to get to the unconstrained optimal solutions. One wants to avoid optimization domains that are very restricted and that only include solutions whose performance is far below that of the unconstrained optimal solutions. The second criterion is that the decision rule must be acceptable to the decision makers. A corollary is that the rule must be applicable in practice.

We identify three different characteristics of decision rules, and discuss how each of these affects the two criteria. First of all, a decision rule is characterized by the indicator variables involved. Obviously, adding additional indicator variables makes the rule more flexible and thus allows for optimal solutions closer to the unconstrained optimal solutions. For example, we could expand an irrigation rule based on “starting a new irrigation cycle every 10 days” to include a delay in case soil moisture is above some threshold. Soil water would be an additional indicator variable and its inclusion would increase the range of possible decision strategies. Consider now the second criterion of acceptability to and applicability by decision makers. Suppose that the person who implements the decision rule does not have the model (this is the case in general). The state variables in the decision rule must then be available from measurements, since they are not calculated. A farmer without access to measurements of soil moisture could not use the more complex rule. The availability of measured values may be a major constraint on the state variables in the decision rule.

The mathematical form of the decision rule, for given indicator variables, is a second important characteristic, which can strongly affect the range of possible decision strategies. For example, a more complex version of the above decision rule involving soil water would be to delay irrigation if soil water is above threshold 1 up to day d, then after that day to use a threshold 2. There is only one indicator variable, soil moisture, but the rule is more flexible than having a single threshold.

The third characteristic of decision rules for optimization is the range of values for each parameter in the rule. In most cases one might simply opt for a range which allows all reasonable values. However, it may sometimes be worthwhile to restrict the range of possible values. One reason would be to improve the performance of the optimization algorithm. If one is sure that the optimal value is in some restricted range, it would be worthwhile to limit the optimization domain to that range.

6. Simulation-based optimization

6.1. The stochastic optimization problem

We address in this section the general problem of designing optimal crop management strategies for a large variety of contexts (climates, soils). As we have seen in the previous sections, it is common to define such solutions that maximize the expected value of an objective function $J = U(Y) = U(f(\pi, C))$ for the random context C:

$$J^* = \max E\left[J(\pi, C)\right] \quad \text{with} \quad \pi \in \Pi. \tag{13}$$

Since Π, the domain π of candidate strategies is potentially infinitely large, a realistic formulation of this optimization problem consists of searching for the best values for the parameters of some pre-defined strategy. In this case a strategy π is completely characterized by the vector of strategy parameters $\theta = (\theta^1, \ldots, \theta^p)$ and $\pi = \{\pi(\theta)/\theta \in \Theta\}$, where Θ is the value domain of θ. This family of candidate management strategies is typically built by some experts of the domain, and the strategy parameters generally represent thresholds, dates, quantities, etc.

Example 3

Consider the MODERATO simulation model of irrigation management for maize crops described in Chapter 19. In this model, irrigation strategies π are described by parameterized decision rules, for example: "The main irrigation period starts from day θ_1 as soon as the soil water deficit reaches θ_2. An amount of water θ_3 is applied. Once an irrigation cycle ends, a new cycle starts when the soil water deficit reaches θ_4. An amount θ_5 is applied. For the irrigation cycle following day θ_6, if the soil water deficit is greater than θ_7 before this irrigation cycle starts, a last irrigation cycle is performed; otherwise the irrigation period ends. An amount θ_8 is applied". The domain of $\theta = (\theta_1, \ldots, \theta_8)$ defines a set of strategies. Within this domain, we can search for irrigation strategies that maximize expected net profit.

In that framework, the objective function J to optimize is a function of θ and C, and optimizing a strategy thus leads to the stochastic optimization problem:

$$J^* = \max E\left[J(\theta, C)\right] \quad \text{with} \quad \theta \in \Theta, \tag{14}$$

which is one of the most difficult problems of mathematical programing. A particular instance of this optimization problem is obtained when a specific context C is considered. In that case the problem (14) becomes the more classical deterministic optimization problem:

$$J^* = \max J(\theta, C) \quad \text{with} \quad \theta \in \Theta, \tag{15}$$

This corresponds to the first use of crop models for management mentioned above, where we want to determine a sequence of actions maximizing an objective function J, in order to answer the question: " knowing the context (climate, soil, etc.), what would have been the best management?"

Example 4

Jallas et al. (1998) are interested in determining the best *a posteriori* irrigation decisions when weather data of the year are available. Decisions are modelled as a fixed length set of (date, dose) values.

6.2. The simulation-based optimization approach

The simulation-based optimization approach consists in solving (14) by means of simulation of the objective function J (Azadivar, 1999; Swisher et al., 2000; Fu, 2001), without any additional information concerning the structure of the function J or the probability distribution of C.

An intuitive approach for finding an approximate solution of Eq. (14) is to solve the associated deterministic optimization problem

$$J^* = \max J^N(\theta) \quad \text{with} \quad \theta \in \Theta, \tag{16}$$

where J^N is an estimate of the criterion $E[J(\theta, C)]$ obtained by averaging the objective function $J(\theta, C)$ over a large number of values C_i:

$$J^N(\theta) = \frac{1}{N} \sum_{i=1}^{N} J(\theta, C_i) \tag{17}$$

When the C_i are fixed (i.e. at evaluation of J^N, the same C_i are used), the objective function J^N becomes a deterministic function of θ, and traditional optimization algorithms can be used. This technique, called "sample path optimization", or "sample average approximation" thus converts a stochastic problem into a deterministic one. It was originally developed for solving continuous parameter simulation optimization problems (Gurkan et al., 1994), and has been recently studied in a variety of contexts (Homem-de-Mello, 2003). When, for instance, the context variables C represent weather series,

$E[J(\theta, C)]$ can thus be estimated by simulating $J(\theta, C_i)$ on N fixed historical series and averaging the result. $N = 30$ is generally considered as sufficient for accurately estimating the expected value of classical objective functions J such as net margin for instance, but in practice smaller values of N are often chosen. However, as illustrated in Exercise 1, depending on the desired precision, a good estimate of $E[J(\theta, C)]$ may require a large number N of simulation runs.

When the size of the optimization domain is very large or when simulation runs are slow, it can be more efficient to avoid this uniform and systematic estimation of the expected criterion and to allocate a computational effort to a candidate θ that depends on the current estimate of the expected value. This adaptive allocation can be facilitated by the use of a stochastic generator of samples C_i like random weather generators. In that case, J^N becomes a random variable. This is the main principle of stochastic simulation-based optimization methods.

6.3. Optimization problems with discrete domains

(a) Problem statement

For discrete domains, (14) can be written as:

$$J^* = \max_{\mathrm{i}=1,\ldots,K} E\left[J(\theta_i, C)\right], \tag{18}$$

where $\Theta = \{\theta_1, \ldots, \theta_K\}$ is the discrete optimization domain of size K.

While considering discrete domains, two cases can occur. Either you can enumerate all the elements of the domain, or you need to make a random search among these candidates. Things differ depending on the size of K.

(b) Complete enumeration methods

For small discrete domains where a complete enumeration is possible, the only methodological question lies on the choice of the number N_i of simulation runs used for estimating the objective function for the parameter value $\theta_i, i = 1, \ldots, K$. You can make a uniform allocation $N_i = N_{\mathrm{tot}}/K$ where N_{tot} is the total number of simulations you can do. When simulation runs are fast, as is the case for most of the crop models, N_{tot} can be very large (up to 10^6 in few hours) and this uniform method is generally sufficient for determining the best solution.

When uniform allocation poses a problem (K is too large or N_{tot} needs to be kept to a fairly small value), specific stochastic methods have been designed for selecting the optimal parameter value over a small finite set with a minimal number of simulation runs (Ho et al., 1992; Goldsman and Nelson, 1994; Hsu, 1996). The main idea of these approaches is to define N_i as a function of the current mean and variance estimates of $J(\theta_i, C)$. For instance, in the OCBA method (Chen et al., 2000), the N_i values are defined by:

$$\frac{N_b}{N_i} = \sigma_b \sqrt{\sum_{j=1, j\neq b}^{k} \frac{1}{\sigma_j^2} \rho_{ij}^2}, \quad i \neq b$$

Example 5

One wants to optimize the choice of variety and sowing date for winter wheat crop production. 10 varieties v are considered. Possible sowing dates d are between 16/09 and 01/11, which corresponds to 45 possible days. The optimization domain is thus $\Theta = \{(v_i, d_j), i = 1 \ldots 10, j = 1 \ldots 45\}$ that contains 450 candidates (see Fig. 2).

	Variety 1	Variety 2	Variety 3	Variety 4	Variety 5	Variety 6	Variety 7	Variety 8	Variety 9	Variety 10
15-sept										
16-sept										
17-sept										
18-sept										
19-sept										
20-sept										
...										
27-oct										
28-oct										
29-oct										
30-oct										
31-oct										
01-nov										

Figure 2. Optimization domain for the wheat crop production problem.

with

$$\rho_{ij} = \left(\frac{\sigma_j / \Delta_j}{\sigma_i / \Delta_i} \right)^2, \quad i, j = 1, \ldots, K, \quad i, j \neq b, \tag{19}$$

$$\Delta_i = J_b - J_i,$$

where θ_b is the best current value of θ and σ_i^2 is the estimated variance of $J(\theta_i, C)$. This method has been designed for optimizing the probability of correct selection of the best candidate given a total number $N_{\text{tot}} = N_1 + \cdots + N_K$ of simulation runs. OCBA, like similar approaches, is interesting when the computation of $J^N(\theta)$ for large N is very expensive, or when one needs to solve the optimization problem very rapidly.

(c) Local search methods

When the optimization domain is very large, a complete search of the domain is no longer possible. In that case, we have to consider heuristic search methods that look for an approximate optimal solution of (18). Two families of methods can be identified: local search and branching methods.

Local search methods move iteratively and randomly from a current point of the domain Θ to another in its neighbourhood, with probability that depends on the respective values J of these different points. In a deterministic framework, where C is known in advance (Eq. 15), $J(\theta, C)$ is not a random variable and some random transitions have to be put explicitly in the algorithm. Such recent local search methods are genetic algorithms (Michaelewicz and Schoenauer, 2001), simulated annealing (Fleischer, 1995) or tabu search (Glover and Laguna, 1997). We develop here the case of the simulated annealing algorithm, that starts with a random exploration of the domain, and then performs hill-climbing with an increasing probability:

Simulated annealing (for deterministic problems)

```
(1)  Initialize $n = 0$, $\theta_0 \in \Theta$
(2)  While simulation effort is not exhausted repeat 2a–2e
(2a) Construct the neighbourhood $V(\theta_n) \subset \Theta$
(2b) $\theta'$ is randomly sampled from $V(\theta_n)$
(2c) Calculate $J(\theta_n, C)$ and $J(\theta', C)$
(2d) if $J(\theta', C) \geq J(\theta_n, C)$
     then $\theta_{n+1} = \theta'$
     else draw randomly $\varepsilon_n$ in [0,1]
        If $\varepsilon_n < \exp\left(\frac{J(\theta_n, C) - J(\theta', C)}{T_n}\right)$
            Then $\theta_{n+1} = \theta'$
            Else $\theta_{n+1} = \theta_n$
(2e) $n \leftarrow n+1$
```

In this algorithm, T_n is a positive factor called the temperature that decreases slowly to 0, for example, using $T_n = 1/\log n$. The neighbourhood $V(\theta_n)$ is a set of points "close to θ_n" in Θ. $V(\theta_n)$ is typically defined by making small changes in the components of θ_n components, or by introducing a distance on Θ.

These local search methods can be guaranteed to converge to the set of global optimal solutions of the deterministic optimization problem. They have been applied to agricultural simulation models in recent years (Mayer et al., 1998a). They all have performed well for the problem of optimizing management decisions on existing historical climatic series (e.g. Li and Yost, 2000; Mayer et al., 1996; 1998b, 2001; Parsons, 1998; Reddy et al., 1995), and appear to be quite robust.

For solving the general stochastic optimization problem defined by Eq. (18) where the context variable C is assumed to be unknown, two approaches are possible. One is to use a deterministic estimate J^N Eq. (17) with a fixed number N of *a priori* samples like historical weather series or soils for instance. In that case, all the previous search methods developed for deterministic problems can be used, replacing J by J^N. Or one can use some local search methods like random search or the stochastic ruler algorithm (Andradottir, 1998), that have been specifically developed for the case where J^N is stochastic. We present here only the random search algorithm.

Example 5 (cont'd)

A neighborhood of size 4 can be defined for the elements of the optimization domain $\Theta = \{(v_i, d_j), i = 1\ldots 10, j = 1\ldots 45\}$:

$$V(\theta = (v_i, d_j)) = \{(v_{i-1}, d_j), (v_{i+1}, d_j), (v_i, d_{j-1}), (v_i, d_{j+1})\} \quad \text{(see Fig. 3).}$$

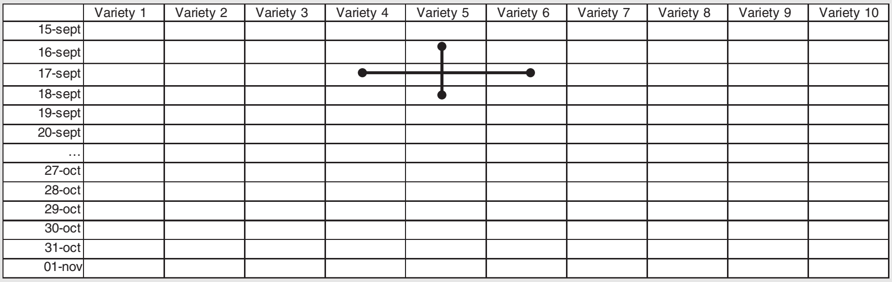


Figure 3. Neighborhood of size 4 for the wheat crop production problem.

Random search algorithm (for stochastic problems)

```
(1)  Initialize $n = 0$, $\theta_0 \in \Theta$
(2)  Repeat 2a–2d until the stopping criterion is satisfied
(2a) Construct the neighborhood $V(\theta_n) \subset \Theta$
(2b) $\theta'$ is randomly sampled from $V(\theta_n)$
(2c) Calculate $J^N(\theta_n)$ and $J^N(\theta')$
(2c) $\theta_{n+1} = \theta'$ if $J^N(\theta') > J^N(\theta_n)$, $\theta_{n+1} = \theta_n$ otherwise.
(2d) $n \leftarrow n+1$
     Return $\theta_n$.
```

Possible stopping criteria are to stop when $\theta_{n+1} = \theta_n$ on several consecutive n, or when the perceived progress between θ_n and θ_{n+1} is too small: $\left|J^N(\theta_{n+1}) - J^N(\theta_n)\right| \leq \varepsilon$.

In this algorithm, J^N is calculated on N randomly selected samples of C through the use of a random generator. One can see that random search is a kind of simulated annealing, where the transition probability is not explicitly represented but is replaced by the noise associated with the estimate $J^N(\theta)$ of $E[J(\theta, C)]$.

(d) Branching methods

Branching methods belong to the family of global optimization algorithms that are designed to look for local and global solutions of the optimization problem (18). Branching methods divide recursively the whole domain Θ into subdomains of smaller size. Once again, traditional branching methods developed in a deterministic framework like the branch-and-bound algorithm can be used for solving (18) with approximation (17). However, stochastic versions of branching methods for discrete global optimization problems have been recently proposed. Norkin et al. (1998) have developed a stochastic version of the branch-and-bound method. Shi and Ólafsson (1997) have proposed the nested partition (NP) method that hierarchically partitions and evaluates by randomly sampling the search space.

6.4. Optimization problems with continuous domains

(a) Problem statement

When $\theta \in IR^p$ is a multi-dimensional continuous variable, the optimization problem (20) becomes a continuous stochastic optimization problem:

$$J^* = \max E[J(\theta, C)] \quad \text{with} \quad \theta \in \Theta \subset IR^p. \tag{20}$$

Example 6

We want to optimize the starting date of an irrigation period with the MODERATO simulation model of corn crop management. The objective function to be maximized is the expected value of the net revenue R after harvest defined as the gross revenue minus irrigation costs. The choice of the starting date is modeled with a simple decision rule that starts irrigation when the accumulated thermal units above 6°C since sowing becomes greater than a threshold θ. The range of θ is fixed to [500°, 1000°]. Once irrigation begins, we assume that a fixed irrigation strategy is applied for the subsequent irrigation decisions. This leads to a one-dimensional optimization problem:

$$R^* = \max_{\theta \in [500^\circ, 1000^\circ]} E\,[R\,(\theta, C)].$$

(b) Local methods

As for discrete domains, local methods consist in exploring Θ by moving from θ_n to its neighbor θ_{n+1}. The gradient search method is a hill-climbing method that consists in moving from one point θ_n to the next θ_{n+1} by following the gradient of a deterministic objective function $J(\theta)$. Gradient search is a very popular approach for continuous domain $\Theta \subset IR^p$. However, solving with simulation multidimensional and nonlinear optimization problems is not an easy task, even for deterministic problems. Indeed, we can only simulate the objective function J but not its derivative functions which have to be approximated numerically, and the function J can be quite chaotic or conversely very flat as a function of the inputs θ.

Coupled with approximation (17), the gradient search method can be implemented with the following algorithm, where $\Delta\theta^i$ is the gradient step size for parameter θ^i:

Gradient search algorithm

(1) Initialize $\theta_0 \in \Theta$

(2) Repeat 2a–2b until the stopping criterion is satisfied

(2a) estimate $\frac{\partial E[J(\theta_n,C)]}{\partial \theta_n^i}$ by

$$\frac{\partial \tilde{J}(\theta_n)}{\partial \theta_n^i} = \frac{J^N\left(\theta_n^1,\ldots,\theta_n^i+\frac{\Delta\theta_n^i}{2},\ldots,\theta_n^p\right) - J^N\left(\theta_n^1,\ldots,\theta_n^i-\frac{\Delta\theta_n^i}{2},\ldots,\theta_n^p\right)}{\Delta\theta_n^i},$$

$$i=1,\ldots,p$$

(2b) $\theta_{n+1}^i = \theta_n^i + \Delta\theta_n^i \frac{\partial \tilde{J}(\theta_n)}{\partial \theta_n^i}, \quad i=1,\ldots,p$
Return θ_n.

A possible stopping criterion is to stop when the perceived progress between θ_n and θ_{n+1} is too small: $||\theta_{n+1}-\theta_n|| \le \varepsilon$ or $||J^N(\theta_{n+1}) - J^N(\theta_n)|| \le \varepsilon'$.

Example 7

We consider the MODERATO strategy given in Example 3. All the parameters are considered constant except θ_1 and θ_2. These two continuous variables are respectively constrained to be in the intervals [0; 2000°C day] and [0; 150 mm]. The function $J^N(\theta_1,\theta_2)$ is plotted on Figure 4, with the sequence of points θ_n generated by a typical gradient search method.

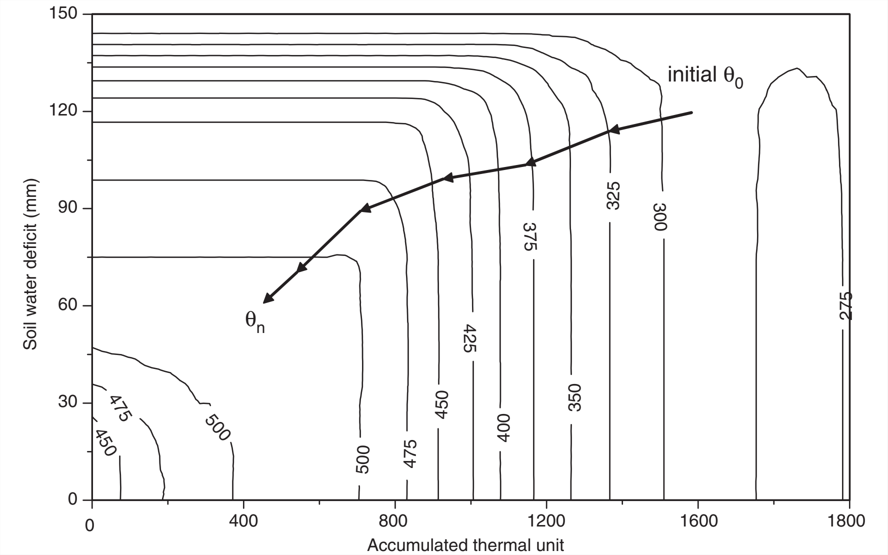


Figure 4. Plot of the parameter trajectory obtained using a gradient method to optimize two parameters of a MODERATO strategy with $N = 49$ historical weather series. The underlying iso-contour map was obtained using a systematic grid as explained in 6.4.c.

In the more general stochastic case where samples Ci are not *a priori* fixed but sampled during simulation with a random generator, efficient stochastic approximation (SA) algorithms have been developed, based on the original work by Kiefer and Wolfowitz (1952). SA algorithms are gradient search procedures that approach $\theta*$ using some original type of gradient estimation technique (e.g. perturbation analysis).

In the Kiefer–Wolfowitz method, the stochastic gradient estimate is defined:

$$\theta_{n+1}^{i} = \theta_{n}^{i} + a_n.\frac{J^N(\theta_n + c_n.e^i) - J^N(\theta_n - c_n \cdot e^i)}{2c_n} \qquad \forall i = 1, \ldots, p \tag{21}$$

where N is equal to 1 or set to a small integer, e_i is the vector with 1 for the ith component and 0 for the others, a_n and c_n are two series of positive real numbers. Some common values for the series are $a_n = a/n$ and $c_n = c/n^b$, where $a > 0, c > 0$ and $b \in$ [0, 0.5].

We can see that stochastic approximation is a variant of the gradient search method from deterministic optimization. Convergence to locally optimal parameters for optimization problem (20) can be guaranteed under appropriate conditions. This approach was successfully used by Cros et al. (2001) to derive the best values of some parameters involved in a grazing management strategy.

Also designed for continuous problems, the Nelder–Mead (simplex) method is an alternative to SA that is not based on the gradient, but instead uses a geometric figure to move from one point to another in the search space (Nelder and Mead, 1965). Botes et al. (1996), Mayer et al. (1996) and Parsons (1998) used this Nelder–Mead simplex method for determining optimal decisions for agricultural systems.

(c) Global methods

Global optimization algorithms are designed to explore systematically the value domain of $\theta \in IR^p$, in order to find local and global optimal solutions. For small values of p, the simplest systematic method for such problems is a grid-search, which consists in calculating the objective function J on all points located on a grid obtained by discretizing the domains $\Theta_i, i = 1, \ldots, p$ of the p management options, for a given precision for each variable (Bergez et al., 2002). A uniform discretization is obtained by

$$\theta^{i,j} = \theta_{\min}^{i} + j\Delta_i, \quad j = 1, \ldots, K_i \tag{22}$$

where Δ_i is the precision required for option i, and $K_i = \frac{\theta_{\max}^i - \theta_{\min}^i}{\Delta_i}$ is the size of the new discretized domain of Θ_i.

The grid search method thus transforms a continuous optimization problem into a discrete one, of size $K = K_1, K_2, \ldots, K_p$. For large-dimension problems ($p > 4$) and high precision requirements (small Δ_i) this method is not efficient since the number of grid points grows exponentially with p.

A more useful approach is therefore used to give priority to the evaluation of points θ within promising regions of IR^p that are likely to contain the optimal solution. Its main advantage is to maximize, for a given budget of time or simulation runs, the chance

Example 6 (cont'd)

We discretize the domain [500°, 1000°] into 6 values giving a precision for θ of 100°. The expected net margins R for these 6 θ values are estimated by simulation on 49 weather series (see Fig. 5). The value of the threshold parameter $\theta = 700°$ is the approximate optimal solution of this problem, with $R^* = 680€$.

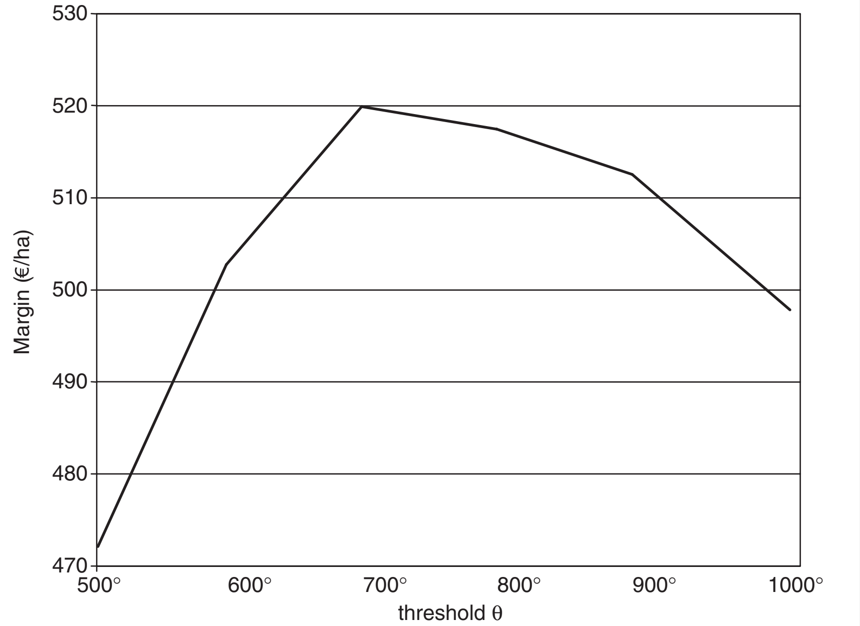


Figure 5. One-dimensional continuous problem solved by discretizing the domain [500°, 1000°] into 6 values. Expected margins are estimated by simulation using 49 weather series.

of finding a good solution. DIRECT (Jones et al., 1993) or MCS (Huyer and Neumaier, 1999) are such algorithms dedicated to deterministic problems and based on a hierarchical decomposition of the domain Θ. At each iteration of the search, a promising region is selected from a list of "pending regions". This selected region, a p-dimensional rectangle, is then broken down into k smaller ones. Each of these k new pending regions is then evaluated on some sampled points and ranked in the pending list. From the initial domain, these algorithms generate a sequence of rooted trees of rectangles (k-trees). The maximum depth of the principal tree is achieved when pending regions cannot be broken down any further (the maximum precision is reached).

DIRECT or MCS can be used for solving (20) when the context variable C is known in advance. These algorithms can also be adapted to the stochastic simulation framework, as in the P2P algorithm (Bergez et al., 2004) which is based on a hierarchical decomposition of Θ into 2^p-trees and an estimation J^N of the expected criterion by (17).

Example 7 (cont'd)

Using the P2P algorithm for optimizing the 2-parameters θ_1 and θ_2 of the irrigation strategy gives the result shown in Figure 6. One can see that P2P leads to a set of optimal parameters as previously shown in Figure 4.

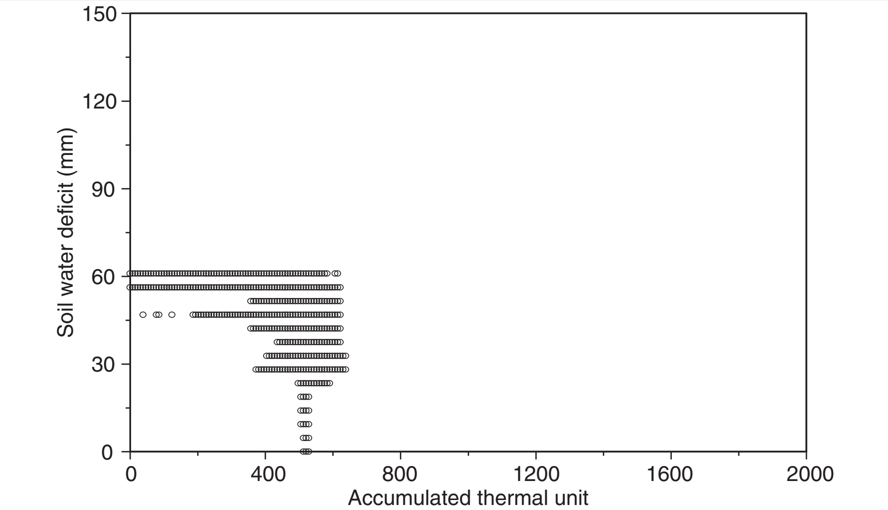


Figure 6. Optimal region obtained by P2P with a hierarchical decomposition of the two-dimensional domain.

7. Simulation-based control

Crop management problems for which management strategies are represented as a sequence of decision rules can be seen as a control problem for which efficient methods exist. The main advantage of this approach is that unlike simulation-based optimization methods, it allows an optimization of the structure of the decision rules, and not only of their parameters.

7.1. Modeling crop management as a Markov control problem

Let us consider a crop management problem that can be divided into a sequence of N sequential decision stages, typically from sowing to harvest. We assume that in each decision stage $i = 1, \ldots, N$ the crop process is characterized by state variables U_i, (with a continuous domain Θ_i^U in the general case) and that the values of the decision variables D_i (within domain Θ_i^D) have to be chosen. U_{N+1} describes the final state. The Markov control model relies on the assumption that the controlled dynamics of the state U_i follow Markovian transition probabilities:

$$U_{i+1} \sim dP_i(U_{i+1}|U_i, D_i), \quad i = 1, \ldots, N \tag{23}$$

This means that the distribution of the random variable U_{i+1} only depends on the current state U_i and on the decision D_i that was applied at stage i. From the crop model defined by Eq. (2), such a Markovian model can be obtained by choosing $U_i = U(t_i)$ and $D_i = D(t_i)$ for some instant t_i. Stochastic dynamics results from the context variable $C(t)$ that includes weather variables such as temperature or rainfall. Assuming the daily series $C(t)$ follows Markovian dynamics is not generally correct, but this approximation is reasonable for the $C(t_i)$ variables when $t_{i+1} \gg t_i$.

We also assume in Markov control models that during each transition from stage i to $i+1$, a deterministic return $r_i(U_i, D_i, U_{i+1})$ is obtained, and that the global criterion to be maximized is the expected value of the objective function:

$$J = r_1 + \cdots + r_i + \cdots + r_N. \tag{24}$$

7.2. Bellman's optimality equations

One can show that optimal solutions of Markov control problems are sequences of N optimal substrategies π_i that map U_i to D_i, $i = 1, \ldots, N$. By definition, once one is in the state U_i of the crop at stage i, the optimal decision to apply is given by $D_i = \pi_i(U_i)$.

This proposition is easy to understand. Assume that the crop process is in a state $U = u$ at the last stage N. The optimal decision to execute in that state is clearly the one that maximizes the expectation of the final return r_N, so we have:

$$\begin{aligned}\pi_N(u) &= \arg\max_{d \in \Theta_N^D} E\left[r_N\left(u, d, u'\right)\right]\\ &= \arg\max_{d \in \Theta_N^D} \int_{u'} r_N\left(u, d, u'\right) dP_N\left(u' | u, d\right), \quad \forall u \in \Theta_N^U\end{aligned} \tag{25}$$

and the optimal average gain obtained in this state u will be:

$$V_N(u) = \max_{d \in \Theta_N^D} E\left[r_N\left(u, d, u'\right)\right], \quad \forall u \in \Theta_N^U \tag{26}$$

Example 8

We consider here the use of the simulation model DÉCIBLÉ for optimizing wheat crop management strategies (Chatelin et al., 2005) defined as a sequence of $N = 3$ decision steps, respectively sowing, first and second nitrogen applications. The corresponding state and decision variables are presented in Table 4. The state variables for the 2 nitrogen applications are chosen for their capacity to summarize the past trajectory of the process, and thus to approach the Markov property as closely as possible. The sowing time dS is introduced as a random state variable. Note the relative definition of the dN_1 and dN_2 domains, meaning that a policy does not specify an absolute date for nitrogen application, but rather a date relative to the start of tillering or stem elongation. The objective function is net margin.

$$J = \text{Yield} - \alpha(qN_1 + qN_2) - \beta qS.$$

Table 4. State spaces and decision spaces of a Markov control model for wheat crop management.

	Sowing	1st nitrogen application	2nd nitrogen application
State Variables	• Sowing time $dS \in$ [01/10,15/12]	• Tillering $dT \in$ [15/11,01/04] • Number of plants $NP \in$ [0,200]	• Residual soil nitrogen $Ns \in$ [0,100] kg ha^{-1} • Start of stem elongation $d1cm \in$ [15/02,15/05] • Aerial biomass $ba1cm \in$ [0,200] g m^{-2}
Decision Variables	• Seed rate $qS \in$ [100,200] g/m^2 • Wheat variety $vS \in$ {soissons, artaban, ..}	• Date $dN1 \in$ [$dT-5$, $dT+20$] • Quantity $qN1 \in$ [0,100] kg ha^{-1}	• Date $dN2 \in$ [$d1cm-5$, $d1cm+20$] • Quantity $qN2 \in$ [0,200] kg ha^{-1}

Assume now that the crop process is in state u at stage $N-1$. The optimal decision to execute in that state is the one that maximizes the sum of the final returns $r_{N-1}+r_N$, knowing that in any resulting state u' at stage N, the optimal policy $\pi_N(u')$ will be applied, giving a return $r_N = V_N(u')$. We thus have:

$$\begin{aligned}\pi_{N-1}(u) &= \arg\max_{d\in\Theta^D_{N-1}} E\left[r_{N-1}(u,d,u') + V_N(u')\right]\\ &= \arg\max_{d\in\Theta^D_{N-1}} \int_{u'} \left(r_{N-1}(u,d,u') + V_N(u')\right) dP_{N-1}(u'\,|u,d),\\ &\quad \forall u \in \Theta^U_{N-1}\end{aligned} \tag{27}$$

and the optimal average gain obtained in this state at stage $N-1$ will be:

$$V_{N-1}(u) = \max_{d\in\Theta^D_{N-1}} E\left[r_{N-1}(u,d,u') + V_N(u')\right], \quad \forall u \in \Theta^U_{N-1}. \tag{28}$$

The optimization can be continued for stages $i = N-2$, $i = N-3, \ldots$ until $i = 1$. This process defines a general set of equations called Bellman's optimality equations (Puterman, 1994). Solutions of these optimality equations are N management policies $\pi_i : \Theta_i^U \to \Theta_i^D$, $i = 1, \ldots, N$. Following these policies from the initial state U_1 will define on the fly a sequence of decisions D_i that maximize the expected return $E[r_1 + \cdots + r_i + \cdots r_N]$.

7.3. Solving Markov control problem with stochastic dynamic programming

When crop management problems have been modelled as Markov control problems, the classical approach for automatically generating optimal strategies is to apply stochastic dynamic programming (Kennedy, 1990). Crop simulation models are used for estimating the Markov model parameters within the Bellman optimality equations, that is the net returns r_i and the Markovian transition probabilities $dP_i(U_{i+1}|U_i, D_i)$ of the crop dynamics (Epperson et al., 1993; Bergez et al., 2001b). The state and decision variables U_i and D_i are generally discretized into finite sets $\Theta_i^U = \{1, \ldots, n_i^U\}$ and $\Theta_i^D = \{1, \ldots, n_i^D\}$. Rewards r_i are then obtained by averaging corresponding simulated $r_i(U_i, D_i, U_{i+1})$ values, and transition probabilities are estimated by simulation with a maximum-likelihood approach:

$$P_i\big(u'\,|u,d\big) \approx \frac{N_i\big(u,d,u'\big)}{\sum_{u''} N_i\left(u,d,u''\right)}, \quad u = 1, \ldots, n_i^U, \quad u' = 1, \ldots, n_{i+1}^U, \quad d = 1, \ldots, n_i^D, \tag{29}$$

where $N_i(u, d, u')$ is the number of simulated transitions at stage i between states u and u' with decision d. Note that special attention has to be taken when defining discretized states, since the discretization of continuous processes generally leads to non-Markovian dynamics. In practice, the size of the grids has to be chosen carefully taking into consideration the quality of the corresponding optimal management strategy, the required memory size and the computation time of the dynamic programming algorithm used for solving Bellman's equations. This algorithm, also called Value Iteration, is defined as follows:

Stochastic dynamic programming algorithm (Value Iteration)

(1) Initialize $V_{N+1} = 0$
(2) For $i = N, \ldots, 1$

$$\begin{aligned} V_i(u) &= \max_{d=1,\ldots,n_i^D} \sum_{u'=1}^{n_{i+1}^U} P_i\big(u'\,|u,d\big)\big(r_i\big(u,d,u'\big) + V_{i+1}\big(u'\big)\big), \\ \pi_i(u) &= \arg\max_{d=1,\ldots,n_i^D} \sum_{u'=1}^{n_{i+1}^U} P_i\big(u'\,|u,d\big)\big(r_i\big(u,d,u'\big) + V_{i+1}\big(u'\big)\big) \quad u = 1, \ldots, n_i^U \end{aligned} \tag{30}$$

In practice, for each i and d, the transition probabilities $P_i(u'|u, d)$ can be encoded in a transition probability matrix, which is simply a two-dimensional array whose element at the lth row and mth column is $p_{lm} = P_i(m|l,d)$:

$$P_i^d = \begin{bmatrix} p_{11} & \cdots & p_{1n_{i+1}^U} \\ \vdots & p_{lm} & \vdots \\ p_{n_i^U 1} & \cdots & p_{n_i^U n_{i+1}^U} \end{bmatrix}.$$

The rewards $r_i(u,d,u')$ can be similarly represented by vectors $R_i^d = [r_l]$ in $IR^{n_i^U}$, with $r_l = E[r_i(l,d,u')] = \sum_{u'=1}^{n_{i+1}^U} P_i(u'|l,d) r_i(l,d,u')$. The value iteration algorithm (Eq. (30)) can then be written simply in vector notation:

$$V_i = \max_{d=1,\ldots,n_i^D} \left\{ R_i^d + P_i^d V_{i+1} \right\}, \quad i = N, \ldots 1, \tag{31}$$

where the optimal decisions are computed for each component of V_i.

7.4. Reinforcement learning of optimal management strategies

Crop management problems are often efficiently approximated by Markov models and optimal strategies obtained by solving these control problems are generally quite efficient. However, such an indirect method (model estimation + dynamic programming) is often inappropriate when faced with large state and decision spaces, where it is difficult to compute and store transition probabilities.

Some promising improvements of this method have been shown recently with the reinforcement learning approach (Bertsekas and Tsitsikli, 1996; Sutton and Barto, 1998; Gosavi, 2003). Reinforcement learning, also called neuro-dynamic programming, directly approximates the solution of the Bellman equations during simulation, without having to estimate rewards and probabilities. Today, reinforcement learning is one of the major approaches to solving sequential control problems with unknown transition probabilities and/or with large state variable domains.

The most studied reinforcement learning algorithm is Q-learning. The principle of Q-learning with a finitehorizon is to learn by simulation for each state $u \in \Theta_i^U$ and decision $d \in \Theta_i^D$ an estimate of the Q_i-value function at stage i:

$$Q_i(u,d) = E\left[r_i\left(u,d,u'\right) + V_{i+1}\left(u'\right)\right], \quad \forall u \in \Theta_i^U, d \in \Theta_i^D. \tag{32}$$

The value $Q_i(u,d)$ represents the expected sum of the future returns assuming the decision d is applied in the current state u at stage i, and that for the subsequent stages $j > i$ an optimal strategy is followed. Equation (32) shows that once these estimates have been learned, the optimal policy can be obtained through:

$$\pi_i(u) = \arg\max_{d \in \Theta_i^D} Q_i(u,d), \quad \forall u \in \Theta_i^U. \tag{33}$$

Q-learning regularly updates the Q_i estimates after each observed transition (u,d,u',r), from stage $i = 1$ to $i = N$, along T simulated trajectories. These trajectories are obtained by choosing an initial state $u \in \Theta_i^U$ and at each stage i either the current optimal decision according to Eq. (33) or a random decision $d \in \Theta_i^D$:

Finite-Horizon Q-learning algorithm

```
1) Initialize $Q_i = 0, i = 1, \ldots, N$
2) For $t = 1, \ldots, T$
2a) Choose $u_1 \in \Theta_1^U$
2b) For $i = 1, \ldots, N$
      • Choose $d_i \in \Theta_i^D$
      • $(u_{i+1}, r_i) =$ Simulate $(u_i, d_i)$ at stage $i$
      • Update $Q_i(U_i, d_i, u_{i+1}, r_i)$
3) Return $Q_i$
```

In this algorithm, Simulate (u, d) is a function simulating a random transition from state u when action d is applied and Update $Q_i(u, d, u', r)$ is the Q-learning update rule for Q_i, defined in the case of discrete spaces Θ_i^U and Θ_i^D as:

Update $Q_i(u, d, u', r)$ (discrete representation):

$$Q_i(u,d) \leftarrow Q_i(u,d) + \varepsilon\left(r + \max\nolimits_{d' \in \Theta_{i+1}^D} Q_{i+1}\left(u', d'\right) - Q_i(u,d)\right). \tag{34}$$

The parameter ε is a small learning rate that decays toward 0 as the number of observed transitions increases.

When Θ_i^U and Θ_i^D are very large or continuous domains, this simple learning rule may not be efficient and a more useful approach consists in using parameterized representations of Q_i:

$$Q_i(u,d) = f_i\left(\varphi_1(u,d), \ldots, \varphi_p(u,d); \theta_i\right), \tag{35}$$

where the functions f_i are linear or non-linear transformations that map important features $\varphi_k(u,d)$ of the state and decision variables to Q_i-values. Here the parameters θ_i are "weights" that have to be optimized during simulation:

Update $Q_i\langle u, d, u', r\rangle$ (parameterized representation):

$$\theta_i \leftarrow \theta_i + \varepsilon\left(r + \max_{d' \in \Theta_{i+1}^D} Q_{i+1}\left(u', d'\right) - Q_i(u,d)\right) \nabla_{\theta_i} f_i \tag{36}$$

where $\nabla_{\theta_i} f_i$ is the gradient of f_i with respect to θ_i in (u, d).

With such parameterized representations, very large sequential decision problems can be solved approximately. The main drawback of this approach is that the near-optimal policies that are obtained are not really suitable with respect to intelligibility and ease of applicability by farmers or agronomists. To overcome this one can use additional procedures that automatically extract simpler structures such as decision trees from the Q$_i$-value functions, which lead directly to strategies represented as a set of decision rules (Garcia, 1999).

8. A comparison between optimization and control

We have seen two general methods for solving stochastic optimization problems by simulation: simulation optimization and control-based optimization. Here we discuss the advantages and disadvantages of these approaches with regard to the use of crop models for optimizing management decisions.

No doubt the simplest approach is to optimize the parameters of a management strategy. For problems with many discrete parameters, algorithms like the genetic algorithm or tabu search generally furnish approximate solutions. If the parameters are continuous, the various stochastic optimization methods that we have discussed can be used. These methods require no assumptions about the crop model or the decision model, except that the latter must be defined up to a vector of parameters.

Example 9

We consider the problem described in Example 6, where we want to optimize the starting date of an irrigation period. This problem is treated here as a sequential decision problem under uncertainty: each day, farmers observe the physiological stage of development of the crop and the soil water deficit, a measure of the soil water content. On the basis of these observations they must decide each day whether to continue waiting or to start the first irrigation period. The state of this decision process is defined every day by the two state-variables Δ and σ, where δ is the soil water deficit, and σ is the accumulated thermal units above 6°C since sowing. Both variables are continuous. The ranges of δ and σ are respectively the intervals Δ and Σ. In any state (δ, σ) there are only two possible decisions: wait until the next day (W), and start irrigation today (I). Once action I is selected, we assume that a fixed strategy is applied for the subsequent irrigation decisions (the same as in Example 6). The only optimization here concerns the decision rule for starting irrigation, defined as a function that maps each possible state (δ, σ) in $\Delta \times \Sigma$ to an action W or I.

We used both stochastic dynamic programming (DP) and reinforcement learning (RL) to calculate optimal decision rules. The MODERATO simulation model for corn crop management was used to define the dynamics of the system. For DP we discretized the domains Δ and Σ on a regular grid. We used simulation for estimating the discrete transition probabilities from one grid point to the others, and then stochastic dynamic programming algorithms for computing approximate optimal decision rules. RL does not require an *a priori* estimation of the transition probabilities and an approximate optimal decision rule on the discretized domain is directly obtained by simulation and learning.

Figure 7 represents the average value (over 1000 runs) of the best decision rules obtained by dynamic programming and reinforcement learning (Bergez et al., 2001b). The first conclusion seems to be that RL performs better than DP when a SMALL number of simulations are available. When the number of trajectories T is sufficiently large ($T > 100\,000$), all the policies are equivalent. Surprisingly, better results are obtained with fewer grid points. For both approaches, approximate decision rules are complex mappings from $\Delta \times \Sigma$ to $\Delta\{W, I\}$.

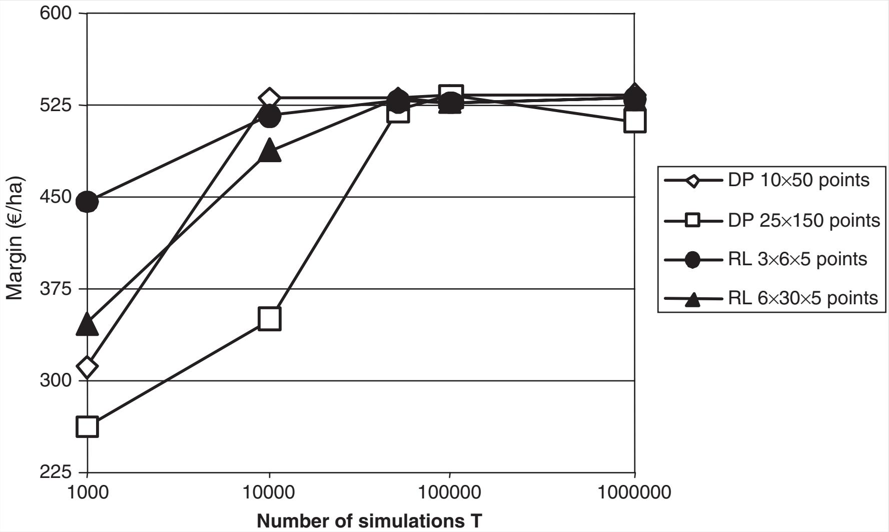


Figure 7. Average value of the optimal strategies obtained by dynamic programming and reinforcement learning as a function of the number of trajectories T and the grid sizes.

One can note that the optimal values obtained by dynamic programming or reinforcement learning are very close to the average value of the optimal structured decision rule obtained in Example 6.

The use of optimal control methods with simulation is more demanding. In particular, the decision problem must have the form of a Markov decision process. However, this will often be approximately true for decision problems for a single field, even though climate transition probabilities are not exactly Markovian. On the other hand, this approximation will, in general, not be acceptable for management decisions at the farm level, where the decisions can no longer be expressed as a simple sequence of decisions. A further disadvantage here is the complexity of reinforcement learning algorithms. Choosing the algorithm, programming it and analyzing convergence are all long and difficult tasks that require experience.

Despite these drawbacks, the optimal control methods do have a great advantage, namely, they do not require the form of the decision rules be specified in advance. They can be used to suggest decision rules even in cases where one has little idea what form the optimal rules will take, as might be the case for instance, when new criteria or new constraints appear. In that case one might envision a mixed approach. Optimal control methods would first be used to suggest the form of the decision rules, then optimization methods would be used to determine the best parameters for those decision rules.

References

Affholder, F., Scopel, E., 2001. Modèle de culture et diagnostic agronomique régional. In : Malézieux, E., Trébuil, G., and Jaeger, M., (Eds), Modélisation des agroécosystèmes et aide à la décision, INRA-CIRAD, coll: Repères, 107–125.

Andradóttir, S., 1998. Simulation optimization. In : Banks, J. (Ed), Handbook of Simulation. Principles, methodology, advances, applications, and practice. John Wiley & Sons Inc., New York, 307–334.

Attonaty, J.M., Chatelin, M.H., Poussin, J.C., Soler, L.G., 1994. OTELO: un simulateur à base de connaissance pour raisonner équipement et organisation du travail. Cahiers des Chambers d'Agriculture 66(7), 37–49.

Aubry, C., Chatelin, M.H., Verjux, N., 1996. Déciblé: guide de l'utilisateur. ITCF.

Aubry, C., Papy, F., Capillon, A., 1998. Modelling decision-making process for annual crop management. Agricultural Systems 56(1), 45–65.

Azadivar, F., 1999. Simulation optimization methodologies. Proceedings of the 1999 Winter Simulation Conference, 5–8 December, 1999, Squaw Peak, Phoenix, AZ, USA, 93–100.

Batchelor, W.D., Jones, J.W., Boote, K.J., Pinnschmidt, H.O., 1993. Extending the use of crop models to study pest damage. American Society of Agricultural Engineers 36, 551–558.

Bergez, J.-E., Debaeke, Ph., Deumier, J.-M., Lacroix, B., Leenhardt, D., Leroy, P., Wallach, D., 2001a. MODERATO: an object-oriented decision model to help on irrigation scheduling for corn crop. Ecological Modeling 137(1), 43–60.

Bergez, J.E., Eigenraam, M., Garcia, F., 2001b. Comparison between dynamic programming and reinforcement learning: a case study on maize irrigation management. EFITA 2001, Third European Conference of the European Federation for Information Technology in Agriculture, Food and the Environment, 18–20 June 2001, Montpellier (FR), 343–348.

Bergez, J.E., Deumier, J.M., Lacroix, B., Leroy, P., Wallach, D., 2002. Improving irrigation schedules by using a biophysical and a decisional model. European Journal of Agronomy 16, 123–135.

Bergez, J.E., Garcia, F., Lapasse, L., 2004. A Hierarchical Partitioning Method for Optimizing Irrigation Strategies. Agricultural Systems 80, 235–253.

Bertsekas, D.P., Tsitskli, J.N., 1996. Neuro-dynamic programming. Athena Scientific, Belmont, USA.

Botes, J.H.F., Bosch, D.J., Oosthuizen, L.K., 1996. A simulation and optimization approach for evaluating irrigation information. Agricultural Systems 51, 165–183.

Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M.H., Ruget, F., Nicoullaud, B., Gate, P., Devienne-Barret, F., Antonioletti, R., Durr, C., Richard, G., Beaudoin, N., Recous, S., Tayot, X., Plenet, D., Cellier, P., Machet, J.M., Meynard, J.-M., Delécolle, R., 1998. STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parametrization applied to wheat and corn. Agronomie 18, 311–346.

Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche, D., Zimmer, D., Sierra, J., Bertuzzi, P., Burger, P., 2003. An overview of the crop model. European Journal of Agronomy 18, 309–332.

Cabelguenne, M., Debaeke, P., Puech, J., Bosc, N., 1996. Real time irrigation management using the EPIC-PHASE model and weather forecasts. Agricultural and Water Management 32, 227–238.

Chatelin, M.H., Aubry, C., Poussin, J.C., Meynard, J.M., Massé, J., Verjux, N., Gate, P. and Le Bris, X., 2005. DéciBlé, a software package for wheat crop management simulation, Agricultural Systems 83, 77–99.

Cros, M.-J., Duru, M., Garcia, F., Martin-Clouaire, R., 2001. Simulating rotational grazing management. Journal of Environment International 27, 139–145.

Doré, T., Sebillotte, M., Meynard, J.M., 1997. A diagnostic method for assessing regional variations in crop yield. Agricultural Systems 54: 169–188.

Epperson, J.E., Hook, J.E., Mustafa, Y.R., 1993. Dynamic programming for improving irrigation scheduling strategies of maize. Agricultural Systems 42, 85–101.

Fleischer, M.A., 1995. Simulated annealing: past, present, and future. Proceedings of the 1995 Winter Simulation Conference, 3–6 December 1995, Hyatt Regency Crystal City, Arlington, VA, USA, 155–161.

Fu, M.C., 2001. Simulation Optimization. Proceedings of the 2001 Winter Simulation Conference, 9–12 December 2001, Crystal Gateway Marriott, Arlington, VA, USA, 53–61.

Garcia, F., 1999. Use of reinforcement learning and simulation to optimize wheat crop technical management. In Proceedings of the International Congress on Modelling and Simulation (MODSIM'99), Hamilton, New Zealand, 801–806.

Ghaffari, A., Cook, H.F., Lee, H.C., 2001. Simulating winter wheat yields under temperate conditions: exploring different management scenarios. European Journal of Agronomy 15: 231–240.

Glover, F., Laguna, M., 1997. Tabu Search. Kluwer Academic Publishers, Boston.

Goldsman, D., Nelson, B.L., 1994. Ranking, selection and multiple comparisons in computer simulation. Proceedings of the 1994 Winter Simulation Conference, 11–14 December 1994, Walt Disney World Swan Hotel, Orlando, FL, USA, 192–199.

Gosavi, A., 2003. Simulation-based optimization: parametric optimization techniques and reinforcement learning. Kluwer Academic Publishers.

Gurkan, G., Ozge, A.Y., Robinson, S.M., 1994. Sample-path optimization and simulation. Proceedings of the 1994 Winter Simulation Conference, 11–14 December 1994, Walt Disney World Swan Hotel, Orlando, FL, USA, 247–254.

Hearn, A.B., 1994. OZCOT: a simulation model for cotton crop management. Agricultural Systems 44, 257–299.

Ho, Y.C., Sreenivas, R., Vakili, P., 1992. Ordinal optimization of discrete event dynamic systems. Discrete Event Dynamical Systems 2, 61–88.

Homem-de-Mello, T., 2003. Variable-sample methods for stochastic optimization. ACM Transactions on Modeleling Computer and Simulation 13(2), 108–133.

Hsu, J.C., 1996. Multiple comparisons: theory and methods. Chapman & Hall, London, England.

Huyer, W., Neumaier, A., 1999. Global optimization by multilevel coordinate search. Journal of Global Optimization, 331–355.

Jallas, E., Sequeira, R., Boggess, J.E., 1998. Evolutionary algorithms for knowledge discovery and model-based decision support. 3rd IFAC/CIGR workshop on AI in Agriculture, 24–26 April 1998, Makuhari, Chiba, Japan, 120–125.

Jones, D.R., Perttunen, C.D., Stuckman, B.E., 1993. Lipschitzian optimization without the Lipschitz constant. Journal of Optimization Theory and Applications 79, 157–181.

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., 2003. An overview of APSIM, a model designed for farming systems simulation. European Journal of Agronomy 18, 267–288.

Kennedy, J.O., 1986. Dynamic Programming: application to agricultural and natural resources, Elsevier Applied Science, London.

Kiefer, J., Wolfowitz, J., 1952. Stochastic estimation of the maximum of a regression function. Annals of Mathematical Statistics 23, 462–466.

Kropff, M.J., Bouma, J., Jones, J.W., 2001. Systems approaches for the design of sustainable agro-ecosystems. Agricultural Systems 70, 369–393.

Leenhardt, D., Trouvat, J.L., Gonzalès, G., Pérarnaud, V., Prats, S., Bergez, J.E., 2004. Estimating irrigation demand for water management on a regional scale. I. ADEAUMIS, a simulation platform based on bio-decisional modelling and spatial information. Agricultural and Water Management 68, 207–232.

Maton, L., Leenhardt, D., Goulard, M., Bergez, J.E., 2005. Assessing the irrigation strategies over a wide geographical area from structural data about farming systems. Agricultural Systems 86, 293–311.

Matthews, R., 2002. Crop management. In : Matthews, R., Stephens, W., (Eds), Crop-soil simulation models: applications in developing countries. CABI editions, pp. 29–53.

Mayer, D.G., Belward, J.A., Burrage, K., 1996. Use of advanced techniques to optimize a multi-dimensional dairy model. Agricultural Systems 50, 239–253.

Mayer, D.G., Belward, J.A., Burrage, K., 1998a. Optimizing simulation models of agricultural systems. Annals of Operations Research 82, 219–231.

Mayer, D.G., Belward, J.A., Burrage, K., 1998b. Tabu search not an optimal choice for models of agricultural systems. Agricultural Systems 58, 243–251.

Mayer, D.G., Belward, J.A., Burrage, K., 2001. Robust parameter settings of evolutionary algorithms for the optimisation of agricultural systems models. Agricultural Systems 69, 199–213.

Meynard, J.M., David, C., 1992. Diagnostic que l'élaboration du rendement des cultures. Cahiers Agriculture 1, 9–19.

Michalewicz, Z., Schoenauer, M., 2001. Evolutionary algorithms. Encyclopedia of Operations Research and Management Science, 2nd edition. Kluwer Academic Publishers, Boston, pp. 264–269.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Computational Journal 7, 308–313.

Norkin, W.I., Pflug, G.C., Ruszczyński, A., 1998. A branch and bound method for stochastic global optimization. Journal of Mathematical Programming 83, 425–450.

Pang, X.P., 1997. Yield and nitrogen uptake prediction by CERES-maize model under semiarid conditions. Soil Science Society of America Journal 61, 254–256.

Papy, F., 2000. Farm models and decision support: a summary review. In: Colin, J.P., and Crawford, E.W. (Eds), Research on agricultural systems: accomplishments, perspectives and issues. Nova Science Publishers, Inc., New York, pp. 89–107.

Papy, F., Attonaty, J.-M., Laporte, C., Soler, L.-G., 1988. Work organization simulation as a basis for farm management advice. Agricultural Systems 27, 295–314.

Parsons, D.J., 1998. Optimizing silage harvesting plans in a grass and grazing simulation using the revised simplex method and a genetic alogorithm. Agricultural Systems 56, 29–44.

Puterman, M.L. 1994. Markov Decision Processes, Wiley, New York.

Reddy, V.R., Acock, B. and Whisler, F.D., 1995. Crop management and input optimization with GLYCIM: differing cultivars. Computers and Electronics in Agriculture 13, 37–50.

Roy, B., 1996. Multicriteria methodology for decision aiding. Kluwer, Dordrecht.

Shi, L., Ólafsson, S., 1997. An integrated framework for deterministic and stochastic optimization. Proceedings of the 1997 Winter Simulation Conference, 7–10 December 1997, Renaissance Waverly Hotel, Atlanta, GA, USA, 358–365.

Stockle, C.O., Donatelli, M., Nelson, R., 2003. CropSyst, a cropping systems simulation model. European Journal of Agronomy 18, 289–307.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: an introduction. MIT Press, Cambridge.

Swaney, D.P., Jones, J.W., Boggess, W.G., Wilkerson, G.G., Mishoe, J.W., 1983. Real-time irrigation decision analysis using simulation. Transaction of the ASAE 26, 562–568.

Swisher, J.R., Hyden, P.D., Jacobson, S.H., Schruben, L.W., 2000. A survey of simulation optimization techniques and procedures. Proceedings of the 2000 Winter Simulation Conference, 10–13 December 2000, Wyndham Palace Resort & Spa, Orlando, FL, USA, 119–128.

Van Evert, F.K., Campbell, G.S., 1994. CropSyst: a collection of object-oriented simulation models of agricultural systems. Agronomy Journal 86, 325–331.

Welch, S.M., Jones, J.W., Brennan, M.W., Reeder, G., Jacobson, B. M., 2002. PCYield: model-based decision support for soybean production. Agricultural Systems 74, 79–98.

Exercises

1. An irrigation strategy was simulated with MODERATO for $N = 49$ consecutive weather series, from year 1955 to 2003. The net margin J_i results are given in Table 1, $i = 1, \dots, 49$.
   (a) Plot the empirical probability density function of J. What are the mean m and variance s^2 of this distribution?
   (b) Assuming that the J_i are independent random variables with a normal distribution $N(m, s^2)$, how many years must be simulated in order to make the error in the estimation of m lower that 5% with a probability >0.9?

2. Let π_1 and π_2 be two candidate irrigation strategies as in the previous exercise. Table 2 gives the net margin results obtained with π_1 and π_2 for a sequence of $N = 49$ historical weather series.
   (a) Assuming a normal distribution for J^N, estimate the probability that π_1 is better than π_2 based on the first 5 weather series.
   (b) Same question for $N = 10$ and $N = 49$.
3. We consider the two previous candidates π_1 and π_2. Assuming a normal distribution for J^N, calculate from Table 2 the probability of correct selection of π_1 for all the

Table 1. Net margin J obtained by simulating an irrigation strategy with 49 weather series.

Year	Margin (€/ha)	Year	Margin (€/ha)	Year	Margin (€/ha)
1955	584	1975	841	1995	653
1956	669	1976	786	1996	653
1957	839	1977	862	1997	628
1958	668	1978	837	1998	636
1959	701	1979	753	1999	739
1960	884	1980	888	2000	574
1961	957	1981	793	2001	623
1962	923	1982	839	2002	710
1963	919	1983	998	2003	700
1964	841	1984	959		
1965	921	1985	766		
1966	971	1986	901		
1967	870	1987	785		
1968	828	1988	670		
1969	819	1989	707		
1970	726	1990	816		
1971	887	1991	803		
1972	850	1992	784		
1973	758	1993	701		
1974	880	1994	708		

Table 2. Simulated net margins for two irrigation strategies for 49 weather series.

Run	Margin (€/ha)		Run	Margin (€/ha)		Run	Margin (€/ha)	
	π_1	π_2		π_1	π_2		π_1	π_2
1	474	486	21	739	760	41	418	411
2	451	466	22	542	562	42	406	403
3	826	768	23	724	711	43	450	421
4	484	505	24	740	710	44	515	505
5	523	535	25	673	647	45	649	612
6	757	727	26	894	847	46	392	401
7	806	814	27	606	624	47	472	479
8	827	801	28	645	625	48	551	499
9	818	765	29	884	816	49	627	589
10	763	738	30	779	786			
11	790	764	31	571	594			
12	886	830	32	862	787			
13	797	771	33	679	686			
14	600	612	34	572	572			
15	824	821	35	571	571			
16	609	627	36	629	638			
17	750	774	37	651	658			
18	687	675	38	518	532			
19	530	562	39	662	613			
20	814	789	40	694	639			

allocations (N_1, N_2) such that $N_{\text{tot}} = N_1 + N_2 = 50, N_i > 5$. What is the optimal allocation?

4. We consider the irrigated corn crop management strategy described in Example 3. Table 3 gives the domain for all the continuous management options of this strategy and the precision required for each of these variables.
   (a) Calculate the size of the grid and the number of simulation runs of the crop model if each possible management strategy has to be evaluated on 49 historical weather series and 3 different soils.
   (b) Divide the required precision by 2 for each management option and calculate the new size of the grid and number of simulation runs that are necessary.
5. We consider a simple problem with $N = 2$ stages, 2 states and 2 possible decisions at each stage. The transition matrices P_i^d and the rewards vectors R_i^d are:

$$P_1^1 = \begin{bmatrix} 0.1 & 0.9 \\ 0.2 & 0.8 \end{bmatrix}, \quad P_1^2 = \begin{bmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \end{bmatrix}, \quad P_2^1 = \begin{bmatrix} 0.4 & 0.6 \\ 0.5 & 0.5 \end{bmatrix}, \quad P_2^2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \end{bmatrix}$$

$$\text{and } R_1^1 = \begin{bmatrix} 15 \\ 10 \end{bmatrix}, \quad R_1^2 = \begin{bmatrix} 7 \\ 11 \end{bmatrix}, \quad R_2^1 = \begin{bmatrix} 11 \\ 5 \end{bmatrix}, \quad R_2^2 = \begin{bmatrix} 4 \\ 14 \end{bmatrix}.$$

Table 3. Domains and discretization step sizes of the management parameters of a corn crop management strategy as described in Example 3.

Name	Meaning	Unit	Min	Max	Δmin
θ_1	Accumulated thermal unit to start the irrigation campaign	ºC day	0	2000	5
θ_2	Soil water deficit to start the irrigation	mm	0	150	3
θ_3	Irrigation applied at the first irrigation	mm	5	50	2
θ_4	Soil water deficit to start a new irrigation cycle	mm	50	130	3
θ_5	Irrigation depth applied after the first irrigation round	mm	5	50	2
θ_6	Accumulated thermal units to stop the irrigation	ºC day	1400	1800	5
θ_7	Soil water deficit to stop irrigation	mm	50	130	3
θ_8	Irrigation applied at the last irrigation round	mm	5	50	2

(a) Calculate the optimal policy of this Markov control problem with the stochastic dynamic programming algorithm (Eq. (30)). Compare this solution to the 4 *a priori* sequences of decisions $(d_1, d_2) = (1,1), (1,2), (2,1)$ and $(2,2)$. Which management strategy has the larger expected total reward $J = r_1 + r_2$?

(b) Calculate the memory size required for storing the Markov model parameters and the number of elementary mathematical operations executed by the dynamic programming algorithm as a function of N, n_i^U and n_i^D.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 7

Using crop models for multiple fields

D. Leenhardt, D. Wallach, P. Le Moigne, M. Guérif, A. Bruand
and M.A. Casterad

1. Introduction **209**
2. Physical input data **211**
3. Management practices **222**
4. Remote sensing data **228**
5. Obtaining the outputs for multiple fields **233**
6. Evaluation **235**
References **238**
Exercises **245**

1. Introduction

The original use of crop models was to calculate crop growth and development for a single field with supposedly homogeneous soil, climate, initial conditions and management practices. This is indeed still a basic use of crop models. However, there is also increasing interest in studies that concern multiple fields (see Hartkamp et al., 1999; Hansen and Jones, 2000; Russell and Van Gardingen, 1997; Leenhardt et al., 2004a,b). In some cases each field can be treated independently, but it is the combined result from all fields that is of interest. Examples include the calculation of crop yields or forage yields on a regional or national basis (e.g. Lal et al., 1993; Rosenthal et al., 1998; Chipanshi et al., 1999; Donet, 1999; Faivre et al., 2000; Yun, 2002), the calculation of water requirements for agriculture within the area served by a water provider (e.g. Sousa and Santos Pereira, 1999; Heinemann et al., 2002; Leenhardt et al., 2004a) or total emission of nitrogen oxides from agricultural land in a region. In other cases, it is necessary to model not only individual fields but also interactions between fields or between a field and non-crop surroundings. For example, the problem addressed might involve nitrogen, herbicide or pesticide pollution of streams or ground water due to runoff or leaching from agricultural land (e.g. Beaujouan et al., 2001; Gomez and Ledoux, 2001). Another example would involve transfer of genetically modified pollen from one field to surrounding fields. In all of these problems there is a need to use the crop model for multiple fields, perhaps hundreds or thousands of fields, with different soils, climate and management.

A major problem in all these studies is obtaining the input data necessary to run the crop model. This is considered in the next two sections. The first concerns physical input data (climate, soil characteristics and initial conditions). We discuss the types of data that are usually available and review methods that have been proposed for associating values of the input variables with each point in space. For climate, we also consider approaches that can be used for prediction or for scenario testing. The second group of input variables considered, in Section 3, are management variables (choice of crop, sowing date, irrigation, etc.). These are often unavailable even for the past. One simple approach is to assume that management practices are fixed for a region. A more complex approach is to assume that management practices are determined by decision rules, which relate practices to other input variables and to the state of the crop (See also Chapter 6).

In Section 4, we discuss the relation of remote sensing data to the problem of running a crop model for multiple fields. Remote sensing provides detailed spatially explicit data for an entire area. However, the data provided are not directly the data needed by crop models. This section describes how the remote sensing data can be used. In Section 5, we discuss how one obtains the outputs that are sought, for example national yield. In Section 6, we consider the problems of evaluation that are specific to the case where the model is used for multiple fields.

We must distinguish a large number of different situations. The problem may involve interactions between fields or not. Each type of input data may be available at each field, only for some fields or only indirectly. Table 1 presents a number of studies that have been reported, categorized by some of these choices.

Table 1. Different situations involving the use of crop models for multiple fields.

Output	Objective	Description of management practices	Interactions between fields	References with examples
Sum over representative fields	Prediction for current season	Actual past practices, future decision rules	No	Faivre et al. (2000) Launay (2002) Leenhardt et al. (2004)
	Diagnosis	Actual past practices	No	Sousa and Pereira (1999) Donet et al. (2001) Heineman et al. (2002)
	Scenario testing	Decision rules or hypothetical decisions	No	Lal et al. (1993) Priya and Shibasaki (2001)
Result from geographical area	Prediction for current season	Actual past practices, future decision rules	Yes	
	Diagnosis	Actual past practices	Yes	Gomez and Ledoux (2002)
	Scenario testing	Decision rules or hypothetical	Yes	Beaujouan et al. (2001)

2. Physical input data

2.1. Weather data

2.1.1. Available data

The required data for crop models typically include precipitation, temperature (minimum and maximum), potential evapotranspiration (or the parameters necessary to compute it) and solar radiation. These data are measured at specific locations (meteorological stations) not at the location of every field. For example, the density of the French national meteorological network corresponds to 1 station for 500 km^2 on the average. There is in addition a rainfall network that provides data with a delay of a month or more, with a density of 1 station per 100 km^2.

Obtaining meteorological data at locations other than weather stations is a problem that is of importance in many ways, not only for crop models. Two main approaches are used, namely zoning and interpolation.

2.1.2. Defining zones

The zoning approach involves dividing a region into zones considered homogeneous for climate. The same weather data are then used for all locations within a given zone. The weather data are generally those of a meteorological station included in the zone and considered as a representative of the zone. The definitions of the zones are determined just once and then are maintained over time.

An example of the zoning method, based on multivariate statistical analysis, is given in Ripert et al. (1990). More recently, the French Meteorological Services has defined the climatic zones of France (Fig. 1). These zones are based on the expertise of local meteorologists.

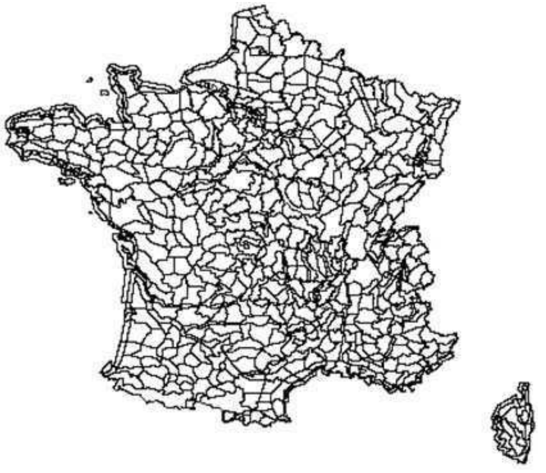


Figure 1. The division of France into homogeneous climatic areas.

2.1.3. Interpolating weather data

In the zoning approach, each location within the zone has the same climate data, with an abrupt change at the zone boundaries. An alternative approach is to interpolate climatic data between weather stations, so that the climatic variables are smooth functions over space. Creutin and Obled (1982) present a review of various interpolation methods including nearest neighbor, arithmetic mean, spline functions, optimal interpolation, kriging or an interpolation method based on empirical orthogonal functions.

Interpolation is generally done separately for precipitation and temperature. Furthermore, the calculations must be redone every day if daily data is used. Although such calculations are very time consuming, they are used by the National French Meteorological services.

An example of an interpolation technique is the Aurelhy method (Bénichou and Le Breton, 1987), which is considered the reference method for interpolating precipitation over France. The first step is to do a principal component analysis (PCA) to identify the major factors that describe topographical variability. In the second step, precipitation is regressed on the first components of the PCA.

There has been some effort to use other information to improve the estimation of the spatial variability of precipitation. If weather radar information is available, it can be used to provide information about precipitation at all locations, though the accuracy of the information may sometimes be a problem: the precision may vary among studied areas due to the distance to the radar, echo effects or topography problems. There have also been studies using satellite thermal infrared images. One study showed that surface temperature is related to precipitation (Seguin et al., 1989). Another found a relation between temperature at the cloud surface and the duration of a cold cloud responsible for precipitation (Laurent et al., 1998). However, these relations seem to apply better to West Africa than to temperate countries where the relations between precipitation and clouds are more complex.

In temperate countries, there is more reliance on 3D numerical weather prediction (NWP) than on remote sensing to aid in interpolation. An example is the SAFRAN (système d'Analyse Fournissant des Renseignements Atmosphériques à la Neige) approach, which bases interpolation on NWP modeling combined with observations. SAFRAN is a meteorological application of objective analysis (Brun et al., 1992) that was developed to initialize the French operational snow model CROCUS (Brun et al., 1989, 1992) to forecast avalanches in the Alps. It is based on the optimal interpolation technique (described in Durand et al., 1993) which combines observations and large-scale analysis provided by an NWP model, to analyze air temperature, relative humidity and wind near the surface, precipitation, total cloudiness and incoming radiation (shortwave and long wave). The system has been applied to the whole of France (Le Moigne, 2002) to provide input data for the Interface Soil Biosphere Atmosphere (ISBA) land surface model of Météo-France (Noilhan and Planton, 1989).

2.1.4. Predicting future weather

The prediction of near- or medium term weather for specific locations is a major goal of meteorological services. We will not consider this problem here.

In crop models, the more common problem is to make predictions not for a specific year but on the average over the different possible meteorological conditions at a site. Two main approaches are used, namely (i) using past data directly, (ii) using a weather generator based on past data.

A common approach when one has past data, for say n years is to run the model for the future n times using each weather record in turn and to assume that each result is equally likely. The result is n different model results, each assumed to have equal probability.

A simpler approach is to identify an "average" past climate year and use that for future weather. This simplifies the calculations (one runs the model for only a single weather series), but is unrealistic in that weather uncertainty is replaced by an average value.

In some cases one is not interested in average future results, but rather in results for specific conditions. For example, a water manager might want to make predictions in order to see whether water storage is sufficient for worst case conditions. In this case, one could use for example only 10% of the driest years from the past for prediction.

Finally, an expanding use of crop models is to evaluate the consequences of climate change. In this case, one does not assume that future climate will be similar to past climate. One can still use past climate series to represent future weather, but now one adds specific changes to the data. For example, to imitate global warming one could simply increase all temperatures by say 2°.

An alternative to the direct use of past climate data is to use a weather generator. This is simply an algorithm that generates the values of climate variables according to some probability distribution. A major effort required here is to create the weather generator. Consider for example only the generation of solar radiation. A very simple approach would be to divide the year into 10 day periods and for each period identify the minimum and maximum values in the past records. Then, the generator could generate solar radiation values for each day from a uniform distribution with the given minimum and maximum values. In practice, the generators also take into account the correlations between different variables. For example, the smaller values of solar radiation are usually associated with lower temperature, and rainfall events are usually associated with relatively low solar radiation. It is important, in developing or choosing a weather generator, to make sure that it reproduces the aspects of major importance. For example, many weather generators are based on the probability of rainfall events, and generate rainfall days at random using that probability. This need not necessarily give good agreement with other aspects of the weather record. For example, it may not give good agreement with the distribution of the lengths of periods between rainfall events. If the lengths of dry periods are of special concern, then the generator should probably be built explicitly for this purpose.

It requires substantial amount of past data to build a reliable weather generator, so it should not be imagined that a weather generator is a solution to the problem of insufficient data. Rather, its usefulness is that it allows one to transform a finite sample into an infinite number of different climate scenarios. Application of a weather generator for a region poses another crucial problem: correlations between adjacent fields exist and should not be ignored in the simulations of weather series. Interpolating in space the parameters of the weather generator in order to obtain a weather generator adapted to each field, or generating weather scenarios only for locations with past data and interpolating them, do not solve this problem.

Weather generators produce climates with properties similar to past climate. There is also interest in scenarios representing global climate change. Climate scenarios can be defined by arbitrary changes in temperature and precipitation, or on the basis of the output from general circulation models (GCMs). Such scenarios have been used with crop models to determine impacts on agriculture (e.g. Adams et al., 1990; Rozenweig, 1990). Barrow (1993) proposed two methods for constructing climate change scenarios and furnished a series of scenarios. These were used to investigate the effects of climate change on the development, yield and distribution of a variety of crops throughout Europe using crop growth models (e.g. Bindi et al., 1993; Semenov et al., 1993; Wolf, 1993). A similar approach has been used with hydrological models. For example, Etchevers et al. (2002) studied the impact of climate change on the Rhone river watershed. To estimate the climate 60 years in the future they used the climate general circulation model ARPEGE, but with modified air temperature and precipitation amounts. In another study, Noilhan et al. (2002) generated climate scenarios using global atmospheric climate models (GCMs) with the assumption of a doubling of atmospheric CO_2 concentration.

2.2. Soil properties

2.2.1. Required information

The required soil data for crop models typically include soil depth and soil physical properties, such as bulk density, soil water content at field capacity and wilting point for the whole soil profile or, if soil layers are identified, for each of them. Some models require the available water capacity of the soil directly, which is a combination of these properties. These data can be measured *in situ* or at the laboratory on soil samples. It is clear that it is impossible to sufficiently sample any area, whatever its size (but *a fortiori* large areas), to account for all spatial variability. Therefore, a spatial estimation method is necessary. As for weather data, two main approaches are used, namely zoning (i.e. soil mapping) and interpolation.

2.2.2. Defining zones

Soil surveys are the most common basis for estimating the spatial distribution of soil properties over an area. Furthermore, since soil properties are considered as stable over time, even old soil surveys can be used.

Standard soil survey procedure is to classify soils according to appearance and measured attributes, to define the geographic zone of each class, and to describe in detail for each class representative profiles from one or more sites (e.g. Soil Survey Staff, 1951; Boulaine, 1980; Bouma et al., 1980; Brus et al., 1992). Either implicitly or explicitly the properties and behavior at these "representative" sites are assumed to apply approximately to the whole area of the class. In general, there is at most one representative profile per soil unit. When representative profiles are not identified, one could choose sites within each soil unit at random (stratified sampling).

The capability of soil maps to provide soil mechanical properties was first investigated in the 1960s by engineers (Morse and Thornburn, 1961; Kantey and Williams, 1962; Thornburn et al., 1966; Webster and Beckett, 1968). Even though the maps that Thornburn and his colleagues evaluated had been made for agricultural purposes rather than engineering, the information provided about mechanical properties was deemed useful. However Webster and Beckett (1968) showed that the maps were not useful for predicting soil chemical properties. Beckett and Webster (1971) suggested that, in general, if the criteria for classification are not the properties that one wants to predict or not closely related to them, then any success in predicting those properties will be fortuitous.

Leenhardt et al. (1994) showed that the mean squared error in predicting soil water properties from soil survey information is a sum of terms related to the accuracy of soil stratification and the choice of representative sites. They found that the scale of the soil survey is a key factor in determining accuracy. Maps at the scales 1/10 000 and 1/25 000, where the criteria for classification were intrinsic soil properties, gave good results. The soil survey at a scale of 1/100 000 performed poorly, partly because it was based mainly on variables not directly related to the soil properties of interest. Choosing representative profiles and predicting from them was important only when the initial classification performed poorly. Nevertheless, as it is impossible to know in advance whether a soil map categorizes a given property well, choosing representative profiles provides information. It could however be replaced without loss of precision by stratified random sampling, where the classification proved effective. Finally, Leenhardt et al. (1994) found that the usefulness of soil surveys is much greater for spatial estimation of soil properties that are used during the soil survey, e.g. particle size distribution, or for estimation of properties that are strongly correlated with the latter, e.g. the different points of the water retention curve. On the other hand, the accuracy of soil surveys for predicting soil layer thickness, and therefore available water capacity, is much lower. Indeed, it seems that, though the succession of horizons in the profile is important for classifying the soil, the variation in thickness of the horizons is not taken into account by soil surveyors.

2.2.3. Interpolation

Because of the drawbacks of soil maps, discussed by Beckett and Webster (1971), attention has switched from classification to geostatistical prediction to take into account any spatial dependence within classes and the gradual nature of change of soil across mapped boundaries. In 1990, Voltz and Webster showed that standard kriging techniques are unsuited to cases where the soil changes abruptly, and in these circumstances soil classification outperforms it. Finally, where land management requires estimation of several soil properties over the study area the task is multivariate, and soil classification appears to be easier to comprehend than multivariate geostatistics. Consequently, the classical soil map approach to prediction is likely to remain of value in the right circumstances and environment for a long time to come.

2.2.4. Obtaining non-measured soil characteristics

Soil maps can be the basis for obtaining soil properties at each location, but in general the properties recorded (usually soil type and soil texture) are not those needed for crop models (for example, soil depth and available water capacity or water retention curves and hydraulic conductivity). A common solution to this problem is to use "*pedotransfer functions*" (PTFs), which relate basic soil properties that are considered as easily accessible to less often measured soil properties (Bouma, 1989; van Genuchten and Leij, 1992) (Table 2). Most PTFs are "*continuous-pedotransfer functions*" (continuous-PTFs) which predict hydraulic properties as continuous functions of more commonly measured soil properties. Other PTFs called "*class-pedotransfer function*" (class-PTF) are functions which predict hydraulic properties from the soil class (very often texture class). Finally, a "*pedotransfer rule*" can be also used. This is a relationship based on expert opinion between soil composition and the predicted property (Daroussin and King, 1997). For a recent review of research in this area, see the review by Wösten et al. (2001).

Pedotransfer functions are derived using databases, which contain both the input data (readily available soil characteristics) as well as the output data (soil hydraulic properties). Several large databases such as the USDA Natural Resource Conservation Service pedon database (USDA Natural Resource Conservation Service, 1994), WISE (Batjes, 1996), UNSODA (Leij et al., 1996, 1999) and HYPRES (Lilly, 1997; Lilly et al., 1999; Wösten et al., 1999) and much smaller databases (Wösten et al., 2001) have been used for development of PTFs. PTFs have been developed for the United States (Rawls, 2004), Europe (Wösten and Nemes, 2004) and tropical soils (Tomasella and Hodnett, 2004). Since water retention at different water potentials is much easier to measure than hydraulic conductivity, the number of soils with measured water retention properties in databases is considerably greater than the number of soils with measured hydraulic conductivity. As an example, in the European database HYPRES, there are 1136 soil horizons with both water retention and hydraulic conductivity and 2894 soil horizons with only water retention (Wösten et al., 1999). A result is that PTFs developed for water retention properties are much more numerous than those that predict hydraulic conductivity (Bastet et al., 1998).

Early PTFs predicted the water retention properties by predicting individual points of the water retention curve. Among these PTFs, those of Renger (1971), Gupta and Larson (1979), Rawls et al. (1982) are continuous-PTFs and those of Hall et al. (1977) and Bruand et al. (2002, 2003) are class-PTFs. Other PTFs assume that all the water retention curves have the same mathematical form so that the PTFs need only predict the parameters of that model. Among these PTFs, those of Cosby et al. (1984), and Vereecken et al. (1989) are continuous-PTFs while those of Wösten et al. (2001) are class-PTFs. Most studies during the last decade were concerned with second type of PTF because it provides a mathematical model directly for the entire water retention curve (Rawls et al., 1992; Minasny et al., 1999; Wösten et al., 2001). Despite their possible inaccuracies, class-PTFs which predict individual points of the water retention curve are easy to use because most require little soil information and are well adapted to prediction of water retention over large areas (Wösten et al., 1995; Lilly et al., 1999; Wösten et al., 1999; Bruand et al., 2003). Although they only give water retention at certain potentials, it is easy to fit a mathematical model to these predictions and thus to obtain water retention as a continuous function of water potential (Table 3).

Table 2. Input variables for PTFs for prediction of water retention at different potentials.

PTFs		Input variables at potential (hPa) of				
		−100	−330	−1000	−3300	−15 000
Renger (1971)			Clay			Clay
			Silt			Silt
Hall et al. (1977)	Top- and sub-soil	Clay				Clay
		Silt				
		OC				
		ρ_b				
Gupta and Larson (1979)		Sand	Sand	Sand		Sand
		Silt	Silt	Silt		Silt
		Clay	Clay	Clay		Clay
		OM	OM	OM		OM
		ρ_b	ρ_b	ρ_b		ρ_b
Rawls et al. (1982)	Model 1	Clay	Clay	Clay		Clay
				Silt		
		Sand	Sand			
		OM	OM	OM		OM
	Model 2	Sand	Sand	Sand		
		OM	OM	OM		
		$\theta_{15\,000}$	$\theta_{15\,000}$	$\theta_{15\,000}$		
	Model 3	Sand				
		OM		OM		
		θ_{330}		θ_{330}		
		$\theta_{15\,000}$		$\theta_{15\,000}$		
Cosby et al. (1984)		Clay	Clay	Clay	Clay	Clay
		Silt	Silt	Silt	Silt	Silt
		Sand	Sand	Sand	Sand	Sand
Vereecken et al. (1989)		Clay	Clay	Clay	Clay	Clay
		Sand	Sand	Sand	Sand	Sand
		OC	OC	OC	OC	OC
		ρ_b	ρ_b	ρ_b	ρ_b	ρ_b
Bruand et al. (1996)		ρ_b	ρ_b	ρ_b	ρ_b	ρ_b

Clay – clay content; Silt – silt content; Sand – sand content; OC – organic carbon content; OM – organic matter content; ρ_b – bulk density; θ_{330} and $\theta_{15\,000}$; volumetric water content at −330 and −15 000 hPa, respectively.

Table 3. Volumetric water contents at different water potentials using the non-continuous class-PTFs based on texture (FAO triangle) and bulk density and parameters of the van Genuchten's (1980) model (Bruand et al., 2003).

Texture class	Class of D_b^c	D_h^b	Volumetric water content $\theta_{\log(-h)}$ cm^3 cm^{-3}							Parameters of van Genuchten's model				
			$\theta_{1.0}$	$\theta_{1.5}$	$\theta_{2.0}$	$\theta_{2.5}$	$\theta_{3.0}$	$\theta_{3.5}$	$\theta_{4.2}$	θ_s	θ_r	n	α	R^2
Very fine	]1.2–1.3]	1.25	0.531	0.514	0.490	0.465	0.428	0.418	0.329	0.527	0.0100	1.0849	0.0098	0.964
		1.15	0.484	0.473	0.451	0.428	0.393	0.384	0.303	0.481	0.0001	1.0868	0.0083	0.966
	]1.3–1.4]	1.35	0.493	0.486	0.467	0.447	0.416	0.401	0.321	0.488	0.0002	1.0930	0.0042	0.971
		1.25	0.456	0.450	0.433	0.414	0.385	0.371	0.298	0.452	0.0006	1.0923	0.0043	0.973
	]1.4–1.5]	1.45	0.489	0.477	0.464	0.445	0.422	0.386	0.318	0.481	0.0001	1.1055	0.0028	0.987
		1.35	0.455	0.444	0.432	0.415	0.393	0.359	0.296	0.448	0.0001	1.1066	0.0027	0.988
Fine	]1.3–1.4]	1.35	0.459	0.429	0.419	0.390	0.369	0.332	0.270	0.449	0.0007	1.0975	0.0088	0.977
		1.25	0.425	0.398	0.388	0.361	0.341	0.325	0.250	0.415	0.0010	1.0927	0.0086	0.952
	]1.4–1.5]	1.45	0.441	0.422	0.400	0.381	0.348	0.323	0.274	0.441	0.0002	1.0802	0.0194	0.992
		1.35	0.410	0.393	0.373	0.355	0.324	0.301	0.255	0.410	0.0007	1.0811	0.0180	0.993
	]1.5–1.6]	1.55	0.383	0.378	0.366	0.350	0.326	0.295	0.259	0.383	0.0006	1.0854	0.0062	0.999
		1.45	0.358	0.353	0.342	0.328	0.305	0.276	0.242	0.358	0.0001	1.0864	0.0059	0.999
	]1.6–1.7]	1.65	0.381	0.363	0.353	0.333	0.312	0.302	0.264	0.384	0.0003	1.0558	0.0377	0.986
		1.55	0.358	0.341	0.332	0.313	0.293	0.284	0.248	0.361	0.0002	1.0560	0.0367	0.986
	]1.7–1.8]	1.75	0.366	0.364	0.341	0.315	0.310	0.292	0.263	0.377	0.0005	1.0518	0.0560	0.981
		1.65	0.345	0.343	0.322	0.297	0.292	0.276	0.239	0.352	0.0001	1.0583	0.0333	0.974

Medium fine	]1.4–1.5]	1.45	0.381	0.365	0.348	0.313	0.264	0.220	0.193		0.377	0.1402	1.3325	0.0068	0.997
		1.35	0.355	0.340	0.324	0.292	0.246	0.205	0.180		0.352	0.1309	1.3332	0.0068	0.997
Medium	]1.5–1.6]	1.55	0.356	0.340	0.312	0.274	0.231	0.206	0.175		0.360	0.1125	1.2472	0.0170	0.999
		1.45	0.334	0.318	0.292	0.257	0.216	0.193	0.164		0.338	0.1036	1.2423	0.0176	0.999
	]1.6–1.7]	1.65	0.350	0.338	0.319	0.286	0.241	0.193	0.152		0.350	0.0120	1.1862	0.0078	0.999
		1.55	0.329	0.318	0.299	0.268	0.226	0.181	0.143		0.329	0.0088	1.1820	0.0082	0.999
	]1.7–1.8]	1.75	0.322	0.310	0.299	0.282	0.261	0.226	0.184		0.317	0.0002	1.1231	0.0049	0.992
		1.65	0.304	0.292	0.282	0.266	0.246	0.212	0.173		0.299	0.0005	1.1245	0.0048	0.992
	]1.8–1.9]	1.85	0.311	0.300	0.287	0.272	0.265	0.239	0.181		0.302	0.0003	1.1276	0.0026	0.959
		1.75	0.294	0.284	0.271	0.257	0.250	0.226	0.172		0.286	0.0009	1.1240	0.0028	0.959
Coarse	]1.6–1.7]	1.65	0.315	0.277	0.210	0.182	0.142	0.114	0.089		0.352	0.0334	1.2429	0.0843	0.996
		1.55	0.296	0.260	0.197	0.171	0.133	0.121	0.084		0.339	0.0328	1.2286	0.1123	0.993
	]1.7–1.8]	1.75	0.280	0.252	0.193	0.154	0.121	0.100	0.086		0.294	0.0695	1.4180	0.0339	0.999
		1.65	0.264	0.238	0.193	0.154	0.100	0.094	0.081		0.272	0.0711	1.5179	0.0257	0.996
	]1.8–1.9]	1.85	0.303	0.281	0.257	0.226	0.183	0.165	0.128		0.310	0.0008	1.1434	0.0304	0.996
		1.75	0.287	0.266	0.243	0.214	0.173	0.156	0.121		0.294	0.0008	1.1435	0.0307	0.996

D_b^c – bulk density measured on clods of around 1 cm^3; D_b^h – bulk density of the horizon inferred from D_b^c.

The accuracy of PTFs has been discussed in several studies (e.g. Tietje and Tapkenhinrichs, 1993; Kern, 1995; Wösten et al., 1995; Bastet et al., 1999; Schaap, 2004). A common measure of accuracy is root-mean square error (*RMSE*) defined as:

$$RMSE = \sqrt{\sum \left(\theta_{\mathrm{m}} - \theta_{\mathrm{p}}\right)^2 / n},$$

where θ_{m} and θ_{p} are, respectively measured and predicted volumetric water contents and n the total number of observations. Analysis of the literature showed that *RMSE* varied from 0.02 to 0.11 $\mathrm{m}^3\,\mathrm{m}^{-3}$. The smallest *RMSE* values were obtained in studies where either a preliminary grouping of soils was applied or one or more measured points of the water retention curve were used as predictors (Wösten et al., 2001; Bruand, 2004). The largest *RMSE* of 0.11 $\mathrm{m}^3\,\mathrm{m}^{-3}$ was obtained in a study where the soil texture was used as a sole predictor (Table 4).

PTFs have also been used to predict saturated hydraulic conductivity and more recently unsaturated hydraulic conductivity (Nemes and Rawls, 2004). The accuracy of several of these PTFs was evaluated by Tietje and Hennings (1996) on a set of 1161 soils from Lower Saxony in Germany. In fact saturated hydraulic conductivity K_{s} is closely related to the characteristics (size, shape, connectivity, tortuosity) of macropores in the soil that result from biological activity and from tillage practices. The PTFs studied by Tietje and Hennings (1996) on the other hand, are based on soil characteristics such as particle size distribution or organic matter content, which are related to the total porosity but are only distantly related to the presence of macropores. This explains the poor accuracy of prediction found for the PTFs.

Other PTFs are based on the concept of "effective porosity" which in most studies refers to the air-filled porosity at -330 hPa (Ahuja et al., 1989). These PTFs relate K_{s} to effective porosity (ϕ_{e}) by the equation

$$K_{\mathrm{s}} = a(\phi_{\mathrm{e}})^b,$$

where a and b are two parameters. The validity of these PTFs was discussed by Franzmeier (1991) and Tomasella and Hodnett (1997). They showed that a and b are not in fact constant, but rather vary according to the characteristics of the soil studied. The prediction of the unsaturated hydraulic conductivity is still very difficult today, progress being limited by the small number of available data.

The rapidly increasing demand for PTFs in the last decade has led to the utilization of available databases that are not adequate for the purpose. We recommend a wiser utilization of PTFs. There is still a need for acquiring measured hydraulic properties to enrich the databases. The new measured hydraulic properties will enable the improvement of available PTFs and the development of innovative new PTFs. The lack of data is particularly appreciable for the unsaturated hydraulic conductivity in the range of water potential between 0 and –50 hPa, i.e. close to saturation. Indeed, within this range of water potential, the unsaturated hydraulic conductivity varies over several orders of magnitude.

Table 4. Accuracy of water retention PTFs (modified after Wösten et al., 2001).

Source	Water potential (hPa)	RMSE ($m^3\ m^{-3}$)	PTF input variables
Ahuja et al. (1985)	−330	0.05	Clay, silt, organic matter content, bulk density
	−15 000	0.05	
Bruand et al. (1996)	−330	0.03	Bulk density
	−15 000	0.03	
Gupta and Larson (1979)	−15 000	0.05	Clay, silt, organic matter content, bulk density
Koekkoek and Bootlink (1999)	−100	0.05	Clay, silt, sand, organic matter content, bulk density
	−15 000	0.05	
Lenhardt (1984)	−330	0.07	Clay
	−15 000	0.05	
Minasny et al. (1999)	−330	0.07	Clay, silt, sand, bulk density, porosity, mean particle diameter, geometric standard deviation
	−15 000	0.07	
Pachepsky et al. (1996)	−330	0.02	Clay, silt, sand, bulk density
		0.02	
	−15 000		
Paydar and Cresswell (1996)	A[a]	0.02	Slope of the particle size distribution curve + one measured point on the water retention curve (WRC)
Paydar and Cresswell (1996)	A	0.03	Clay, silt, coarse sand, fine sand, organic matter content
Schapp et al. (1998)	A	0.11	Texture class only
Schapp and Leij (1998)	A	0.10	Clay, silt, sand
Sinowski et al. (1997)	−300	0.04	Clay, silt, sand, bulk density, porosity, median particle diameter and standard deviation
	−15 000	0.04	
Tomasella and Hodnett (1998)	A	0.06	Clay, silt, sand

[a] A – average RMSE along the measured water retention curve obtained after estimating parameters of a water retention equation and using this equation to compute water contents at all potentials where the water retention was measured.

2.3. Initial conditions

The main initial conditions required for crop models are soil moisture and nitrogen in each soil layer. Often this information is not available for the past and so the initial values have to be estimated. The same of course is true for future or hypothetical cases.

A common approach for initial water is to assume that water at sowing is some fixed percentage of maximum available water, for example 80% for all layers. A modification that may be more realistic for some environments is to fix initial conditions not at sowing but several months before. For example, consider initial water conditions at the time corn is sown (around April) in southwestern France. In this region, there are usually one or more rainfall events during winter that completely fill the soil profile. It is sufficient then to start simulations at the beginning of winter with a very rough estimate of initial soil water. At some point before sowing the profile will be filled, both in reality and according to the model, and this provides the correct initial condition for the subsequent calculations.

Initial soil nitrogen at sowing may also be difficult to obtain. Here again it may sometimes be useful to start simulations some time before sowing. In France, for example, there are tables for calculating soil nitrogen at the end of winter as a function of soil type, the previous crop species, its yield and nitrogen applications to the previous crop.

3. Management practices

3.1. Required information

The management decisions that must be specified for each field include the crop species, variety, sowing date, fertilization dates and rates, irrigation dates and amounts, etc. These inputs are particularly difficult to obtain for fields where they have not been observed because they depend on individual farmer decisions rather than on physical properties, and these may not vary at all smoothly with location. Thus, mathematical interpolation may not be a reasonable approach. One possible approach is to use a unique set of management decisions, for example recommended practices, for an entire region (Hansen and Jones, 2000; Yun, 2002). However, ignoring the spatial variability of practices can lead to prediction errors (Yun, 2002). In the following sections, we present approaches that account for the spatial variability of crop species and agricultural practices within a region.

3.2. Crop species

The information as to which crop species were planted in previous years is usually easily obtained for a single field or a small number of fields. However, this is no longer the case when running a crop model over a region or an entire country.

3.2.1. Determining the crop planted in every field

When a model is used as a diagnostic tool to analyze past results, it is necessary to know past land use. For the recent past, the two main sources of information are statistical sampling data and remote sensing data from the period in question. Obtaining information about the more distant past is more difficult. One can use remote sensing data if available or use a combination of statistical data and surveys.

3.2.1.1. Recent past

One important source of information is agricultural statistics. The countries of the European Union use several different procedures (Gallego, 1995). "Local statistics" are based on an administrative unit (e.g. municipality, small agricultural region, parish). These data are collected by local administrators or by agricultural organizations. These data are available every year but may be quite imprecise. "Farm census" data result from contacting every farm. The area devoted to each crop is just one of the pieces of information acquired. Only in small countries is the census done annually, otherwise it is usually redone every five or ten years. "Sampling surveys" contact only a sample of farms, and then use statistical techniques to extrapolate to the entire population. In the United States, the National Agricultural Statistics Service (NASS) has been using area frames for agricultural surveys since the early 1960s (Cotter and Nealon, 1987).

All the above methods provide an estimation of the crop acreage over a region, but do not indicate specifically which fields were planted with which crops. To overcome this, new approaches have been developed, based on technologies such as remote sensing (Campbell, 2002; Chuvieco, 2002; Lillesand et al., 2003) and Geographic Information Systems (Burroughs and McDonnell, 1998).

Remote sensing allows one to directly identify the spatial distribution of crop species, and also to monitor the development and state of the vegetation. The capability of remote sensing to provide precise and useful land use maps depends of various factors: cover type, crop development, cloudiness conditions, dates of the images, method of analysis. Although one image taken when the crop is fully developed may be sufficient, generally several images during the season are required to provide reliable identification of crop species. Furthermore, though automatic classification is possible, supervised classification usually gives better results.

An example of combining remote sensing and statistical data is the MARS (Monitoring Agriculture with Remote Sensing) project conducted in Europe, since 1988 (Gallego, 1995). Another one is the AgRISTARS program (Agricultural and Resources Inventory Surveys Through Aerospace Remote Sensing) (Pinter et al., 2003). Satellite data and ground data are combined by means of a regression estimator.

3.2.1.2. More distant past

For the more distant past, remote sensing shots may provide valuable information about the evolution of land use cover of a region (Ducrot and Gouaux, 2004). Unfortunately the necessary past images, which must be sufficient in number, taken at appropriate periods and have sufficient resolution, are often not available. Therefore, other techniques are required. Mignolet et al. (2004) propose a method for reconstituting the evolution over time of cropping systems and their spatial distribution. This process makes use of both expert opinion and regional or national agricultural statistics that are compared, step by step, in order to build the most reliable database for the time and space scales considered. Data mining and statistical cartography methods proposed by Mari et al. (2000) are used to identify the major crop rotations of a region, their evolution in the past years and their spatial distribution.

3.2.2. Predicting which crops will be planted

The prediction of crop species depends on the time at which prediction is required. If for example, a prediction of national yield is required shortly before harvest time, then the agricultural statistics for the current year or remote sensing data may be available, and the approaches described above are applicable.

If on the other hand early prediction is required, then different procedures are required. For example, the study of Leenhardt et al. (2004a) concerns a water manager in southwestern France who requires predictions of future water use starting in early summer. At this time statistical survey information is not yet available and satellite imagery cannot yet distinguish between the various summer crops, and particularly between corn, which is irrigated, and sunflower, which is not irrigated in southwestern France.

One possible approach in this case is simply to assume that at a regional scale the change in land use from one year to another is negligible. Such an assumption would be reasonable for a region where single-crop farming dominates and no major changes in economic or regulatory factors have occurred.

A second possibility is to use declared intentions of farmers, where such information is available. The European Agricultural Policy involves asking farmers to declare which crops they intend to cultivate in each field. A minor problem here is that climatic conditions may lead to some changes in plan. A major difficulty is obtaining this information, which is protected by privacy laws. The information is made available in the form of a computer database, but this only concerns data aggregated by district and furthermore there is considerable delay before this is done.

A third possibility is explored by Leenhardt et al. (2005). This approach has two stages. First, one obtains an estimate of land use in the preceding year. Then, one uses information about crop rotations to give the probability of having various crops in a field this year, given the crop last year. The crop rotation information is based on Ter-Uti, which is a systematic land-use sampling program in France. This sampling system is being extended to the entire European Community under the name of LUCAS (Land Use/Cover Area frame statistical Survey) (SCEES, 2005). The same locations are sampled each year. The approach proposed by Mari et al. (2000), allowing the identification of the major crop rotations of a region (see Section 3.2.1), might be worthwhile combined with early remote sensing information, which can not identify the exact crop but which can limit the possibilities to a group of crops.

In many cases, one wants to study scenarios that imply a change in choice of crop. For example, the scenarios might concern changes in climate or in the economic or regulatory context. A number of possible approaches exist (see for example, Veldkamp and Lambin, 2001). One approach is to assume that each farmer maximizes some objective function, for example net profit, subject to constraints (for example, available labor). The problem is then to determine which crop species (and perhaps management decisions) this implies. A simple approach is to suppose that farmers have a choice between a limited number of systems (crops, management), each of which is associated with a certain yield and a certain use of resources. Linear programming can then be used to find the optimal crop and management.

3.3. Sowing date

For past data one could simply seek to obtain the sowing date for each field, but this can be very difficult for large numbers of fields. Even if one is willing to address direct inquiries to each farmer many may not respond. For example, Leenhardt and Lemaire (2002) had a 39% answer rate to a postal survey, and this is considered a very good return rate.

Information that is generally available is a recommended sowing period for each crop, each variety and each region. One also has in general climate information and statistical information about farm structure and land use.

3.3.1. Estimating or predicting sowing date

Sowing dates could be based on the recommendations that exist for each variety in each region, but within the possible sowing period the actual sowing date will depend on available manpower, the state of the soil and climatic conditions. This suggests two possible approaches, either using a fixed average sowing date or calculating a sowing date for each field based on information about farm organization and climate. An example of calculation of sowing date is the SIMSEM model of sowing date proposed by Leenhardt and Lemaire (2002).

3.3.1.1. Overview of SIMSEM

The model SIMSEM is based on a water balance model, a farm typology and expert knowledge. The farm typology used (SICOMORE[1]) defines, for each small agricultural region (SAR[2]), a limited number of farm types (around 6). Each type is characterized by the area of the various soil–crop associations per farm and the availability of time and equipment. Table 5 presents the main characteristics of a sample of farm types. SIMSEM calculates, for each SAR, the distribution of sowing dates of each summer crop following a three-step procedure: (i) the determination of possible sowing days using a soil water model, (ii) the determination of the time required to sow each crop, and (iii) the determination of the days on which each crop is sown.

3.3.1.2. Determining possible sowing days using a soil water model

A water balance model is run at a daily time step over the months of the sowing period to determine, for each soil type, which days are possible sowing days. To determine if sowing is possible, a decision rule based on soil water status and precipitation is used. The rule is: "If the soil water content (SWC) is below $x\%$ of the soil available water capacity (SAWC), and if it does not rain more than y mm this day, then sowing can occur." Threshold values x and y were obtained, for the study region, after analysis of past sowing dates.

---

[1] SICOMORE (SImulation éCOnomique de MOdèles REprésentatifs de l'agriculture) (CRAMP, 1988) is a method used by the regional agricultural extension service of the Midi-Pyrénées region of France to identify farm types in the region. The farm typology, which is an output of SICOMORE, is obtained from a minimum of standard data from statistical, administrative and professional surveys.

[2] The SAR are the sub-divisions of the corresponding agricultural regions (ARs) that respect the administrative limits of the French Departments. The ARs, on the other hand, are based on physical and economic criteria. They are homogeneous agricultural areas that date from the 1940s (République Française, 1946a,b, 1949).

Table 5. Soil–summer crop associations, manpower and number of cows of the farm types with irrigation present in a small agricultural region (SAR) (Haut-Armagnac), as described by SICOMORE. Among the six farm types are two types of mixed crop–livestock farms (CLF) and four types of field crop farms (FC).

Farm type	Number of farms	Area (ha) of different crop–soil associations				Man-power (equivalent full-time workers)	Number of cows
		Maize × loam	Sunflower × clay	Soybean × clay	Sorghum × Clay		
CLF1	150	10.1	4.9			2.4	22
CLF2	112	10.4				1.8	10
FC1	181	66.3	29.8	2.6		2.2	
FC2	58	29	25.9	2.4		2.2	8
FC3	105	3			5.1	2.4	
FC4	423				4.7	1.5	7

3.3.1.3. Determining the time required to sow each crop

The second step of the SIMSEM procedure is primarily based on the information given by the farm typology: the type and area of various crop–soil associations for each farm type, the kind and size of its livestock, and the amount of manpower available. However, complementary information (and very specific to the region considered) was provided by experts from local technical institutes: the earliest possible date for sowing the various summer crops, the priority between crops for sowing, the time necessary to sow for various soil types, and estimations of daily working time. For each farm type and for each crop the following outputs were calculated:

- The time T_{cs}^{f} (in days) necessary, in farm type *f*, to sow crop *c* on soil *s*. This is a function of the area A_{cs}^{f} of crop *c* (in hectares) on soil *s* represented in farm type *f*, the time t_s (in hours/hectare) necessary to sow on soil *s* and the amount of time per day devoted to sowing t_{d} (in hours/day):

  $$T_{cs}^{f} = A_{cs}^{f} t_s / t_{\mathrm{d}} \tag{1}$$

  The time devoted to sowing is calculated as

  $$t_{\mathrm{d}} = (L^f t_{\mathrm{d}}^{\max}) - (t^{\mathrm{a}} N^{\mathrm{a}}) \tag{2}$$

  where L^f is the manpower available for farm type *f* (equivalent number of full time workers), $t_{\mathrm{d}}^{\max}$ is the maximum working time per full time worker per day (hours), t^{a} is the time (in hours/animal/day) necessary to take care of the animals and N^{a} is the number of animals on the farm. (In this area, only cows are considered since other kinds of livestock are not time demanding at the sowing period of summer crops.)

- The area A^r_{cs} (in hectares) of this crop–soil association to be sown for the small agricultural region r is

$$A^r_{cs} = A^f_{cs} N^r \tag{3}$$

where N^r is the number of farms of type f in the region r.

- The area a^r_{cs} (in hectares/day) that can be sown in one day in the region r is then

$$a^r_{cs} = A^r_{cs} / T^f_{cs} \tag{4}$$

The various inputs used in the above equations come either from the SICOMORE typology (A^f_{cs}, L^f, N^{a}, N^r) or from experts (t_s, $t_{\mathrm{d}}^{\max}$, t^{a}).

3.3.1.4. Combining possible sowing days and required time to sow

Days suitable for sowing (defined in step one) are taken in chronological order. Each crop–soil combination has a priority. Sowing begins with the highest priority combination. Each day, a^r_{cs} hectares are sown, until the total area A^r_{cs} is sown. Then the same procedure is repeated for the crop–soil association with the next highest priority, etc. The order of priorities may vary between regions depending on economic criteria (relative contribution of crop to farm profit), physiological criteria (heat needs of crop for emergence) or physical criteria (capacity of soils to store heat). The result of the simulation is a temporal distribution of sowing dates for each summer crop in each small agricultural region. The estimated sowing dates are not field specific.

3.4. Irrigation

For irrigated crops, dates and amounts of irrigation are required inputs to the model. When numerous fields are concerned, such information can be very difficult to obtain. In general, each field has different irrigation dates and amounts. However, it may be reasonable to assume that this diversity results from a small number of underlying irrigation strategies applied to a diversity of situations. (Decision strategies are discussed in detail in Chapter 6. An irrigation strategy corresponds to a set of rules which relate irrigation decisions to various aspects of the situation such as soil characteristics, climate, state of crop development etc.) It is therefore more reasonable to try to characterize irrigation strategies than to characterize the diversity of irrigation dates and amounts over the study region. Furthermore, the irrigation strategy of a given farmer is likely to be relatively stable (at least over short or medium periods) while irrigation dates and amounts change each year. Simulating irrigation strategy appears then as a good solution when irrigation dates and amounts must be approximated for many fields. It is also a reasonable approach when one deals with future or hypothetical situations.

A number of studies have treated the representation of irrigation strategies. Chapters 6 and 19 of this book contain examples. A very common option encountered in the literature (e.g. Herrero and Casterad, 1999; Ray and Dadhwal, 2001) is to suppose that irrigation is applied as needed to satisfy crop water needs, as indicated in the FAO guidelines (Smith, 2000). This simple assumption allows one to calculate irrigation amounts without tackling the difficult problem of individual farmer behavior. Very few studies concern actual irrigation strategies over a region. For such studies one can use interviews with farmers, but often the irrigation strategy of a farmer is implicit and so cannot be obtained from a simple direct question. One asks questions concerning a farmer's practices and reasoning and then deduces the strategy. In general, however, farmers can only provide irrigation information about the recent past. Furthermore, the interviews generally are quite long (1–2 h). Given the number of farmers (for example, more than 500 farmers in a 500 km^2 area in southwestern France), sampling is necessary. Two solutions can then be applied to generalize the results over a region. The first consists in defining some average irrigation strategy that is assumed to apply to the entire population of farmers who irrigate (Leenhardt et al., 2004a). This average irrigation strategy may be adjusted by comparing simulated and observed water consumption data for the whole region. The second approach consists in building a typology of irrigation strategies. Then one would seek a relation between the choice of type of strategy and various explanatory variables (e.g. soil type, level of equipment, farm type) (Maton et al., 2005). Note that this would describe the main strategies but would not specify where exactly they are located. This would only be possible if the explanatory variables could be geo-referenced using soil maps, climate zones or administrative divisions.

4. Remote sensing data

Remote sensing is of great interest when one wants to use a model for multiple fields for past or current conditions, because it gives detailed spatially explicit information on soils and crops over an entire area.

The main informative spectral domain is that corresponding to solar radiation (400–1600 nm) which allows one to estimate various canopy properties such as canopy structure (leaf area index LAI or fraction of absorbed radiation FAPAR), leaf chlorophyll, leaf water content, etc. One way of obtaining estimates is by using a radiative transfer model. Such models express reflectance as a function of canopy properties. If reflectances are measured, inverting the model gives estimates of canopy properties (Baret et al., 2000). Alternatively, one can use empirical equations which relate canopy properties directly to functions of reflectances called vegetation indices. Satellite images are available in the solar domain at various spatial resolutions (IKONOS: 1 m, SPOT-HRV: 10 m, MERIS-MODIS: 300 m, SPOT-VGT: 1 km). The frequency of images from the same location goes from once a month to once a day and decreases as the area covered by an image decreases. Other spectral domains like thermal infrared or passive and active microwaves are related to other characteristics (crop biomass and water content, soil water content), but have lower spatial resolution.

We have already discussed the use of remote sensing to identify the crop in a field. In that case the spectral signature is exploited directly, being compared statistically to signatures for different crops. In this section we consider how remote sensing data, through their relation to canopy properties, can help in obtaining the input data necessary to run a crop model. The different ways of taking advantage of this external information use the fact that the biophysical variables that are available by remote sensing include key state variables simulated by crop models. Using these data to estimate values of model input variables or to modify values of model state variables is known as "data assimilation" (Delécolle et al., 1992; Pellenq and Boulet, 2004). See Chapter 5 for a detailed explanation of assimilation techniques.

4.1. Use of remote sensing data for estimating input variables

Figure 2 shows the general procedure. Suppose that the input variable sought is planting date. The basic idea is to find the planting date that leads to the best agreement between reflectance calculated using model predictions of soil and canopy properties and the observed reflectance (Bouman, 1992; Moulin, 1995; Guérif and Duke; 1998, Prévôt et al., 2003). As shown in the figure, this requires a radiative transfer model, which calculates the reflectance in the visible or near-infrared from soil and canopy properties.

The use of remote sensing data for obtaining sowing dates was explored by Moulin (1995) in a simulation study. She used the AFRC (Agricultural and Food Research Council)-Wheat model coupled to the SAIL radiative transfer model to generate artificial remote sensing observations for wheat canopies. She then used the procedure of Figure 2 to go backwards and determine sowing date from the remote sensing data. The study explored the impact of various factors on the quality of the sowing date estimate. The number and dates of images had a large influence on the shape of the cost function that is minimized in the assimilation procedure. The most favorable configurations corresponded to measurements made during the rapid canopy growth period. The precision of the remote sensing observations affected the precision of the sowing date estimate. A 5% error in the observed values led to an error of 2 days in sowing date (and a 4.8% error in

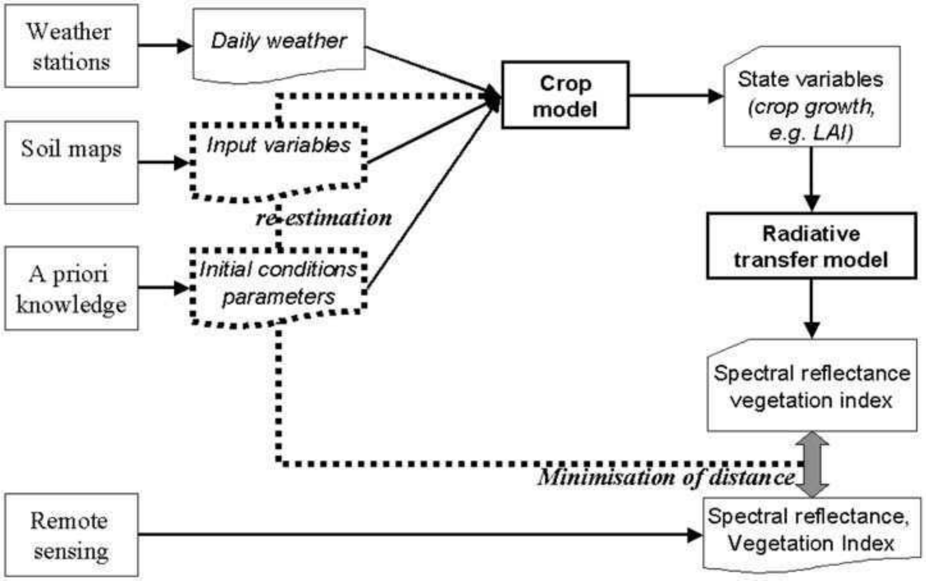


Figure 2. Data, actions and models involved in the assimilation of remote sensing data into a crop model.

final biomass). A 20% error in observed values led to an error of 20 days in sowing date and an error of 20.3% in final biomass.

She also applied the method to a wheat field in the French Beauce region, using SPOT satellite images at four dates in 1992. First, it was assumed that sowing date was unknown within the range from day of year (DOY) 266 to DOY 326. For this range of planting days, the relative error for final biomass prediction ranged from +17.6 to −9.8%. The re-estimation of the sowing date using the 4 SPOT-HRV data led to a sowing day of 299 with an error of +3 days. The associated relative error for biomass prediction was −2.8%. Figure 3 shows the results with and without re-estimation.

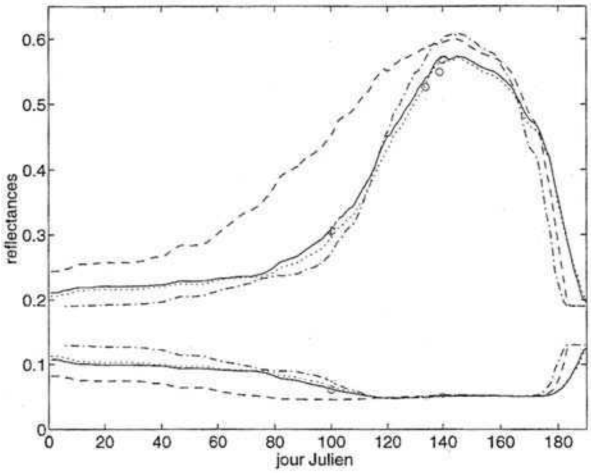


Figure 3. Reflectance profiles (in the near-infrared wavelength on the upper part of the graph, in the red wavelength on the lower part of the graph) simulated by the coupled AFRC-Wheat+SAIL model. Sowing days are DOY 266 (- - -), DOY 299 (-·-·-), the true sowing day (—) and sowing day estimated using remote sensing observations (. . .). The symbols (○) represent SPOT satellite measurements.

Remote sensing data can be used to estimate not only sowing date but also other crop model inputs. In fact, the only obligatory requirement is that the inputs must affect some quantity which can be related to reflectance measurements. However, in practice it may be impractical to estimate several different inputs. An example is provided by the study of Launay (2002), who used the SUCROS crop model coupled with the SAIL radiative transfer model to estimate both sowing date and LAI at 500°C day after sowing (LAI_{500}), for 31 sugar beet fields in Picardie in northern France. LAI_{500} is calculated from the initial relative growth rate of the leaves (Launay et al., 2005): the greater the LAI_{500}, the better the crop establishment in the critical period in spring. This is a major determinant of subsequent growth for crops like sugar beet, less capable than wheat of compensating for poor crop establishment.

This study showed that in the case considered, where few remote sensing data were available during the rapid crop growth period, it did not seem feasible to estimate both sowing date and LAI_{500}. Table 6 shows the errors in the estimated values with or without assimilation of the remote sensing data. When both sowing date and early LAI

Table 6. Root mean square errors (RMSEs) for 31 fields for sowing date and for early LAI when only early LAI or when both early LAI and sowing date are estimated from remote sensing data. When estimation is not from remote sensing, the average value of early LAI or of sowing date over the 31 fields is calculated and that same value is used for every field.

	Only LAI estimated		LAI and sowing date estimated	
	RMSE without assimilation	RMSE with assimilation	RMSE without assimilation	RMSE with assimilation
Sowing date (days)	–	–	9	9
LAI at 500°C.days after sowing (m^2 m^{-2})	0.34	0.29	0.32	0.40

were estimated from remote sensing data, the errors were as large as or larger than without assimilation. On the other hand, assimilation did bring some improvement when only LAI_{500} was estimated from the remote sensing data.

A different example of the use of remote sensing data is given in Launay and Guérif (2005). The purpose of this study was to estimate sugar beet yield in all fields in a region using the model SUCROS. Soil properties were available from a soil map associated with pedotransfer functions (Jamagne et al., 1977). However, it was clear that the soil map information was insufficient, in particular for properties related to soil water availability. First of all, the soil map provided information only to a depth of 1.2 m, whereas several experiments in the region have shown that sugar beet roots reach depths greater than 1.2 m. Second, the soil map did not provide information on chalk type for soils with a chalk substratum, whereas it is known that there are two different types of chalk with different water availability properties. The first is a hard material, with low sensitivity to frost and impenetrable by roots and the second a soft material, cracked and penetrable by roots. The soft chalk can serve as a reservoir for soil water, which can be mobilized by capillary rise or by direct penetration of roots (Maucorps, personal communication).

The problem then is that soil properties, and as a consequence water availability, are poorly known. Remote sensing is a good candidate for providing additional information about soil characteristics. However, the situation is complicated by the fact that the crop model does not take properties like penetrability for roots or capillary rise as inputs. It was decided then to use remote sensing data to estimate root growth rate and maximum rooting depth. These parameters also affect soil water availability to the plant, and so can be used to compensate for errors in soil characteristics. Notice that, here, the objective is not to provide the true inputs for the model, but rather to compensate for errors in inputs that are poorly known.

Figure 4 shows the effect of using remote sensing data for one particular plot, which had a chalky substratum starting at a depth of 0.2–0.3 m. The right graph shows predicted (with and without assimilation) and observed values for LAI. The left graph shows predicted and observed values of the vegetation index VI, which is a combination of reflectance values. The radiative transfer model SAIL was used to convert from LAI to VI. The prediction without assimilation assumes that the soil description at 1.2 m can be extended to 1.8 m, and that where chalk is present it has the same water holding properties as silt loam.

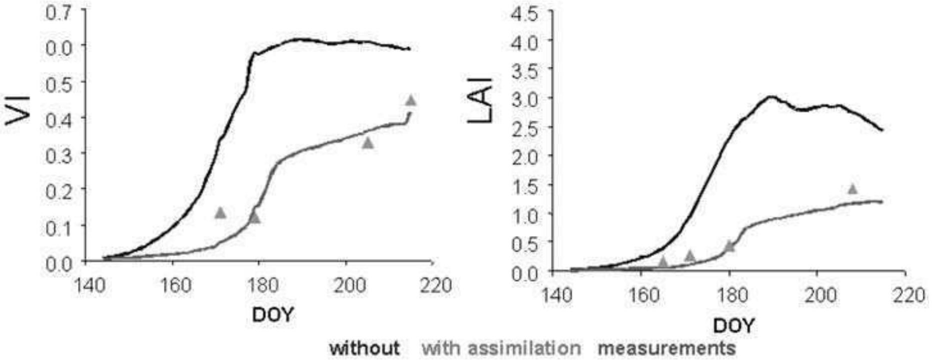


Figure 4. Simulated vegetation index (VI) (left) and leaf area index (LAI) (right) values without (-----) and with (——) assimilation of remote sensing data, compared to observed values (▲) obtained from remote sensing for VI and from ground measurements for LAI.

This led to an overestimation of the soil water holding capacity and to an overestimation of LAI. The available remote sensing data consisted of five SPOT-HRV and airborne images. Using these data, the calculated rate of root growth was 0.010 m day^{-1} (compared to the standard value of 0.012 m day^{-1}) and maximum root depth was 0.81 m compared to the standard value of 1.8 m. Without assimilation the predicted yield was 84.8 t ha^{-1} and with assimilation 53.8 t ha^{-1}. The latter value is much closer to the observed yield of 50.3 t ha^{-1}.

Chapter 17 provides another example of the use of remote sensing data to estimate soil input variables.

4.2. Use of remote sensing data for forcing state variables

So far we have discussed how remote sensing data can be used to obtain values of model input variables at every point where the model will be used (at every crop model support unit, in the language of scale change). An alternative approach is to use remote sensing to obtain values of crop state variables at every crop model support unit, and to use those values to replace the values calculated by the model. The idea is that replacing calculated by measured state variables could compensate for errors in the input variables. Such an approach was proposed by Steven et al. (1983) and Leblon et al. (1991), using very simple models describing net primary production. Delécolle et al. (1992) and Clevers et al. (2002) employed this approach using more detailed crop models.

Clevers et al. (2002) compared two ways of replacing LAI values in their model ROTASK by values obtained by remote sensing. They had estimated LAI values from 5 fields of durum wheat at 3 dates, obtained from SPOT-HRV images using empirical relationships between reflectance and LAI. In the first approach, LAI values were estimated at all dates by interpolation, assuming that LAI followed a logistic curve over time. Then those LAI values were used every day to replace the LAI values calculated by the model. In the second approach, the model values of LAI were only replaced by the values derived from remote sensing at the three acquisition dates. The first approach improved final yield estimates for the 5 fields, but the second approach led to even greater improvement. The first approach would probably have been even better if more measurement dates had been available. Three dates is probably insufficient.

When dealing with a large spatial domain, one usually has to rely on low-resolution satellite data. As a result, the support unit of the observed data may be larger than the support unit of the crop model, and a change of scale is necessary. For example, in a study to estimate regional wheat yield by Faivre et al. (2000), the support unit of the observed data (the pixel) was 1 km^2 (SPOT-VGT) while the support unit of the crop model was the individual field, which was generally much smaller. Using knowledge of land use provided by high-resolution satellite data, they were able to deconvolute the original signal and to recover the specific reflectances of each crop included in the pixel (Faivre and Fischer, 1997). Then, they derived LAI values from the reflectance values using empirical relationships and then interpolated between these values in order to obtain LAI values for each day. Finally, they fed these values into the STICS model, replacing the values calculated by the model for each crop type in each pixel. The results showed very good agreement between estimated and announced values of wheat production.

Whether remote sensing data is used to provide values of input variables or values of state variables at each point in space, special attention must be paid to the quality of the data and of the variables derived from those data (LAI or chlorophyll content estimates). Evaluation of the associated errors is essential in order to study how they propagate to other model outputs. Finally, the number of remote sensing data dates is very important and largely determines how well one can estimate different input variables or state variables.

5. Obtaining the outputs for multiple fields

Crop models are developed for single homogeneous fields. Producing outputs at the regional scale requires a scale change which implies the consideration of new processes and properties, emerging at this new scale and revealed by the extension of the system considered. They influence the soil–plant–atmosphere system but they are not represented in crop models. These processes can concern physical transfers between neighboring units, including water transfer between fields, pathogen propagation, weed or pollen diffusion, etc. The interactions between fields can also result from the multiplicity of actors in a region and from the decisions they make. They arise because, at this scale, human and economic sub-systems cannot be neglected. For example, at the scale of an irrigated area, the water resource must be allocated between farmers. At the farm scale, not only water allocation but also other management decisions are interrelated between fields due to the constraints of labor and equipment. In some cases interactions can be ignored or included implicitly, while in other cases the interactions will be explicitly modeled.

5.1. No explicit representation of interactions

In this case, it is common not to run the model for every field, but rather to divide the region into elementary simulation units and to run the model independently for each unit. The simulation units are determined by defining homogeneous zones for the most sensitive input data. For example, a soil map could be overlaid with a climatic zone map in order to obtain a map of homogeneous pedoclimatic units. This could be overlaid with a map of administrative regions, for which there is cropping and management information, to further sub-divide the units so that each new unit also belongs to a single administrative region.

The final result would be obtained as a sum (for example of water requirements, Leenhardt et al., 2004a,b) or as an average (for example of yield, Donet et al., 2001) over simulation units. To obtain these results, it might be necessary to run the model several times for each elementary simulation unit. If various weather scenarios are considered, one might want to average over them for each elementary simulation unit. Also, an elementary simulation unit might have a distribution of farm types or management practices. Then the model would be run for each.

Even when interactions between fields are not considered explicitly, it may be possible and useful to include them implicitly. An example is given by the SIMSEM model (Leenhardt and Lemaire, 2002; see above). This model calculates a range of sowing dates, which takes into account the fact that different fields (and other activities) on a farm interact because they share the labor resources of the farm. Fields can then be treated as independent, but the range of sowing dates implicitly takes into account interaction due to shared resources.

Another approach is to inject information from remote sensing (generally, values of LAI) into the crop model. The actual growth and development of the crop automatically take into account interactions, and so adjusting the model to data from each field is an implicit way of taking interactions into account. The remote sensing data can be used to re-estimate parameters and/or input values of the crop model (Guérif and Duke, 1998; Launay and Guérif, 2005). They can also be used to force the crop model to be consistent with the observed data over the course of the growing season (Faivre et al., 2000).

5.2. Modeling interactions between fields

Examples of situations where interactions may be explicitly modeled include fields in a landscape with slope, where surface water flow (run-off and run-on) between fields needs to be taken into account, and the study of contamination of fields without genetically modified (GM) crops by fields with GM crops, where the exchange of pollen must be modeled (Colbach et al., 2005).

Also, one often wants to couple crop models with other models, in particular hydrological or meteorological models, in order to study some aspect of the overall behavior of a region. In this case, the other model may provide the modeling of interactions. For example, the hydrological model will calculate water flow, both surface and sub-surface. In general, the major interest in such cases is not on crop behavior. Examples include the studies by Beaujouan et al. (2001) and Gomez and Ledoux (2001). The first studied water quality for a small catchment and the second for a large river basin. They coupled a crop model with a hydrological model to simulate results at the watershed outlet. Coupling crop models with models of three-dimensional hydrology, pollen exchange or farm operations may require restructuring the crop models so that multiple crops can be simulated in the same simulation study and so that information can be exchanged between models at each simulation time step (Hansen and Jones, 2000). Also, the type of interactions taken into account may influence the choice of the simulation units (see for example, Kiniry et al., 2002).

6. Evaluation

6.1. Validation data – few in number and often not totally pertinent

In principle, one can evaluate a model used at the regional level by comparing the results with observed data. A basic problem is the lack of data. We are interested specifically in the results for a particular region, so that there is only one single result for comparison per year (or per season). This is in contrast to evaluating a model on the average for a type of field, where one can compare with numerous fields each year. Furthermore, it can be very difficult to get reliable data at the regional level. Various approximations may be necessary, leading to a fairly high uncertainty in the observed value. As shown in Chapter 2 on evaluation, when there is measurement error the mean squared error of prediction (MSEP) is augmented by the variance of the measurement error. If that variance is large, as it often will be for regional studies, then the measurement error may be the major part of MSEP, masking the error between the model and the true result.

Rabaud and Ruget (2002) discuss two specific problems with validation data that are probably quite common. They used expert estimation of forage production in each region of France to obtain data for comparison with the ISOP model. The first problem is that the expert estimates may be more or less accurate depending on the expert. Mignolet et al. (2004) also noted this problem. With regard to past land use, two different experts asked about the same period and the same region did not provide the same information, especially for the more distant past. The second problem noted by Rabaud and Ruget (2002) is that it may not be feasible to get validation data that is exactly comparable to what the model predicts. In their case each expert reported for an administrative region, while the model was run for specially defined forage regions. Another example of this is provided by the study of Leenhardt et al. (2004b), who simulated total water demand for irrigation in a region. For comparison they had to rely on data relative to farmers who subscribe to a collective irrigation system, while the simulation concerns all farms in the region including farmers who irrigate from their own reservoirs.

6.2. Estimating error at the regional level

Suppose that the regional output of interest is the sum of outputs from elementary units. For example, we might be interested in total crop production or total water use for a number of fields. (The case where the quantity of interest is an average over fields rather than a sum is easily derived from the treatment here.) We wish to estimate the mean squared error of prediction (MSEP) for the regional result. For a general discussion of MSEP, see Chapter 2.

If past results for the region and corresponding model predictions are available, then they can be used to estimate MSEP for the region. We make the simplifying assumption throughout this section that the data available for estimating regional prediction error were not used for model development. Also, for concreteness we suppose that the quantity of interest is regional production (tons of wheat, for example). Then the estimate of MSEP for the regional result is simply

$$\hat{M}SEP = \frac{1}{V}\sum_{v=1}^{V}(P_v - \hat{P}_v)^2 \tag{5}$$

where the index v, $v = 1, \ldots, V$ identifies the year, P_v is the true regional result for year v and $\hat{P}_v$ is the corresponding model prediction. That is, we simply evaluate the mean squared error for past results, and that estimates the mean squared error for future applications of the model.[3]

Suppose, however, that past data for the entire region are unavailable or unreliable. Is there another way of estimating model error on a regional scale? In particular, if we have past data for just a sample of fields from the region, and corresponding model predictions, can we use that information to estimate MSEP for the region? We propose two ways of doing this.

Suppose first that we have results from a random sample of N fields for each of V years. Then a simple estimate of MSEP is obtained by replacing the production values in Eq. (5) by values estimated using the sample. In particular, our estimate of P_v is $S\left(\sum_{i=1}^{N} P_{i,v} / \sum_{i=1}^{N} S_{i,v}\right)$ where $P_{i,v}$ and $S_{i,v}$ are respectively, the true production and the area of the ith sampled field in year v and S is total area of the fields in the area. An estimate of $\hat{P}_v$ is obtained using an analogous equation, with calculated production $\hat{P}_{i,v}$ in place of true production. We assume for simplicity that the total area in question is the same every year. The treatment could easily be generalized to the case where the total area varies between years.

The second approach is more complex but gives insight into what determines overall error. We introduce a statistical description of model error, which takes into account bias, year effects and field effects. Model error in yield (production per unit area) can be written as

$$d(u,v) = Y(u,v) - \hat{Y}(u,v) = \mu + a(u) + b(v) + c(u,v) \tag{6}$$

Here u identifies a particular field and v a particular year. $Y(u,v) = P(u,v)/S(u)$ and $\hat{Y}(u,v) = \hat{P}(u,v)/S(u)$ are respectively, true and calculated values of yield for field u, year v and $S(u)$ is the area of the indicated field. The model error is thus written as an average model error μ, a field effect $a(u)$, a year effect $b(v)$ and an interaction $c(u,v)$. We will treat both fields and years as infinite populations. This is natural for years. For fields it is an approximation that is reasonable if the number of fields in the region is

---

[3] The notation *MSEP* here is equivalent to *MSEP*($\hat{\theta}$) in Chapter 2. The prediction error here refers to a specific model with a specific parameter vector $\hat{\theta}$, as in Chapter 2, but to simplify the notation we do not show the dependence on $\hat{\theta}$ explicitly.

fairly large. The definitions of the terms on the right-hand side of Eq. (6) are

$$\begin{aligned} \mu &= E\,[d(u,v)] \\ a(u) &= E\,[d(u,v)|u] - \mu \\ b(v) &= E\,[d(u,v)|v] - \mu \\ c(u,v) &= d(u,v) - a(u) - b(v) - \mu \end{aligned} \tag{7}$$

It is easily seen from their definitions that

$$E\,[a(u)] = E\,[b(v)] = E\,[c(u,v)|u] = E\,[c(u,v)|v] = 0$$

Furthermore, it can be shown that

$$E\,[a(u)b(v)] = E\,[a(u)c(u,v)] = E\,[b(u)c(u,v)] = 0$$

Finally, we assume that the variances are the same for each field and year, and define

$$\begin{aligned} \sigma_{\mathrm{A}}^2 &= \mathrm{var}\,[a(u)] \\ \sigma_{\mathrm{B}}^2 &= \mathrm{var}\,[b(v)] \\ \sigma_{\mathrm{AB}}^2 &= \mathrm{var}\,[c(u,v)] \end{aligned}$$

We can now derive expressions for the mean squared error of prediction. The error for total regional production in year v is $N_{\mathrm{tot}}E\,[S(u)d(u,v)|v]$ where N_{tot} is the total number of fields in the region. The mean squared error of prediction is that error squared, averaged over years, i.e.

$$\begin{aligned} MSEP_{\mathrm{region}} &= E\{\{N_{\mathrm{tot}}E[S(u)d(u,v)|v]\}^2\} \\ &= N_{\mathrm{tot}}^2E\{[S(u)]^2\}E\{\{E[\mu + a(u) + b(v) + c(u,v)|v]\}^2\} \\ &= N_{\mathrm{tot}}^2E\{[S(u)]^2\}E\{[\mu + b(v)]^2\} = N_{\mathrm{tot}}^2E\{[S(u)]^2\}(\mu^2 + \sigma_{\mathrm{B}}^2) \end{aligned} \tag{8}$$

where we have used the fact that $E\,[a(u)] = E\,[c(u,v)|v] = 0$ and $E\,[b(v)|v] = b(v)$. We have also assumed that errors are independent of field size.

Equation (8) shows that the regional mean squared error of prediction just depends on average model bias and on the year effect in model error. The field effect in model error does not contribute because it is on the average zero, and so it cancels out when we are interested in a sum (or an average) over fields. Given results from a sample of fields, one can use analysis of variance software to estimate μ^2 and σ_{B}^2. Plugging into Eq. (8) would then give an estimate of $MSEP_{\mathrm{region}}$.

6.3. Error propagation studies

The number of input variables for a crop model used for a single field is already appreciable. When the model is used for multiple fields, then the overall number of input variables is multiplied by the number of fields. Furthermore, additional approximations are usually involved in using a model for multiple fields. In particular, if total output is based on sampling fields within the study area, then the sampling is an important additional approximation. Given the very large number of possible sources of error, it is even more important here than for single fields to identify which errors are most important.

Uncertainty and sensitivity analysis (Chapter 3) treat the problems of propagation and decomposition of model error. The specific problem here involves the effect of different errors in multiple field studies, in particular studies where spatial organization is taken into account. For linear models analytical methods exist (for example, Heuvelink et al., 1989). For strongly non-linear models like crop models, these methods do not apply. Crosetto et al. (2000) and Tarantola et al. (2000) propose applications of uncertainty and sensitivity analysis to GIS-based models that need not be linear. They consider how to estimate the precision needed for the various inputs in order to obtain a specified precision for an output.

References

Adams, R.M., Rozenweig, C., Peart, R.M., Ritchie, J.T., McCarl, B.A., Glyer, J.D., Curry, R.B., Jones, J.W., Boote, K.J., Allen, Jr. L.H., 1990. Global climate change and US agriculture. Nature 345, 219–224.

Ahuja, L.R., Naney, J.W., Williams, R.D., 1985. Estimating soil water characteristics from simpler properties or limited data. Soil Science Society of America Journal 49, 1100–1105.

Ahuja, L.R., Nofziger, D.L., Swartzendruber, D., Ross, J.D., 1989. Relationship between Green and Ampt parameters based on scaling concepts and field-measured hydraulic data. Water Resource Research 25, 1766–1770.

Baret, F., Weiss, M., Troufleau, D., Prévot, L., Combal, B., 2000. Maximum information exploitation for canopy characterization by remote sensing. In: Brysson, R.J., Howard, V., Riding, A.E., Simmons, L.P., Stewen, M.D. (Eds), Remote Sensing in Agriculture, Vol. 60. UK, June 26–28, 2000, Aspects of Applied Biology (GBR) Conference, Cirencester, 71–82.

Barrow, E., 1993. Scenarios of climate change for the European Community. European Journal of Agronomy 2, 247–260.

Bastet, G., Bruand, A., Quétin, P., Cousin, I., 1998. Estimation des propriétés de rétention en eau à l'aide de fonctions de pédotransfert (FPT): Une analyse bibliographique. Etude et Gestion des Sols 1, 7–28.

Bastet, G., Bruand, A., Voltz, M., Bornand, M., Quétin, P., 1999. Performance of available pedotransfer functions for predicting the water retention properties of French soils. In: van Genuchten, Th., Leij, F.J. (Eds), Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media. Proceedings of the International Workshop, Riverside, California, October 22–24, 1997, pp. 981–992.

Batjes, N.H., 1996. Development of a world data set of soil water retention properties using pedotransfer rules. Geoderma 71, 31–52.

Beaujouan, V., Durand, P., Ruiz, L., 2001. Modelling the effect of the spatial distribution of agricultural practices on nitrogen fluxes in rural catchments, Ecological Modelling 137, 93–105.

Beckett, P.H.T., Webster, R., 1971. Soil variability: a review. Soils and Fertilizers 34, 1–15.

Bénichou, P., Le Breton, O., 1987. Prise en compte de la topographie pour la cartographie des champs pluviométriques statistiques: la méthode Aurelhy. Agrométéorologie des régions de moyennes montagnes, April 16–17, 1986. Colloque INRA 39, 51–68.

Bindi, M., Castellani, M., Maracchi, G., Miglietgta, F., 1993. The ontogenesis of wheat under scenarios of increased air temperature in Italy: a simulation study. European Journal of Agronomy 2, 261–280.

Boulaine, J., 1980. Pédologie Appliquée. Collection Sciences Agronomiques, Masson, Paris.

Bouma, J., 1989. Land qualities in space and time. In: Bouma, J., Bregt, A.K. (Eds), On Land Qualities in Space and Time. Proceedings of ISSS Symposium, Pudoc, Wageningen, The Netherlands, August 22–26, 1988, pp. 3–13.

Bouma, J., de Laat, P.J.M., Awater, R.H.C.M., van Heesen, H.C., van Holst, A.F., van de Nes, T.J.M., 1980. Use of soil survey data in a model for simulating regional soil moisture regimes. Soil Science Society of America Proceedings 44, 808–814.

Bouman, B., 1992. Linking physical remote sensing models with crop growth simulation models, applied for sugar beet. International Journal of Remote Sensing 13, 2565–2581.

Bruand, A., 2004. Preliminary Grouping of soils. In: Pachepsky, Y., Rawls, W.J. (Eds), Development of Pedotransfer Functions in Soil Hydrology. Development in Soil Science, Vol. 30, Elsevier, Amsterdam, pp. 159–174.

Bruand, A., Duval, O., Gaillard, H., Darthout, R., Jamagne, M., 1996. Variabilité des propriétés de retention en eau des sols: Importance de la densité apparente. Etude et Gestion des Sols 3, 27–40.

Bruand, A., Perez Fernadez, P., Duval, O., Quétin, Ph., Nicoullaud, B., Gaillard, H., Raison, L., Pessaud, J.F., Prud'Homme, L., 2002. Estimation des propriétés de rétention en eau des sols: Utilisation de classes de pédotransfert après stratifications texturale et texturo-structurale. Etude et Gestion des Sols 9, 105–125.

Bruand, A., Perez Fernandez, P., Duval, O., 2003. Use of class pedotransfer functions based on texture and bulk density of clods to generate water retention curves. Soil Use and Management 19, 232–242.

Brun, E., Martin, E., Simon, V., Gendre, C., Coléou, C., 1989. An energy and mass model of snow cover suitable for operational avalanche forecasting, Journal of Glaciology 35, 333–342.

Brun, E., David, P., Sudul, M., Brunot, G., 1992. A numerical model to simulate snow-cover stratigraphy for operational avalanche forecasting, Journal of Glaciology 38, 13–22.

Brus, D.J., de Gruijter, J.J, Breeusma, A., 1992. Strategies for updating soil survey information: a case study to estimate phosphate sorption characteristics. Journal of Soil Science 43, 567–581.

Burroughs, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Spatial Information Systems and Geostatistics, Clarendon Press, 2nd Edition. Oxford. 333 pp. ISBN: 0-19-823366-3.

Campbell, J.B., 2002. Introduction to Remote Sensing, 3rd Edition. The Guilford Press, New York. 564 pp. ISBN: 1-57230-640-8.

Chipanshi, A.C., Ripley, E.A., Lawford, R.G., 1999. Large-scale simulation of wheat yields in a semi-arid environment using a crop-growth model. Agricultural Systems 59, 57–66.

Chuvieco, E., 2002. Teledetección ambiental. La observación de la Tierra desde el Espacio. 1st Edition. Ariel Ciencia, Barcelona (Spain). 586 pp. ISBN: 84-344-8047-6.

Clevers, J.G.P.W., Vonder, O.W., Jongschaap, R.E.E., Desprats, J.F., King, C., Prevot, L., Bruguier, N., 2002. Using SPOT-HRV data for calibrating a wheat growth model under mediterranean conditions. Agronomie 22, 687–694.

Colbach, N., Fargue, A., Sausse, C., Angevin, F., 2005. Evaluation and use of a spatio-temporal model of cropping system effects on gene escape from transgenic oilseed rape varieties: Example of the GeneSys model applied to three co-existing herbicide tolerance transgenes. European Journal of Agronomy 22, 417–440.

Cosby, B.J., Hornberger, G.M., Clapp, R.B., Ginn, T.R., 1984. A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. Water Resource Research 20, 682–690.

Cotter, J., Nealon, J., 1987. Area Frame Design for Agricultural Surveys. National Agricultural Statistics Service, US Department of Agriculture, Washington DC.

Creutin, J.D., Obled, C., 1982. Objective analysis and mapping techniques for rainfall fields: an objective comparison. Water Resource Research 18, 413–431.

CRAMP, 1998. SICOMORE: SImulation éCOnomique de MOdèles REprésentatifs de l'agriculture. Chambre Régionale d'Agriculture de Midi-Pyrénées, Auzeville. 27 p.

Crosetto, M., Tarantola, S., Saltelli, A., 2000. Sensitivity and uncertainty analysis in spatial modeling based on GIS, Agriculture, Ecosystems & Environment 81, 71–79.

Daroussin, J., King, D., 1997. A pedotransfer rules database to interpret the soil geographical database of Europe for environmental purposes. In: Bruand, A., Duval, O., Wösten, J.H.M., Lilly, A. (Eds), The Use of Pedotransfer in Soil Hydrology Research in Europe. Proceedings of the Second Workshop of the Project "Using Existing Soil Data to Derive Hydraulic Parameters for Simulation Modelling in Environmental Studies and in Land Use Planning," Orléans, France, October 10–12, 1996. INRA Orélans and EC/JRC Ispra.

Delécolle, R., Maas, S.J., Guérif, M., Baret, F., 1992. Remote sensing and crop production models: present trends. ISPRS Journal of Photogrammetry and Remote Sensing (NLD) 47, 145–161.

Donet, I., 1999. Etude sur l'interpolation de l'ETP quotidienne en points de grille, Report SCEM/SERV/AGRO. Toulouse, February.

Donet, I., Le Bas, C., Ruget, F., Rabaud, V., 2001. Guide d'utilisation d'ISOP. Agreste Chiffres et Données Agriculture, 134.

Ducrot, D., Gouaux, P., 2004. Caractérisation des agro-systèmes de la plaine alluviale de la Garonne et des côteaux du Gers, mise en évidence de leurs changements aux cours des vingt dernières années. In: Les Systémes de Production Agricole: Performances, Évolutions, Perspectives, Colloque de la Société Française d'Economie Rurale, Lille, France, November 18–19, 2004.

Durand, Y., Brun, E., Mérindol, L., Guyomarc'h, G., Lesaffre, B., Martin, E., 1993. A meteorological estimation of relevant parameters for snow models. Annals of Glaciology 18, 65–71.

Etchevers, P., Golaz, C., Habets, F., Noilhan, J., 2002. Impact of a climate change on the Rhone river catchment hydrology, Journal of Geophysical Research 107, 101029–101048.

Faivre, R., Bastié, C., Husson, A., 2000. Integration of VEGETATION and HRVIR into yield estimation approach. In: Gilbert Saint (Ed), Proceedings of "Vegetation 2000, 2 years of Operation to Prepare the Future." Space Application Institute and Joint Reasearch Center, Ispra, Lake Maggiore, Italy, April 3–6, pp. 235–240.

Faivre R., Fischer, A., 1997. Predicting crop reflectances using satellite data observing mixed pixels. Journal of Agriculture, Biology and Environmantal Statistics 2, 87–107.

Franzmeier, D.P., 1991. Estimation of hydraulic conductivity from effective porosity data for some Indiana soils. Soil Science Society of America Journal 55, 1801–1803.

Gallego, F.J., 1995. Sampling frames of square segments, Report EUR 16317 EN. Joint Research Centre, European Commission, pp. 68, ISBN 92-827-5106-6.

van Genuchten, M.Th., Leij, F.J., 1992. On estimating the hydraulic properties of unsaturated soils. In: van Genuchten, M.Th., Leij, F.J., Lund, L.J. (Eds), Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils. University of California, Riverside, October 11–13, 1989, pp. 1–14.

Gomez, E., Ledoux, E., 2001. Démarche de modélisation de la dynamique de l'azote dans les sols et de son transfert vers les aquifères et les eaux de surface. Comptes-rendus de l'Académie d'Agriculture de France 87, pp. 111–120.

Guérif, M., Duke, C., 1998. Calibration of the SUCROS emergence and early growth module for sugarbeet using optical remote sensing data assimilation. European Journal of Agronomy 9, 127–136.

Gupta, S.C., Larson, W.E., 1979. Estimating soil water characteristics from particle size distribution, organic matter percent, and bulk density. Water Resources Research 15, 1633–1635.

Hall, D.G., Reeve, M.J., Thomasson, A.J., Wright, V.F., 1977. Water retention, porosity and density of field soils, Technical Monograph No. 9. Soil Survey of England & Wales, Harpenden.

Hansen, J.W., Jones, J.W., 2000. Scaling-up crop models for climate variability application. Agricultural Systems 65, 43–72.

Hartkamp, A.D., White, J.W., Hoogenboom, G., 1999. Interfacing geographic information systems with agronomic modeling: a review, Agronomy Journal 91, 761–772.

Heinemann, A.B., Hoogenboom, G., Faria de, R.T., 2002. Determination of spatial water requirements at county and regional levels using crop models and GIS. An example for the State of Parana, Brazil. Agricultural Water Management 52, 177–196.

Herrero, J., Casterad, M.A., 1999. Using satellite and other data to estimate the annual water demand of an irrigation district. Environmental Monitoring and Assessment 55, 305–317.

Heuvelink, G.B.M., Burrough, P.A., Stein, A., 1989. Propagation of errors in spatial modelling with GIS. International Journal of Geographical Information Systems 3, 303–322.

Jamagne, M., Betremieux, R., Begon, J.C., Mori, A., 1977. Quelques données sur la variabilité dans le milieu naturel de la réserve en eau des sols. Bulletin Technique d'Information, 324–325, 627–641.

Kantey, B.A., Williams, A.A.B., 1962. The use of soil engineering maps for road projects. Transactions of the South African Institution of Civil Engineers 4, 149–159.

Kern, J.S., 1995. Evaluation of soil water retention models based on basic soil physical properties. Soil Science Society of America Journal 59, 1134–1141.

Kiniry, J.R., Arnold, J.G., Xie. Y., 2002. Applications of models with different spatial scales. In: Ahuja, L.R., Ma, L., Howell, T.A. (Eds), Agricultural system models in field research and technology transfer. Lewis Publishers, Boca Raton, FL, USA, 207–226.

Koekkoek, E., Bootlink, H., 1999. Neural network models to predict soil water retention. European Journal of Soil Science 50, 489–495.

Lal, H., Hoogenboom, G., Calixte, J.P., Jones, J.W., Beinroth, F.H., 1993. Using crop simulation models and GIS for regional productivity analysis. Transaction of the American Society of Agricultural Engineers 36, 175–184.

Launay, M., 2002. Diagnostic et prévision de l'état des cultures à l'échelle régionale: couplage entre modèle de croissance et télédétection. Application à la betterave sucrière en Picardie. Ph.D. Thesis, Institut National Agronomique Paris-Grignon, France, p. 72.

Launay, M., Guérif, M., 2005. Assimilating remote sensing data into a crop model to improve predictive performance for spatial applications. Agriculture, Ecosystems and Environment 111, 321–339.

Laurent, H., Jobard, A., Toma, A., 1998. Validation of the satellite and ground based estimates of precipitation over the Sahel, Atmospheric Research 47–48, 651–670.

Leblon, B., Guérif, M., Baret, F., 1991. The use of remotely sensed data in estimation of PAR use efficiency and biomass production of flooded rice. Remote Sensing of Environment 38, 147–158.

Leenhardt, D., Lemaire, P., 2002. Estimating the spatial and temporal variability of sowing dates for regional water management. Agricultural Water Management 55, 37–52.

Leenhardt, D., Voltz, M., Bornand, M., Webster, R., 1994. Evaluating soil maps for prediction of soil water properties. European Journal of Soil Science 45, 293–301.

Leenhardt, D., Trouvat, J.-L., Gonzalès, G., Pérarnaud, V., Prats, V., Bergez, J.-E., 2004a. Estimating irrigation demand for water management on a regional scale. I. ADEAUMIS, a simulation platform based on bio-decisional modelling and spatial information. Agricultural Water Management 68, 207–232.

Leenhardt, D., Trouvat, J.-L., Gonzalès, G., Pérarnaud, V., Prats, V., Bergez, J.-E., 2004b. Estimating irrigation demand for water management on a regional scale. II. Validation of ADEAUMIS. Agricultural Water Management 68, 233–250.

Leenhardt, D., Cernesson, F., Mari, J.-F., Mesmin, D., 2005. Anticiper l'assolement pour mieux gérer les ressources en eau: comment valoriser les données d'occupation du sol? Ingénierie, N°42, 13–22.

Leij, F., Alves, W.J., van Genuchten, M.Th., Williams, J.R., 1996. The UNSODA unsaturated soil hydraulic database, User's Manual Version 1.0. EPA/600/R-96/095. National Risk Management Laboratory, Office of Research and Development, Cincinnati, OH.

Leij, F.J., Alves, W.J., van Genuchten, M.Th., Williams, J.R., 1999. The UNSODA unsaturated soil hydraulic database. In: van Genuchten, M.Th., Leij, F.J. (Eds), Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media. Proceedings of the International Workshop, Riverside, California, October 22–24, 1997, pp. 1269–1281.

Le Moigne, P., 2002. Description de l'analyse des champs de surface sur la France par le système SAFRAN, Scientific report CNRM/GMME, No 77. Toulouse, June.

Lenhardt, R.J., 1984. Effects of Clay–Water Interactions on Water retention in Porous Media. Ph.D. Thesis. 145 pp. Oregon State University, Corvallis, OR, USA.

Lillesand, T.M., Kiefer, R.W., Ghipman, J.W., 2003. Remote Sensing and Image Interpretation. 5th Edition. John Wiley & Sons, Inc. Hoboken, NJ. 784 pp. ISBN: 0-471-15227-7.

Lilly, A., 1997. A description of the HYPRES database (Hydraulic properties of European Soils). In: Bruand, A., Duval, O., Wösten, J.H.M., Lilly, A. (Eds), The use of Pedotransfer Functions in Soil Hydrology Research. Proceedings of Second Workshop of the Project Using Existing Soil Data to Derive Hydraulic Parameters of Simulation Modelling in Environmental Studies and in Land Use Planning, Orléans, France, October 10–12, 1996, pp. 161–184.

Lilly, A., Wösten, J.H.M., Nemes, A., Le Bas, C., 1999. The development and use of the HYPRES database in Europe. In: van Genuchten, M.Th., Leij, F.J. (Eds), Characterization and Measurement of the Hydraulic Properties of Unsaturated Porous Media. Proceedings of the International Workshop, Riverside, California, October 22–24, 1997, 1283–1204.

Mari, J.F., Le Ber, F., Benoît, M., 2000. Fouille de données par modèles de Markov cachés. Proceedings of the Journées francophones d'ingénierie des connaissances, Toulouse, May 10–12, 2000, pp. 197–205.

Maton, L., Leenhardt, D., Goulard, M., Bergez, J.E., 2005. Assessing the irrigation strategies over a wide geographical area from structural data about farming systems. Agricultural Systems 86, 293–311.

Mignolet, C., Schott, C., Benoît, M., 2004. Spatial dynamics of agricultural practices on a basin territory: a retrospective study to implement models simulating nitrate flow. The case of the Seine basin. Agronomie 24, 219–236.

Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93, 225–253.

Morse, R.K., Thornburn, T.H., 1961. Reliability of soil map units. In: Proceedings of the 5th International Conference of Soil Mechanics Federation Engineering, Vol. 1, Dunod, Paris, pp. 259–262.

Moulin, S., 1995. Assimilation d'observations satellitaires courtes longueurs d'onde dans un modèle de fonctionnement de culture. Ph.D. Thesis, Université Paul Sabatier de Toulouse, France, p. 265.

Nemes, A., Rawls, W.J., 2004. Soil texture and particle size distribution to estimate soil hydraulic properties. In: Pachepsky, Y., Rawls, W.J. (Eds), Development of Pedotransfer Functions in Soil Hydrology, Development in Soil Science, Vol. 30, Elsevier, pp. 47–70.

Noilhan, J., Planton, S., 1989. A simple parameterization of land surface processes for meteorological models, Monthly Weather Review 117, 536–549.

Noilhan, J., Boone, A., Etchevers, P., 2002. Application of climate change scenarios to the Rhone basin, Report No. 4. ECLAT-2, Toulouse Workshop, pp. 58–74.

Pachepsky, Ya.A., Timlin, D., Várallyay, G., 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Science Society of America Journal 60, 727–773.

Paydar, Z., Cresswell, H.P., 1996. Water retention in Australian soils, II. Prediction using particle size, bulk density and other properties. Australian Journal of Soil Research 34, 679–693.

Pellenq, J., Boulet, G., 2004. A methodology to test the pertinence of remote-sensing data assimilation into vegetation models for water and energy exchange at the land surface. Agronomie 24, 197–204.

Pinter, Jr. P.J., Ritchie, J.C., Hatfield, J.L., Hart, G.F., 2003. The agricultural research service's remote sensing program: an example of interagency collaboration. Photogrammetric Engineering & Remote Sensing 69, 615–618.

Prévot, L., Chauki, H., Troufleau, D., Weiss, M., Baret, F., Brisson, N., 2003. Assimilating optical and radar data into the STICS crop model for wheat, Agronomie 23, 297–303.

Priya, S., Shibasaki, R., 2001. National spatial crop yield simulation using GIS-based crop production model. Ecological Modelling 135, 113–129.

Rabaud, V., Ruget, F., 2002. Validation de variations interannuelles d'estimations de productions fourragères appliquées à de petites régions françaises. In: Leenhardt, D., Faivre, R., Benoît, M. (Eds), Actes du séminaire "Spatialisation des modéles de culture," Toulouse, January 14–15, 2002.

Rawls, W.J., 2004. Pedotransfer functions for the United States. In: Pachepsky, Y., Rawls, W.J. (Eds), Development of Pedotransfer Functions in Soil Hydrology. Development in Soil Science, Vol. 30, Elsevier, Amsterdam, pp. 437–447.

Rawls, W.J., Brakensiek, D.L., Saxton, K.E., 1982. Estimation of soil water properties. Transaction of the American Society of Agricultural Engineers 26, 1747–1752.

Rawls, W.J., Ahuja, L.R., Brakensiek, D.L., 1992. Estimating soil hydraulic properties from soils data. In: van Genuchten, M.Th., Leij, F.J. (Eds), Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils. Proceedings of the International Workshop, Riverside, California, October 11–13, 1989, pp. 329–340.

Ray, S.S., Dadhwal, V.K., 2001. Estimation of crop evapotranspiration of irrigation command area using remote sensing and GIS. Agricultural Water Management 49, 239–249.

Renger, M., 1971. Die ermittlung der porengröβenverteilung aus der körnung, dem gehalt an organischer sunstanz und der lagerungsdichte. Zeitschrift für Pflanzenernährung und Bodenkunde 130, 53–67.

République Française, 1946a. Circulaire administrative du Ministère de l'Agriculture n°147 M.E.4 du 06/04/1946.

République Française, 1946b. Circulaire administrative du Ministère de l'Agriculture N°152 M.E.4 du 18/05/1946.

République Française, 1949. Circulaire administrative du Ministère de l'Agriculture N°252 M.E.4 du 30/06/1949.

Ripert, C., Nouals, D., Franc, A., 1990. Découpage de Languedoc-Roussillon en petites régions naturelles. CEMAGREF, Aix-en-Provence, 26 p.

Rosenthal,W.D., Hammer, G.L., Butler, D., 1998. Predicting regional grain sorghum production in Australia using spatial data and crop simulation modelling. Agricultural and Forest Meteorology 91, 263–274.

Rozenweig, C., 1990. Crop response to climate change in the southern Great Plains: a simulation study. Professional Geographer 42, 20–37.

Russel, G., van Gardingen, P.R., 1997. Problems with using models to predict regional crop production. In: Van Gardingen, P.R., Foody, G.M., Curran, P.J. (Eds), Scaling-up from Cell to Landscape. Cambridge University Press, Cambridge, UK, pp. 273–294.

SCEES, 2005. Enquête TerUti-Lucas, Avis de conformité n°218/D131, Comité du label. Available from: http://www.cnis.fr/cnis/arretes/Avis-conformite/2005/teruti.pdf.

Schaap, M.G., 2004. Accuracy and uncertainty in PTF predictions. In: Pachepsky, Y., Rawls, W.J. (Eds), Development of Pedotransfer Functions in Soil Hydrology. Development in Soil Science, Vol. 30, Elsevier, Amsterdam, pp. 33–43.

Schaap, M.G., Leij, F.J., 1998. Database-related accuracy and uncertainty functions. Soil Science 163, 765–779.

Schaap, M.G., Leij, F.L., van Genuchten, M.Th., 1998. Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Science Society of America Journal 62, 847–855.

Seguin, B., Assad, E., Freteaud, J.P., Imbernon, J., Kerr, Y.H., Lagouarde, J.P., 1989. Use of meteorological satellite for balance monitoring in sahelian regions. International Journal of Remote Sensing 10, 1001–1017.

Semenov, M.A., Porter, J.R., Delécolle, R., 1993. Climatic change and the growth and development of wheat in the UK and France, European Journal of Agronomy 2, 293–304.

Sinowski, W., Scheinost, A.C., Auerswald, K., 1997. Regionalization of soil water retention curves in highly variable soilscape, II. Comparison of regionalization procedures using pedotransfer function. Geoderma 78, 145–159.

Smith, M., 2000. The application of climatic data for planning and management of sustainable rainfed and irrigated crop production. Agricultural Forest and Meteorology 103, 99–108.

Soil Survey Staff, 1951. Soil Survey Manual, US Department of Agriculture Handbook 18. US Government Printing Office, Washington, DC.

Sousa, V., Pereira, L.S., 1999. Regional analysis of irrigation water requirements using kriging. Application to potato crop (*Solanum tuberosum* L.) at Tras-os-Montes. Agricultural Water Management 40, 221–233.

Steven, M.D., Biscoe, P., Jaggard, K.W., 1983. Estimation of sugar beet productivity from reflection in the red and infrared spectral bands, International Journal of Remote Sensing 4, 325–334.

Tarantola, S., Giglioli, N., Saltelli, A., Jesinghaus, J., 2000. Global sensitivity analysis for the quality assessment of GIS-based models. In: Heuvelink, G.B.M., Lemmens, M.J.P.M. (Eds), Accuracy 2000. Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Amsterdam, July 2000, pp. 637–641.

Thornburn, T.H., Morse, R.K., Liu, T.K., 1966. Engineering Soil Report, Bulletin 482, Livingston County, Illinois. Engineering Experimental Station, University of Illinois, Urbana.

Tietje, O., Hennings, V., 1996. Accuracy of the saturated hydraulic conductivity prediction by pedotransfer functions compared to the variability within FAO textural classes. Geoderma 69, 71–84.

Tietje, O., Tapkenhinrichs, M., 1993. Evaluation of pedotransfer functions. Soil Science Society of America Journal 57, 1088–1095.

Tomasella, J., Hodnett, M.G., 1997. Estimating unsaturated hydraulic conductivity of Brazilian soils using soil-water retention data. Soil Science 162, 703–712.

Tomasella, J., Hodnett, M.G., 1998. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Science 163, 190–202.

Tomasella, J., Hodnett, M.G., 2004. Pedotransfer functions for tropical soils. In: Pachepsky, Y., Rawls, W.J. (Eds), Development of Pedotransfer Functions in Soil Hydrology. Development in Soil Science, Vol. 30, Elsevier, Amsterdam, pp. 415–429.

USDA Natural Resource Conservation Service, 1994. National Soil Pedon Database. USPA, Lincoln, NE.

Van Genuchten, M.Th., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Science Society of America Journal 44, 892–898.

Veldkamp, A., Lambin, E.F., 2001. Predicting land-use change, Agriculture, Ecosystems and Environment 85, 1–6.

Vereecken, H., Maes, J., Feyen, J., Darius, P., 1989. Estimating the soil moisture retention characteristics from texture, bulk density and carbon content. Soil Science 148, 389–403.

Voltz, M., Webster, R., 1990. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. Journal of Soil Science 41, 473–490.

Webster, R., Beckett, P.H.T., 1968. Quality and usefulness of soil maps. Nature 219, 680–682.

Wolf, J., 1993. Effects of climate change on wheat production potential in the European Community. European Journal of Agronomy 2, 281–292.

Wösten, J.H.M., Nemes, A., 2004. Pedotransfer functions for Europe. In: Pachepsky, Y., Rawls, W.J. (Eds), Development of Pedotransfer Functions in Soil Hydrology. Development in Soil Science, Vol. 30, Elsevier, Amsterdam, pp. 431–435.

Wösten, J.H.M., Finke, P.A., Jansen, M.J.W., 1995. Comparison of class and continuous pedotransfer functions to generate soil hydraulic characteristics. Geoderma 66, 227–237.

Wösten, J.H.M., Lilly, A., Nemes, A., Le Bas, C., 1999. Development and use of a database of hydraulic properties of European soils. Geoderma 90, 169–185.

Wösten, J.H.M., Pachepsky, Y.A., Rawls, W.J., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. Journal of Hydrology 251, 123–150.

Yun, J.I., 2002. Predicting regional rice production in South Korea using spatial data and crop-growth modeling. Agricultural Systems 77, 23–38.

Exercises

1. The characteristics of a soil are given in the table below (OC: organic carbon content; D_b: bulk density).

Horizon	Depth (cm)	Particle size distribution (%)			OC (g 100 g^{-1})	D_b (g cm^{-3})
		Clay	Silt	Sand		
Ap	0–25	18	45	37	2.2	1.21
E	25–55	19	42	39	0.8	1.33
BT	55–90	28	33	39	0.6	1.45

(a) Estimate the volumetric water content (θ in cm^3 of water per 100 cm^3 of soil) at a pressure head of 100 cm (θ_{100}) and 15 000 cm ($\theta_{15\,000}$) for every horizon using (i) the continuous pedotransfer functions (continuous-PTFs):

$$\theta_{100} = 27.9 + 0.41\,(\%\text{clay}) + 0.15\,(\%\text{silt}) - 8.32\,(D_b)$$
$$\theta_{15\,000} = 1.48 + 0.84\,(\%\text{clay}) - 0.0054\,(\%\text{clay})^2$$

that predict points of the water retention curve, and (ii) another continuous-PTF that predicts the entire curve of the van Genuchten model:

$$\theta = (\theta_s - \theta_r)[1 + (\alpha h)^n]^{-1} + \theta_r$$

with

h = pressure head in cm

$\theta_s = 0.81 - 0.283\,(D_b) + 0.001\,(\%\text{clay})$

$\theta_r = 0.015 + 0.005\,(\%\text{clay}) + 0.014\,(\%\text{OC})$

$\alpha = \exp[-2.486 + 0.025\,(\%\text{sand}) - 0.351\,(\%\text{OC}) - 2.617\,(\text{D}_b) - 0.023\,(\%\text{clay})]$

$n = \exp[0.053 - 0.009\,(\%\text{sand}) - 0.013\,(\%\text{clay}) + 0.00015\,(\%\text{sand})^2]$

(b) Assuming θ at field capacity (θ_{FC}) and θ at the wilting point (θ_{WP}) equal to θ_{100} and $\theta_{15\,000}$ respectively, estimate the available water capacity (AWC) of every horizon in millimeter of water per centimeter of horizon and with the two types of continuous-PTFs with:

$$\text{AWC} = \theta_{FC} - \theta_{WP}$$

(c) Estimate AWC for the entire soil in millimeter of water.

2. The water content at 100 cm (θ_{100}) and 15 000 cm ($\theta_{15\,000}$) was measured. The results are given in the table below.

Horizon	Water content (in cm^3 per 100 cm^3 of soil) at:	
	100 cm (θ_{100})	15 000 cm ($\theta_{15\,000}$)
Ap	32.5	15.9
E	31.2	14.6
BT	33.3	20.8

Compute AWC for the entire soil. Which is the more accurate PTF?

3. The objective here is to determine the sowing dates of sunflower crops at Mideulville, which is a location situated at equal distance between the villages of St Bonnet-le-Froid and Villesèque in France.
    (a) Determine first the possible sowing dates between April 1 and June 15, 2005, using two methods; (i) A decision rule based on rainfall data only. (ii) Using a soil water balance. In each case do the calculations using (iii) weather data of St Bonnet-le-Froid, (iv) weather data of Villesèque, (v) weather data interpolated for Mideulville.
    (b) Determine actual sowing dates in cases (i, iii), (ii, iii), (i, iv), (ii, iv), (i, v) and (ii, v) above as the median value of all possible sowing dates.
    (c) From the calculations of possible sowing dates using methods (i) and (ii), determine effective sowing dates at Mideulville using the SIMSEM method, assuming the study area includes 3 farm types: CLF1, FC1 and FC2 (cf. Table 5).

Data necessary for calculations:

- Precipitation data for St Bonnet-le-Froid and Villesèque from March 15, 2005. When a date is not in the table, it means "no precipitation."

	Precip_stbonnet	Precip_villesèque
18/3	5	0
19/3	30	25
20/3	20	15
28/3	10	8
29/3	5	0
5/4	5	0
9/4	1	0
10/4	2	0
11/4	1	0
17/4	14	10
18/4	18	15
19/4	9	5
20/4	12	10

21/4	15	12
28/4	14	10
29/4	10	8
30/4	10	7
1/5	5	5
5/5	1	1
9/5	2	2
18/5	25	0
19/5	2	0
30/5	10	9
31/5	12	8
1/6	14	8
2/6	5	2

- Decision rule based on precipitation data

  If the precipitation on day d is less than 5 mm and the total precipitation for the previous 3 days is less than 9 mm, then sowing is possible on day $d + 1$.

- Parameterization for water balance

  We assume that the soil is at maximum available water capacity (AWC) on March 15. The AWC of the top soil layer is 10 mm, the DAW (difficultly accessible water) is 2 mm. The mean daily potential evapotranspiation (PET) for the period considered is 3 mm. Water balance is calculated as follows: $\mathrm{WC}_d = \mathrm{WC}_{d-1} + P_d - \mathrm{PET}_d$, where P is precipitation, WC is the water content of the top soil layer and all variables are expressed in millimeter. WC cannot be lower than DAW and cannot exceed AWC. Any surplus is lost (drained).
  A day is considered as possible for sowing when WC is lower than or equal to 80% of AWC.

- Local data for the calculation of effective sowing dates

Sunflower is sown after maize, and for both the earliest possible sowing date is April 20.

Man-hours required for various tasks

Task	Required time
Sowing	3 h/ha
In loam soil	1.5 h/ha
In clay soil	
Care of livestock	4.5 h/30 cows
Total working time in a day (in sowing period)	12 h

To simplify calculations, round-off the area sown per day to the nearest hectare and assume that a farmer does not sow two different crops on the same day.

4. Model errors for 3 years for production for 4 randomly chosen fields in a region are given in the tables below. The data in the tables were not used for model development. Surface areas per field are also given. The total surface area of fields of interest in the region is 10 000 ha.

   Case 1

		Year 1	Year 2	Year 3
	Area (ha)	$P - \hat{P}(t)$	$P - \hat{P}(t)$	$P - \hat{P}(t)$
Field 1	10	0	−2	3
Field 2	15	2	−1	2
Field 3	12	1	0	3
Field 4	25	2	−1	1

   Case 2

		Year 1	Year 2	Year 3
	Area (ha)	$P - \hat{P}(t)$	$P - \hat{P}(t)$	$P - \hat{P}(t)$
Field 1	10	0	3	−2
Field 2	15	−1	2	2
Field 3	12	1	0	3
Field 4	25	2	−1	1

   (a) Estimate MSEP for regional production using first the data for case 1 then the data for case 2.

   (b) What is mean squared error, averaged over years and fields, for yield for the two cases? On the basis of (a) and (b), comment on the differences between cases 1 and 2.

5. (a) Suppose that a model systematically overestimates yield. Which term in Eq. (6) does this concern? Does this affect MSEP for regional production?

   (b) Suppose that a model is on the average unbiased, but systematically underestimates yield in years with plentiful rainfall and overestimates under water stress conditions. Suppose further that for the region in question, plentiful rainfall or stress, concerns essentially all the fields of the region. Which term in Eq. (6) does this concern? Does this affect MSEP for regional production?

   (c) Suppose that a model is on the average unbiased, and is further unbiased in each year, but has fairly large errors for individual fields. Which term in Eq. (6) does this concern? Does this affect MSEP for regional production?

Section II

Applications

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 8

Introduction to Section II

J.W. Jones, D. Makowski and D. Wallach

1. **Introduction** **251**
2. **Understanding and improving model behavior** **253**
3. **Crop model applications** **253**
4. **Other applications** **255**
References **255**

1. Introduction

The number of crop models developed, published and used has increased greatly over the last thirty years. This same trend has occurred in many other fields as rapidly increasing computing power has made it feasible to study increasingly complex natural or man-made systems. Reasons for developing these models vary considerably, depending on the specific objectives of those who develop them. However, it is useful to consider two broad categories of reasons.

One reason is that model development, as a way of synthesizing knowledge about different aspects of a cultivated field, is viewed as a system. Indeed, crop models have been developed to see whether our knowledge about the physiology of plant growth processes could be combined to predict crop growth and yield response to various climate, soil and management factors (Boote et al., 1996). The second category of reasons relates to the objective of predicting crop performance, aiding in managing crops or aiding in making decisions related to the impact of agriculture on the environment. Models are highly useful tools in agronomy where the effects of management actions depend on complex interactions with soil, climate and other management decisions. For such reasons, crop models are not the end objective; instead they are used as tools to achieve more applied objectives related to cropping systems and their management.

Methods for developing crop models are becoming more widely known and used, and books exist that describe how to conceptualize crops as systems and to develop models that predict their behavior (e.g. de Wit, 1978; Dent et al., 1979; van Keulen and Wolf, 1986; Penning de Vries et al., 1989; Thornley and Johnson, 1990; Peart and Curry, 1998). Chapter 9 gives an overview of crop models. It presents the processes that are generally modeled and the principles behind the model equations. The overview is illustrated using a small number of particular crop models.

Methods for working with dynamic crop models have not received the same attention as those for model development. Thus, the focus of this book is on methods that are useful in understanding, improving and applying the crop models for various purposes, not for developing them. The methods described in the previous section, mostly developed for use in other fields of study, were selected because of their value in working with crop models. This second section concerns specific models or models for specific purposes. The objective of this section is threefold; to illustrate the diversity of systems and problems that has been studied using crop models or similar models, to show that the methodological problems treated in the previous section arise in a wide variety of situations and finally, to illustrate some of the methods presented or discussed in Section I. Table 1 summarizes the objectives of chapters in this section and the methods demonstrated in them. The models selected obviously represent only a small fraction of studies that has been reported.

Table 1. Objectives and methods represented in Section II.

Chapter	Objective of the study	Methods
9	Overview of crop models	–
10	Overview of applications and problems associated with genotypic parameters	–
11	Overview of uses of models for genetic improvement of crops	–
12	Develop and evaluate a corn model to be used for studying irrigation strategies	Model evaluation (MSEP, cross validation). Parameter estimation (weighted least squares).
13	Evaluate a model to be used for predicting kiwifruit characteristics as a function of management	Model evaluation (Kolmogorov–Smirnov statistic)
14	Analyze a model for predicting soil denitrification	Sensitivity analysis (Latin hypercube sampling, first-order sensitivity index)
15	Analyze a model for predicting soil nitrogen transport and transformation	Sensitivity analysis (one-at-time method)
16	Analyze a model for predicting gene flow between fields	Sensitivity analysis (analysis of variance)
17	Use remote-sensing data to estimate parameters and input variables of a crop model with a view toward precision agriculture	Parameter estimation (GLUE). Obtaining model inputs for multiple situations, data assimilation
18	Combine yearly measurements and model predictions to improve accuracy of estimates of soil carbon	Data assimilation (extended and ensemble Kalman filters)
19	Combine a corn crop model and a decision model to propose improved irrigation management	Decision optimization (dynamic programming, reinforcement learning)
20	Determine management strategies adapted to wheat for ethanol production	Decision optimization (multiple criteria analysis)

Nevertheless, these studies demonstrate a wide range of model types, objectives and methods representative of the range reported in the literature. The purpose of this chapter is to introduce these studies and put them in the more general context of model applications.

2. Understanding and improving model behavior

Whether the model is aimed at better understanding or at management, it is important to evaluate and analyze the model.

Model evaluation is important for all models, and is not specific to dynamic crop models. Indeed, many of the methods and criteria described in Chapter 2 come directly from statistics and apply equally to simple linear models. Application to complex dynamic models may however present practical difficulties. Two examples of model evaluation are given in this section. Chapter 12 considers evaluation of a corn model. The evaluation involves the use of cross-validation to estimate the mean squared error of prediction. Also, there is a rather thorough graphical examination of model errors, with a view to identifying specific problems. Chapter 13 presents an evaluation of a kiwifruit model. This example has the particularity that the output of major interest is the distribution of fruit sizes, not the mean value that is the usual variable of interest.

A basic element of model analysis is elucidating the relationship between inputs and outputs. This may be trivial for simple models but not for highly complex models. Crop models are often in this category. Chapter 3 presented the methods of sensitivity analysis, which allows one to analyze complex models. Case studies are presented in this section to demonstrate useful applications of those methods along with discussions on how the methodology was adapted for each. Chapter 14 demonstrates the use of sensitivity analysis using a static denitrification model. Chapter 15 describes the use of sensitivity analysis for a more complex and dynamic model that simulates nitrogen transformation and transport in soils. Chapter 16 concerns a model to predict gene flow in cropping systems. The model potentially requires a great amount of input data, and sensitivity analysis can be used to determine which inputs are truly essential. In Chapter 17, sensitivity analysis is used to determine which parameters and input variables should be estimated using remote-sensing data.

3. Crop model applications

3.1. Prediction

A major use of crop models is for prediction. The determinants of predictive quality and methods of evaluating predictive quality are discussed in Chapter 2. In this section, Chapter 12 is particularly concerned with the quality of prediction of a corn model.

A very promising method for improving prediction quality is by injecting information from the situation of interest into the model. The techniques of data assimilation are presented in Chapter 5. Chapter 18 describes the use of data assimilation techniques for predicting soil carbon.

3.2. Determining optimal management

There are several methods for using models for crop management. A single simulation run using a model predicts the outcome of a particular set of management decisions. To furnish decision support based on a model, one needs to explore a range of possible decision options. Three major approaches are found in the literature:

(a) Evaluating a range of options using simulation. One simulates the consequences of a range of decisions, which allows one to identify those decisions which are the most satisfactory. This assumes that the criteria for judging the quality of a decision, and the range of conditions of interest, are specified. In simple cases one can specify in advance the range of decisions to be tested. In more complex cases with a very wide range of options, the model user may use results from previous simulations to decide which management decisions to test next.
(b) Optimization. One defines an objective function, and then uses an algorithm to perform a directed search for the management decisions that maximize that objective (see Chapter 6). This is simply a more automated version of evaluating options. An example demonstrating the use of optimization is given in Chapter 19, which concerns irrigation management. Note that this is a particularly complex problem, because there can be several irrigation dates and it is the full set of those dates and the associated amounts that one wants to optimize.
(c) Identifying acceptable practices. It may be more informative to provide a set of acceptable management practices rather than a single optimal decision. The study on wheat for ethanol in Chapter 20 takes this approach. Furthermore, this study involves multiple criteria, which are treated individually rather than being lumped into a single composite objective function.

In order to analyze different decisions with a crop model one must first represent those decisions mathematically, and the details of this representation will have an important impact on the results, as explained in Chapter 6. Chapters 19 and 20 use decision rules, which relate decisions to the state of the system and to external variables.

3.3. Large spatial-scale applications

A model can be used for different fields within a region in order to evaluate the cumulative impact or results of agriculture on a regional scale. Chapter 7 discusses the problems and approaches to using a model for multiple fields. There are many examples of this type of use in the literature (to predict total yield for a region or country, for climate change impact assessment, to predict the water required for irrigation within an area served by a water provider, etc.). A comparable problem arises in precision agriculture. Here only a single field is involved, but it is divided into multiple units which must be characterized. In this section, Chapter 17 concerns the use of crop models for precision farming applications.

Running a model for different fields in a geographic area can also be necessary to explore transfers (of water, pollen, insects) between fields. An example involves the transfer of pollen between fields with genetically modified crops and surrounding fields with non-modified crops (see Chapter 16).

A crop model adapted for multiple fields coupled to a hydrological model can be used to evaluate the impact of agriculture on water at the basin-scale drainage. For example, one could estimate the contribution of agriculture to nitrate pollution of stream water. Going further, one could search for spatially distributed management practices that would reduce pollution.

3.4. Characterizing plant varieties and plant breeding

Two different situations can be identified. If the variety already exists, the problem is to determine in which situations the farmer should choose that variety, and if the variety is chosen, what management practices should be adopted. Field experimentation is not well suited to this problem, which involves a large number of factors concerning soil, climate and management. Furthermore, in the early period after a variety becomes available the amount of experimental information will necessarily be limited. Crop models, on the other hand, are well suited to taking into account the multiple input and output variables of importance. Of course, here as well as elsewhere the model also presents difficulties. The problem here is to adapt the model to each specific variety. A standard approach is first to identify and then estimate the variety-specific parameters in the model (see Chapters 10 and 11).

The second situation is during the variety selection process. A major question here is the importance of different traits which can be selected for. In principle, a model could aid in analyzing the role of a particular trait, but this requires knowing the relation between the model parameters and traits that can be selected for. Now that molecular genetic information is becoming available, the problem becomes that of relating this genetic information to model parameters. Adapting crop models for plant breeding is an important and promising new area. For example, Messina et al. (Chapter 11) used the simulated annealing optimization method to determine optimal traits of soybeans for different environments based on genetic information.

4. Other applications

There are many applications of crop models that are not illustrated in this section. A particularly important and promising application that merits mention is the use of crop models as educational tools, whether for university students, extension advisors, farmers or researchers.

However, regardless of the application, there will be a need for methods of evaluating, analyzing, improving and using crop models. The case studies in this section show the ubiquity of the need and the diversity of applications.

References

Boote, K.J., Jones J.W., Pickering, N.B., 1996. Potential uses and limitations of crop models. Agronomy Journal 88(5), 704–716.

Dent, J.B., Blackie M.J., Harrison, S.R., 1979. Systems Simulation in Agriculture. Applied Science Publishers Ltd, London, pp. 180.

de Wit, C.T., 1978. Simulation of Assimilation Respiration, and Transpiration of Crops. Halsted Press, Division of John Wiley and Sons, New York.

Peart, R.M., Curry, R.B. (Eds), 1998. Agricultural Systems Modeling and Simulation. Marcel Dekker, Inc., New York.

Penning de Vries, F.W.T., Jansen, D.M., ten Berge, H.F.M., Bakema, A., 1989. Simulation of Ecophysiological Processes of Growth in Several Annual Crops. Pudoc, Wageningen.

Thornley, J.H.M., Johnson, I.R., 1990. Plant and Crop Modeling: A Mathematical Approach to Plant and Crop Physiology. Oxford University Press, New York, Second Printing by Blackburn Press.

van Keulen, H., Wolf, J. (Eds), 1986. Modeling of Agricultural Production: Weather, Soils and Crops. Pudoc, Wageningen.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 9

Fundamental concepts of crop models illustrated by a comparative approach

N. Brisson, J. Wery and K. Boote

1. Introduction **257**
2. Concepts and their representation **263**
3. User environment **272**
4. Discussion **275**
References **276**

1. Introduction

The purpose of this chapter is to provide an overview of the physiological and physical processes described by crop models and, to a lesser extent, of certain computer aspects of crop models. We begin with some general remarks about crop models, which will help to specify what a crop model is. Then we discuss several fundamental concepts on which crop models are based, and illustrate them with the comparison of five different models (AZODYN, CERES-EGC, CROPGRO, CROPSYST and STICS).

1.1. Some characteristics of crop models

Crop models represent the dynamic functioning of the soil–plant system as it interacts with climate and farming practices. Obviously, crop models do not simulate all processes occurring within the soil–plant system, and the "crop model" approach is based on the more important of these processes. The processes actually taken into account will depend on the output variables and the conditions of interest. For example, such models do not simulate how the crop affects the atmosphere, in contact with the crop. This is the area covered by Soil Vegetation Atmosphere Transfer (SVAT) models (Olioso et al., 1999).

The spatial simulation unit is, in general, the farmed plot or field, but may also be an element of the field (as in the case of precision farming). It is assumed to be homogeneous in terms of soil, climate and management practices. Most of the concepts applied below are based on this hypothesis (soil water balance, crop development and leaf area index (LAI), for example). In practice, of course, field homogeneity is unlikely to occur. Consequently, the model parameters and the variables that describe the state of the system represent averages. It is important to keep this in mind when estimating parameter values and for model evaluation.

The time step for calculating new system state variable values is in general one day, because weather variables are usually available as daily weather inputs. Although one day constitutes the temporal resolution of the processes simulated, that is not the time step for validation. Validation normally concerns how the system functions from the starting date of the simulation through a series of different measurement dates, up to final harvest or maturity.

The state variables of a crop model are the variables that describe the system. The list of state variables differs between models, but the principal state variables are often aboveground biomass, LAI, weight of harvestable product, soil water and soil nitrogen contents. As a consequence, crop models generally have modules whose equations describe development, LAI dynamics, biomass accumulation, yield formation, and water and nitrogen balance.

To run a crop model, it is necessary in general to enter information on the soil, climate, management sequence and initial status of the system. These characteristics are specific to the plot being simulated and, for the initial conditions in particular, may depend on the detailed history of the plot (for example, the nature and management of the preceding crop). They thus require *in situ* measurements. These different input data are sometimes known with a degree of imprecision that is difficult to estimate, and which may propagate to the output variables.

The output variables are state variables or functions of state variables. They include agronomic variables (date of harvest, yield, nutritional quality, water and nitrogen consumption) and environmental variables (leaching of nitrate, etc.). Intermediate variables (biomass, LAI, stress indices, etc.), calculated for example at certain phenological stages may also be important output variables.

Crop models are compartmental in design, the compartments being linked by flows of matter that are driven by energy and that depend on the flow of information. Figure 1 illustrates the flow of carbon in the plant system, driven by radiation energy. Depending on the crop model, the root system can be identified or not. Hydraulic forces and thermal gradients drive flows in the soil system. In general, crop models consider at least flows of water, mineral nitrogen and organic matter. Developmental stages and various stresses act as additional information, which affect the carbon functioning of the system.

Stresses (only of an abiotic nature for most crop models) are functions that reduce plant processes (Fig. 2). They depend on stress variables (fraction of transpirable soil water, nitrogen nutrition index, fraction of root system in waterlogged conditions, etc.), which must then also be calculated. The reduction functions are empirical relationships based on the limiting factor principle (Gary et al., 1996). Nonetheless, they are based on what we know about the effects of these stresses on plant growth and development. For example, water stress acts *via* a hormonal or hydraulic signal on stomatal conductance, which causes a reduction in photosynthesis and hence in radiation use efficiency (RUE). The empirical function relates the reduction in RUE directly to water stress. Similarly, water stress diminishes cell division and expansion, phenomena which cause a reduction in the appearance and expansion of leaves and hence in the rate of increase of LAI. The empirical function then directly relates the reduction in LAI increase to water stress. On the other hand, the regulation involved in interactions between stresses is poorly understood at the scale of the whole plant, and is therefore modelled very simply by using the product or the minimum of reduction factors. More physiological approaches (Farqhuar et al., 1980,

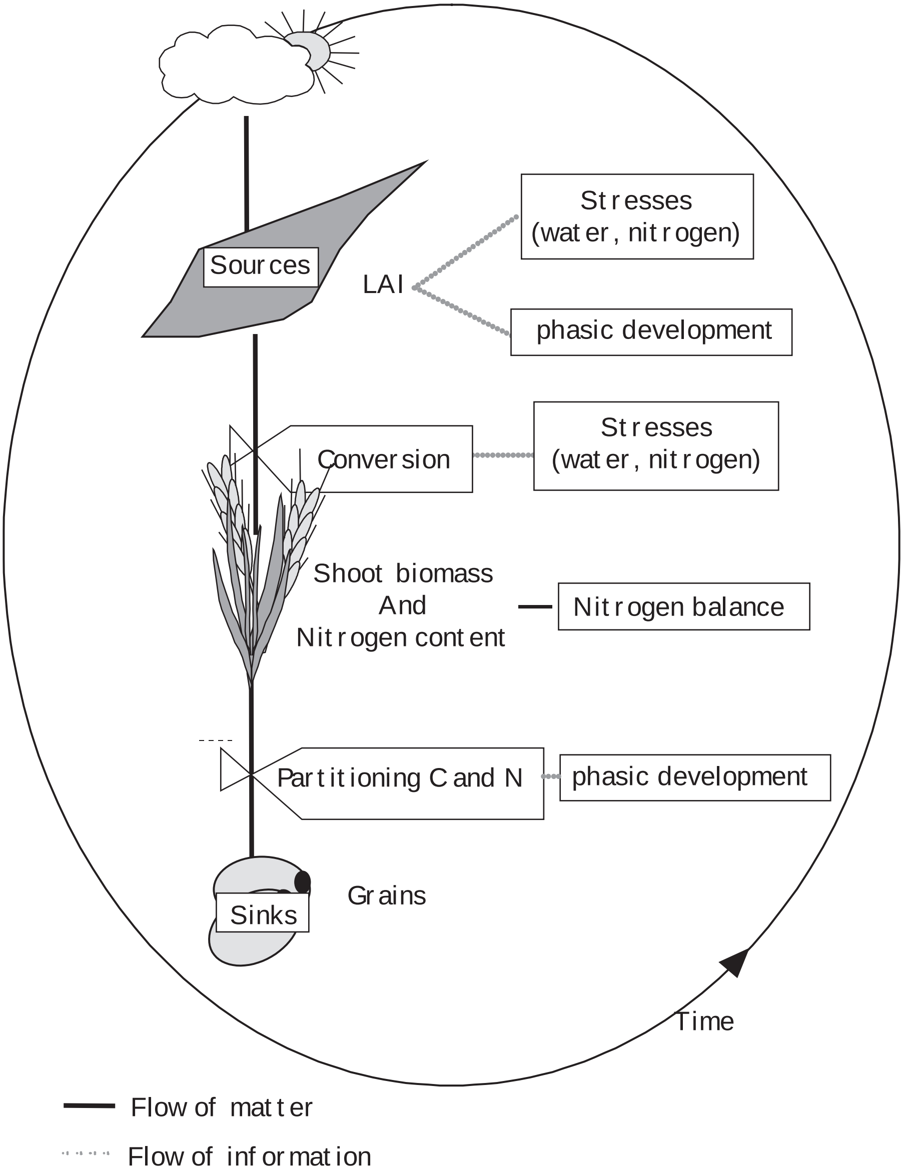


Figure 1. Schematic representation of the main flows of matter and information in the crop models.

for example) could lead to more realistic models for photosynthetic processes, but raise the problem of parameterization.

1.2. Some history

Crop models arose from the work on photosynthesis and respiration by the early 1970s (de Wit et al., 1970), and the establishment of a link between accumulated biomass and

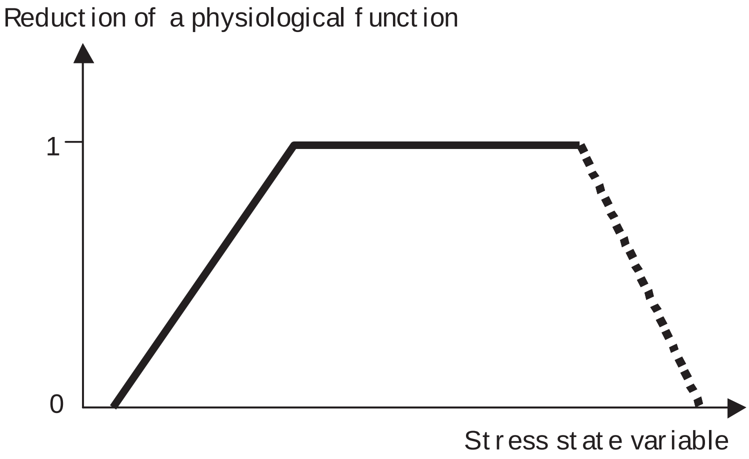


Figure 2. Principle of the limiting factor applied within crop models.

instantaneous processes. Nearly simultaneously relationships were established between canopy architecture and photosynthesis (Baker and Meyer, 1966; Duncan, 1971), using the notion of LAI (introduced by Watson in 1947). In some cases, these complex processes were simplified (Monteith, 1972) and gave rise to new more universal concepts such as the "radiation use efficiency (RUE)" (Spaeth and Sinclair, 1985; Sinclair, 1986), which would ultimately serve as an alternate approach for crop models.

The Dutch school (de Wit, 1978; van Ittersum et al., 2003) produced models derived from SUCROS, aimed at describing in detail the ecophysiology of crops, often for didactic purposes. The British ARCWHEAT model (Weir et al., 1984) can also be placed in this category as well as Duncan's pioneer models for cotton, corn, soybean and peanut (cited in Baker, 1980). The American school, with its GOSSYM/GLYCIM (Whisler et al., 1986; McKinion et al., 1988), CERES (Ritchie and Otter, 1984), and CROPGRO (Boote et al., 1998; Jones et al., 2003) models placed more emphasis on agronomic objectives and started to include farming practices in the inputs. At the same time, EPIC (Williams et al., 1984), the first generic model, was developed in response to agro-environmental concerns including soil erosion, water and nitrogen. CROPSYST (Stöckle et al., 1994, 2003) arose out of EPIC. During the early 1990s, models with a specific environmental objective were introduced, such as DAISY (Hansen et al., 1990) and PASTIS (Lafolie, 1991). APSIM, the Australian model, was derived from CERES (McCown et al., 1996), as was CERES-EGC, which is a modification and extension of CERES to include additional environmental concerns such as the soil nitrogen balance (Gabrielle et al., 1995).

A somewhat different approach is that of the AZODYN model (Jeuffroy and Recous, 1999; Jeuffroy et al., 2000), which is specifically oriented toward nitrogen fertilization and uses diagnostic data (data recorded in the field at a precise date) as input. The distinction is not sharp, but roughly speaking some of the model families listed in Figure 3 are more mechanistic concerning processes in crop or soil sciences, while others are more empirical. These latter models are oriented toward agronomic or environmental objectives. They synthesize knowledge from a range of disciplines. The current trend is

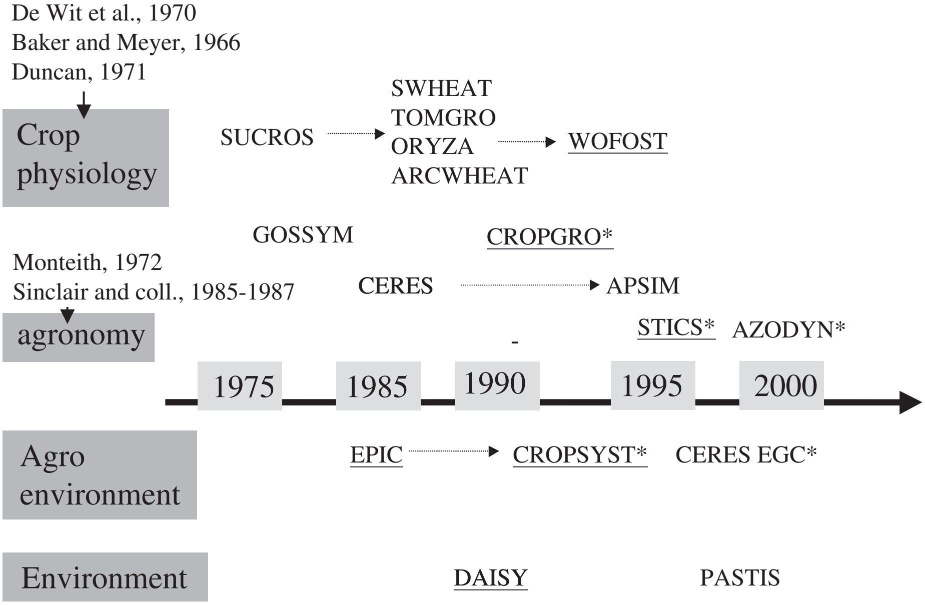


Figure 3. Chronology of crop modeling: underlined – generic models and * – models described in detail in this presentation.

towards developing generic and agro-environmental models that take into account farming practices, an example being STICS (Brisson et al., 1998, 2003). The generic nature of a model does not preclude crop specificity, but it is indicative of the efforts being made towards a common approach based on agro-physiology.

The most often modeled crops are industry-targeted cash crops in temperate and tropical regions (wheat, maize, soybean, cotton, rice, sorghum, etc.). The reasons include the large surface areas concerned and the fact that these crops are compatible with system simplifications and the limited artificial environmental conditions. The crops are annuals and generally homogeneous in nature.

1.3. Five crop models

We have chosen to analyze in detail five models, which illustrate the range of the concepts applied. Each model has a primary objective, which governs the principal modeling choices.

The most specialized model is AZODYN, which simulates winter wheat only, then CERES-EGC which simulates wheat, maize, rapeseed, sugarcane and sorghum. CROPGRO is mainly devoted to grain legumes (soybean, peanut, various species of bean, pea) but also non-legume crops (tomato, pepper and forage crops). CROPSYST simulates all the industry-targeted cash crops and grasslands while STICS is the most generic model, dealing with all the above-mentioned crops plus lettuce, banana, grapevine, catch crops and intercrops. Nevertheless, in order to simplify the presentation in this chapter, we will concentrate on annual determinate crops assumed to have a homogeneous foliage structure.

In all models, the plant sub-system is characterized by its above-ground biomass and LAI. The identification of organs (in terms of numbers and mass) focuses on harvested organs (except for CROPSYST) and vegetative organs (in the case of CERES-EGC and CROPGRO).

Roots are not identified in AZODYN *per se*, while they are characterized by their biomass and/or length in CERES, CROPGRO and CROPSYST and by their length in STICS. The soil is usually defined as extending to bedrock or the water table or to the practical maximum extent of root penetration. Below the root zone it is assumed that only physical processes (migration of water and nitrates) need be taken into account. In the case of AZODYN, the soil is limited to the root zone. In STICS, it is possible to take account of the effect of buried drainage systems. CROPSYST is the only one to consider capillary rise from the water table.

The simplest model in terms of the number of modules is AZODYN, although it does contain a grain quality module. The small number of modules is allowed by the use of diagnostic data measured *in situ* that eliminates the need for calculations of key variables and to its limited field of validity (wheat in the Paris basin). The AZODYN model does not take into account the initial phases of plant establishment, as it starts at the end of winter and does not require a development module, in part because the date of flowering is an input variable. Nitrate leaching is not considered, which leads to a simplified soil description.

With CROPSYST a marked increase in the number of modules appears, because of its "cropping system" objective (e.g. soil erosion and soil structure are calculated) and its generic character. This makes it necessary to include modules, which apply only to certain groups of crops, for example symbiotic nitrogen fixation. The user can choose different representations of water transfer in the soil, namely a tipping bucket or flux-gradient approach. CROPSYST does not include a yield components module.

The environmental objective of the CERES-EGC model led its authors to use a flux-gradient formalism to describe transfers, including heat. This provides a good basis for the simulation of volatilization and denitrification.

The CROPGRO model is more physiologically oriented using, for example, leaf level photosynthesis calculated at an hourly time step. As it simulates mainly indeterminate crops, it has daily cohorts of pods/fruits and seeds complete with numbers and sink strengths. It has the same soil nitrogen uptake, soil organic matter module and tipping bucket water balance as the CERES models.

STICS simulates a broad range of crops and this is made possible by simulation options within several modules (radiation interception, yield formation, etc.). STICS also includes a microclimate module, which makes it possible to take account of the role of the modified atmosphere surrounding the crop.

As shown in Figure 4, the STICS and CROPSYST models are particularly rich in the management techniques, which can be input, including soil preparation and harvesting decisions. Specific management techniques in tropical environments or regarding high value-added crops are taken into account in STICS (plant mulch (also in CROPSYST), plastic mulch, regulation of LAI, thinning, cold shelter).

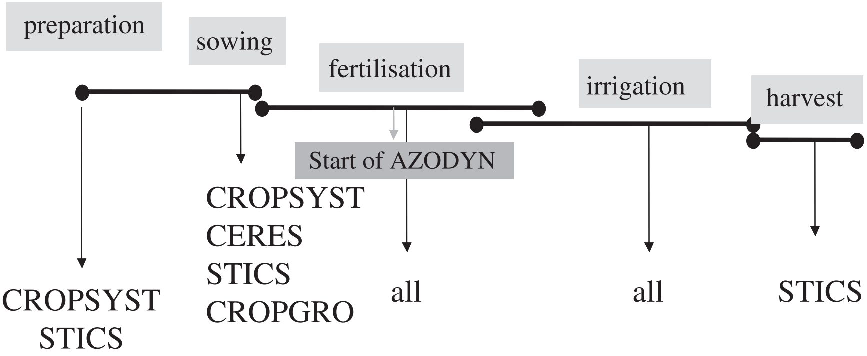


Figure 4. Crop management sequences in AZODYN, CROPSYST, CERES, CROPGRO and STICS.

2. Concepts and their representation

The concepts integrated in crop models correspond to an implicit representation of the soil–plant system in terms of structure and functioning. This "viewpoint" of the system is, in part, a function of the expected utilization of the model. Crop models are not meant to represent in detail the state of knowledge of the functioning of the soil or plant, but rather to provide a synthesis of knowledge in a simplified form. They are based on an implicit ranking of the importance of different mechanisms in terms of the objectives. We have chosen to describe the most widely employed concepts, and then to illustrate them using the reference models. We shall also see that models can employ other concepts, and that such an analysis can lead to defining the area of validity of a model (conceptual validity).

2.1. Crop Development

Crop development is the rate of progress of a plant through its life cycle, i.e. progress from sowing to emergence, emergence to flowering, flowering to maturity, as well as rate of progress between successive leaves on the main axis. In crop models, the same notions are used and the developmental stages correspond to the onset and death of source or sink organs. Crop development is driven by temperature and also influenced by photoperiod. The classical view of crop heat unit accumulation is based on the concept of growing degree-days (Bonhomme et al., 1994). This long accepted concept (Durand, 1967) arises from a linear approximation of the curve for the development rate in response to temperature, as shown in Figure 5. The base temperature (Tb) is the apparent temperature at which the linear approximation gives zero rate of progress toward an event, such as flowering. Thus, strictly speaking, the base temperature (Tb) corresponds to a statistical and not a physiological value. The mean daily temperature is frequently used in this formulation, while other factors affecting the rate of development are modeled as brakes or accelerators on that rate in thermal time (Brisson and Delécolle, 1991). These factors may include the

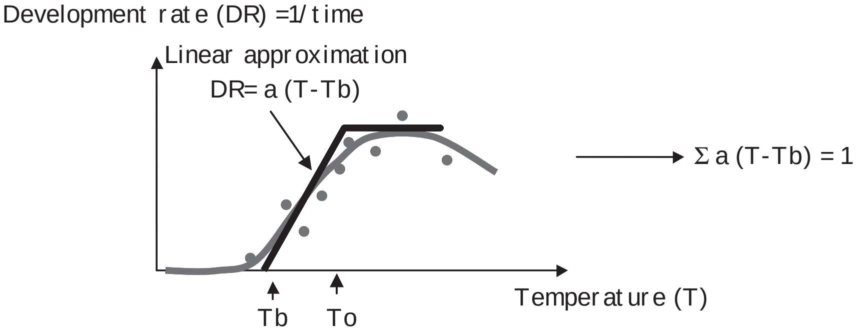


Figure 5. Explanation of the concept of growing degree-days by the representation of the developmental rate or 1/time to a developmental stage (DR) as a function of temperature: experimental points, a curvilinear function fit to those points and an approximation which has a constant slope between a base temperature (Tb) and an optimum temperature (To) (equation DR $= a(T - T\text{b})$) are shown. Using this linear approximation, the accumulation of "development rate" over the developmental stage is 1, which leads to a constant $\Sigma(T - T\text{b}) = 1/a$ value.

photoperiod (CERES, CROPGRO and STICS), vernalization (CERES and STICS) and water deficit (CROPGRO, STICS through the use of the canopy temperature). In the strict degree-day approach above with mean daily temperature, there can be a maximum mean temperature, but possible negative effect of supra-optimum temperatures are not taken into account. The more complex curvilinear function in Figure 5 would allow for negative effects. CROPGRO computes an hourly temperature effect (hourly values are generated from input of daily maximum and minimum) on rate of development that does allow negative effects of supra-optimum temperature. While it is easier to use temperatures under standardized shelter measured by meteorological networks (as is the case for CROPSYST, CERES and CROPGRO), it may be important to move closer to the temperature actually experienced by plants (prediction of canopy temperature, as in STICS).

2.2. LAI dynamics

The canopy is modeled as a homogeneous environment with leaves being uniformly distributed over the land area. Leaf area is described by the leaf area index (LAI) having units of leaf surface area per square meter land area. A consequence of this homogeneous representation is that it allows the use of an optical analogy (Beer's law) to estimate the interception of photosynthetically active radiation (PAR) as shown in Figure 6.

This approach is very successful for homogeneous crops, but poorly suited to canopies in rows or during the first stages of an annual crop because the homogeneity hypothesis cannot apply. The CROPGRO and STICS model have the capabilities to simulate hedgerow canopies in rows, with prediction of light interception dependent not only on LAI, but also on shaded and sunlit leaf area, plant height and width, row spacing, plant spacing, and the direct and diffuse light absorption (Spitters, 1986; Spitters et al., 1986; Boote and Pickering, 1994; Brisson et al., 2004).

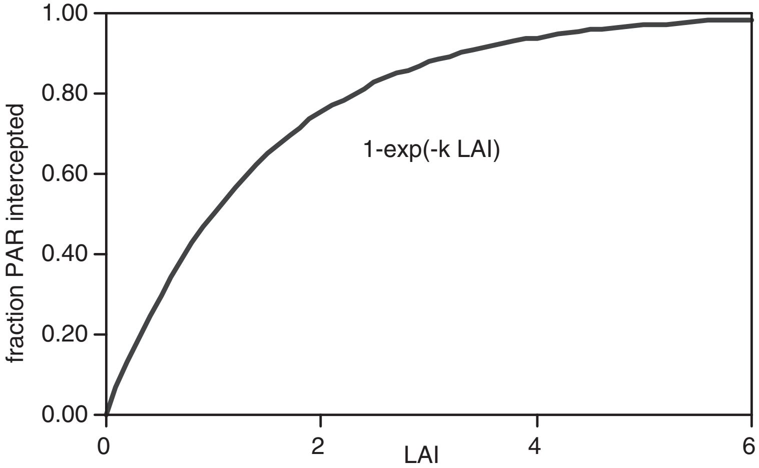


Figure 6. The proportion of intercepted PAR expressed as a Beer's law analogy, *k* being the extinction coefficient.

The specific leaf area (SLA) (ratio of foliage surface to mass) is another concept, which is often employed in crop models. It is used to convert the biomass growth allocated to developing leaves into a surface area, and thus assumes that a portion of the assimilates are allocated to foliage. This approach offers the advantage of integrating some stresses (those acting on biomass) into the control of leaf growth, and in this way mimics the self-regulation of the plant. However, the specific leaf area cannot be considered as a constant, as it is affected by crop age and stresses. For instance, the prediction of SLA in CROPGRO is dependent on direct effects of temperature, water deficit and solar radiation level, as well as crop aging effects. A subset under this approach, assuming constant ratio of leaf area to total vegetative biomass, is sometimes used (as in AZODYN and CROPSYST); however, this assumption causes problems when the stem:leaf ratio evolves. Because of the iterative character of the calculation, leaf area growth computed with the SLA approach is sensitive to initialization (initial LAI).

An alternative to using SLA, is to drive leaf area growth by temperature only (not coupled to biomass) as done in STICS or the Sinclair models. CERES uses a combination of the two approaches successively. CROPGRO uses potential temperature-limited leaf area growth as an upper limit to leaf area expansion only during the period when the first five leaves are produced.

CERES, CROPGRO and STICS allow for the effects of sowing density on increase in LAI, while other models assume that variations in density as applied by farmers are small and that there is no need to introduce this effect, or that inter-plant competition results in similar leaf area indices.

Leaf senescence is approached from the standpoint of lifespan in CROPSYST and STICS. In CERES, daily leaf senescence is a predefined proportion of biomass, while in AZODYN and CROPGRO, senescence depends on the nitrogen status or dynamics of the crop.

Stresses also act on foliage, much more severely than they do on biomass. For this reason, the effects of stresses on biomass are not sufficient to represent the reduction in foliage in the event of stress. Thus water stress (CROPSYST, STICS, CERES, CROPGRO) and nitrogen stress (STICS, AZODYN, CROPGRO) reduce leaf growth and/or accelerate senescence (for AZODYN and CROPGRO *via* the calculation of senescence).

In crop models, LAI is an intermediate state variable, which has a functional role in the sense that it is a variable explaining numerous processes. In the case of CROPSYST, it is the fraction cover (estimated using a parabolic function of the LAI) which is functional.

2.3. Accumulation of biomass

The linear relationship between accumulated biomass in the plant and radiation intercepted by foliage, demonstrated by Monteith (1972), defines the radiation use efficiency (RUE) as the slope of this relationship (Fig. 7). This parameter has become a concept widely employed in crop models (as in AZODYN, CERES and STICS), because it synthesizes (very economically in terms of the number of parameters involved) the processes of photosynthesis and respiration. Its calculation (ratio between above-ground biomass and absorbed radiation) implies that this parameter also takes account of a carbon allocation coefficient between above-ground and below-ground parts of the plant. Obviously, because of underlying physiological processes, this ratio varies in line with stresses, temperature and phenology. To take account of these effects, Sinclair (1986) proposed that RUE should be considered as a physiological function, to which stress indices should be applied (Fig. 7).

In CROPGRO, the photosynthesis and respiration processes are calculated separately. The leaf-level photosynthesis uses rubisco-kinetics from Farquhar et al. (1980), and the growth and maintenance respiration approaches are from Penning de Vries and van Laar (1982).

The efficiency of transpiration (water use efficiency (WUE)) is another synthetic parameter sometimes used to estimate accumulated biomass. It is the ratio of biomass per unit

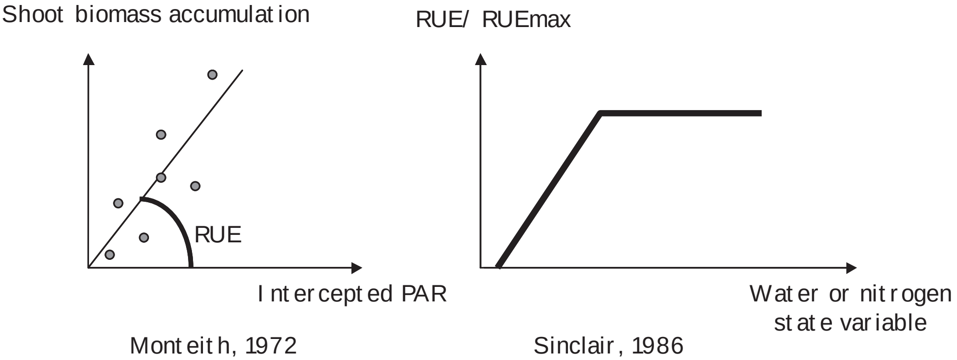


Figure 7. The radiation use efficiency (RUE) is the slope of the linear relationship of shoot biomass accumulation *versus* intercepted PAR (left) that can be modulated by stresses (right).

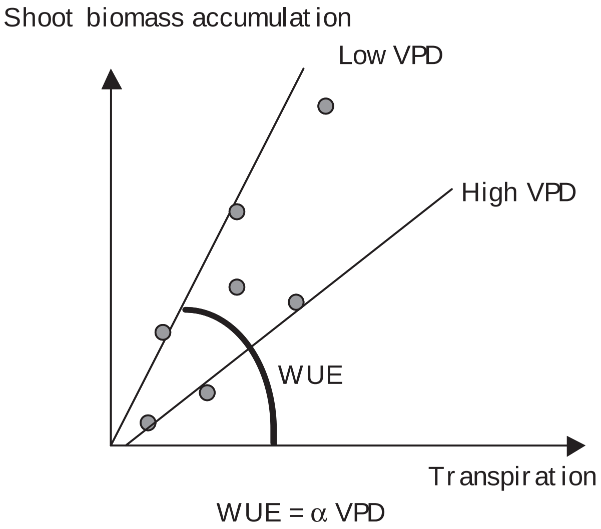


Figure 8. Use of the water use efficiency (ratio of biomass produced per unit of water transpired) in combination with vapor pressure deficit (VPD).

of water transpired. This approach is based on stomatal functioning and must take account of atmospheric moisture (vapor pressure deficit (VPD), in Fig. 8). It is well suited to dry conditions, but raises two problems: the integration of nitrogen stress and the sensitivity of transpiration estimates based on the water balance module. *Via* the intermediary of transpiration, this approach implicitly integrates the driving effects of radiation. CROPSYST uses this WUE parameter in combination with RUE, which acts solely as a threshold.

Factors that limit the accumulation of biomass are numerous: temperature, water and nitrogen in all models, with a double effect of nitrogen stress in CROPSYST (on the calculation of transpiration through a retroactive effect on reducing photosynthesis, and directly on WUE). In addition, the CERES, CROPGRO, STICS and CROPSYST models take account of the effect of CO_2 on RUE or directly on leaf-level assimilation (CROPGRO).

2.4. Dynamics of yield accumulation

Grain yield prediction is a goal of most crop models. The yield and the number of organs harvested are generally calculated independently (until the varietal potential is attained), the elementary weight of these organs being calculated as the ratio between the two.

In 1972, Warren-Wilson proposed that the plant should be considered as a series of compartments playing the role of a source or sink for assimilates. These compartments usually represent organs which can change their function during a cycle: “source and sink” for roots and trunks in perennial plants, or “sink and source” for leaves. Application of this concept to crop models engenders self-regulation of the system between the growth of different types of organs. It is particularly well suited to crops with indeterminate growth and to perennial crops, in which trophic competition exists between vegetative organs and storage organs. Source capacity includes both newly formed assimilates and remobilized resources. Sink strength is usually represented by a continuous or discrete function of the age of the organ. The problems with this approach reside in determining the size of the source capacity and remobilized resources, which are difficult to estimate experimentally. Furthermore, it is often necessary to introduce priorities between organs, thus reproducing the plant's strategy, and this may be speculative. The differential effects of stresses on different plant organs are better understood, but still poorly taken into account by models. One alternative is to impose a constant distribution of assimilates by phenological stage, which is frequently applied in determinate crops. The source–sink approach is used in CERES, CROPGRO and AZODYN in calculating the increase in carbon and nitrogen content of grains.

A second alternative, proposed by Spaeth and Sinclair (1985), is to extend the notion of the final harvest index (HI) (ratio of grain biomass to total shoot biomass) to the dynamic accumulation of biomass in grains, realizing that a linear relationship of the HI as a function of time could be assumed (Fig. 9). This approach has the advantage of globalizing the two sources of assimilates, and is economical in terms of parameters. However, it is important to impose a threshold on this HI dynamic, in order to avoid simulating unrealistic remobilization levels or exceeding the maximum filling allowed by the number of organs (usually fixed beforehand) and the maximum weight of an organ (in a variety). This approach is applied in STICS with respect to the accumulation of carbon (biomass) and nitrogen in grains.

In STICS, CROPGRO and CERES, the number of grains is calculated from genetic parameters and depends on crop growth before flowering (case of cereals) or during the period of pod and seed formation (case of legumes), which is influenced by stresses. In AZODYN, it is the nitrogen status of the plant at flowering which determines the number of grains.

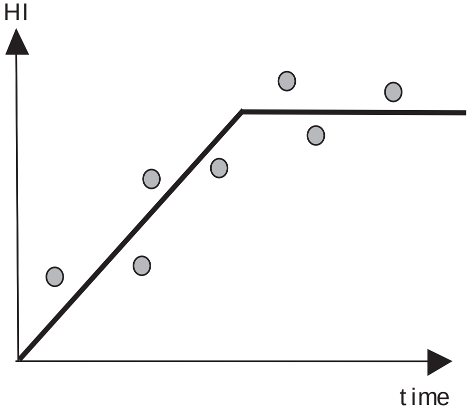


Figure 9. Concept of dynamic harvest index (HI = ratio of grain biomass to total shoot biomass) that can be linearly related to time (Spaeth and Sinclair, 1985) until reaching a maximum.

Whichever approach is used to simulate grain filling, varietal specificities are always taken into account through potential seed size, and thermal (CERES, STICS, AZODYN, CROPGRO) or nitrogen (AZODYN, STICS, CROPGRO) effects may exist.

2.5. *The water balance*

The water balance in crop models has a dual purpose: to estimate soil water content (which, for example, drives nitrogen mineralization of the soil) and water stress indices (which drive the functioning of the plant). The latter objective differentiates crop models from those dedicated to irrigation management, and also forces a distinct separation between evaporation and transpiration. This separation is usually applied at the level of the climatic potential demand (ET_0) based on the partitioning of ET_0 between potential plant transpiration (EP_0) and potential soil evaporation (ES_0). The partitioning (Fig. 10) is generally based on the Beer's law extinction coefficient concept as a function of LAI, although the extinction coefficient is near 0.5 and clearly lower than for photosynthetically active radiation (the latter coefficient is used for canopy assimilation).

This approach is based on estimating ET_0, which comprises both climatic and crop components. However, the ET_0 variable differs from the classical maximal evapotranspiration variable, as defined for example by Itier et al. (1997) because it supposes that all surfaces (soil and foliage) are saturated with water. As for the climatic component, and in view of the problems of accessing meteorological data, models usually propose several calculation choices: Penman-FAO24 (STICS, CERES, CROPGRO), Penman-Monteith-FAO56 (CROPSYST, CERES, CROPGRO) or Priestley–Taylor (CERES, CROPGRO, CROPSYST, STICS). The Penman-FAO24 and Penman-Monteith-FAO56 are described

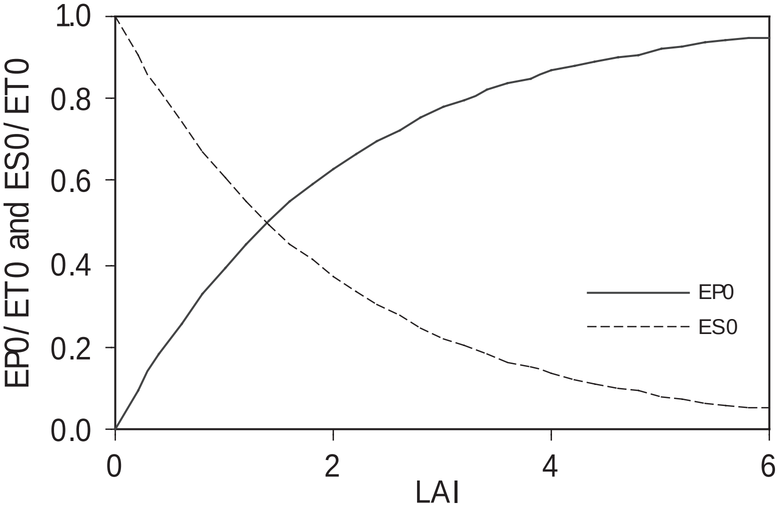


Figure 10. Use of the Beer's law optical analogy applied to LAI to separate maximal evaporation (ES_0) and transpiration (EP_0).

by Allen et al. (1998) while the Priestley–Taylor (1972) option as modified by Ritchie is described by Ritchie (1985). The crop component of ET_0 is usually linked to LAI (STICS, CERES and CROPGRO); it may also be a discrete function of developmental stage (crop coefficients in CROPSYST). In both cases, this crop component takes account of the increase in crop height and its roughness during the cycle, which acts on the degree of the convective component of evapotranspiration. In AZODYN, there is no separation between evaporation and transpiration, and the approach is that of models targeting irrigation (Itier et al., 1997), based on Penman and crop coefficients.

Convection under the plant canopy, which affects maximum transpiration, may be poorly reproduced by this optical analogy (particularly for row canopies); this may justify applying a calculation of the energy balance (optional in STICS).

To calculate the quantity of water actually transpired by the crop, most models are based on a concept which includes the quantity of water physically available in the soil and the capacity of the plant to extract this water thanks to its root characteristics. This is the fraction of transpirable soil water (FTSW) (Sinclair, 1986; Lacape et al., 1998; Pellegrino et al., 2002), which also corresponds to the notion of the maximum available water content (AWC/MAWC) (water amount between the field capacity and the wilting point) (Fig. 11). This approach does not permit a precise localization of root absorption in the soil horizon (at a daily time step, all models hypothesize that transpiration equals absorption), but has the advantage of implicitly taking account of capillary rise within the root zone. However, the threshold of sensitivity may vary over time (root density growth, climatic demand: Brisson, 1998). This global estimate of transpiration is used in AZODYN and STICS, while in CERES and CROPSYST, the calculation of uptake is differentiated in terms of the soil layer (need to simulate capillary rise). This approach, developed originally for the regulation of transpiration, was then extrapolated to a calculation of the functions of water stress for leaf growth or the RUE (Fig. 11). Water uptake per unit root length in CROPGRO and CERES is based on the radial flow equation to roots. Transpiration is limited only if actual root water uptake is less than the crop component of ET_0, although expansive growth is limited at a smaller ratio.

All those water balance modules are sensitive to the holding capacity of the soil, the depth explored by the roots and the climatic evapotranspiration option. Allen et al. (1998)

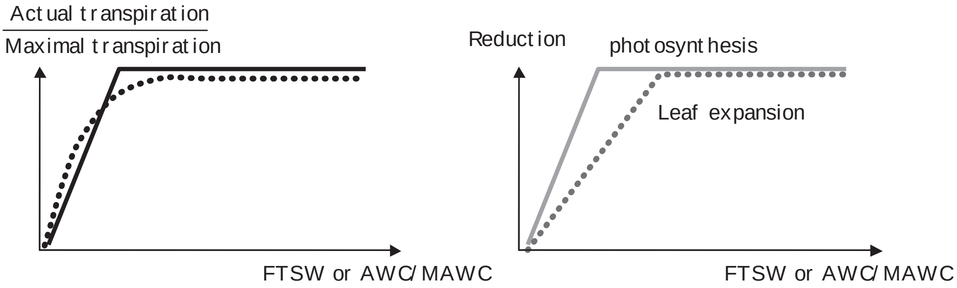


Figure 11. Use of the FTSW concept to calculate actual transpiration using either a bilinear or a curvilinear function (left) and reduction in other physiological functions under the effect of water stress (right).

showed that the Penman-FAO24 predicted too severe water deficits compared to the Penman-Monteith FAO56, and Sau et al. (2004) showed that the Priestley-Taylor function while doing well for mid-range climatic demand conditions, tended to over-predict for cool regions.

2.6. The nitrogen balance

2.6.1. Nitrogen in the plant

Nitrogen requirements depend on biomass accumulation (Lemaire and Gastal, 1997), which generates a close link between carbon and nitrogen dynamics in the plant. The maximum nitrogen accumulation curve defines potential nitrogen accumulation. The critical curve defines the limit of nitrogen concentrations below which the plant restricts its growth. It allows definition of the nitrogen nutrition index (INN) (Fig. 12), which acts as a nitrogen stress state variable in these models. In their construction, feedback effects of nitrogen stress (moderation of nitrogen stress through a slowing in growth) are incorporated in the INN estimate. The presence of reproductive organs may disturb the relationship and it may become inappropriate to use these relationships during a marked nitrogen remobilization stage. Furthermore, although the plateau of the relationship during

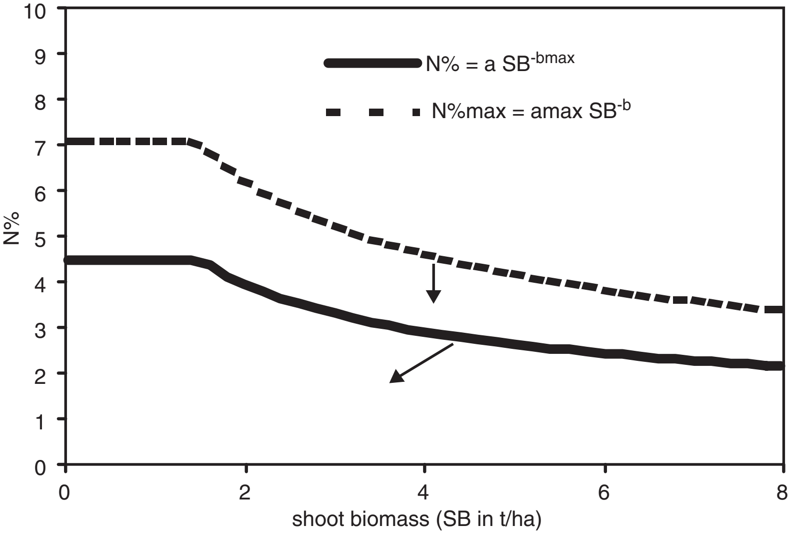


Figure 12. Nitrogen requirement and nitrogen nutrition index (INN) calculation: the nitrogen content of the plant (N%) is inversely proportional to its biomass and there are two functional curves driving this nitrogen behavior. Below the critical curve (solid line) the crop suffers from nitrogen deficiency and its biomass is reduced, while between the critical curve and the maximal curve (dotted line) there is luxury consumption and nitrogen status has no impact on biomass. The maximal curve corresponds to the nitrogen requirement while the critical curve allows the calculation of the nitrogen nutrition index (INN) as the ratio of the actual N% to the critical N%.

the early stages of the crop (fixed nitrogen contents up to a biomass threshold) has little effect when these relationships are used for diagnosis in the field, it may become very important in the context of modeling throughout the cycle, as it determines early plant growth. This approach is adopted by all reference models except CROPGRO, with a few variants for the estimation of nitrogen requirements: maximum curve until flowering, then a function of the INN (STICS), maximum curve with recovery of the possible deficit (difference between maximum and critical curves) (CROPSYST), critical curve × 1.25 until flowering and then an estimate of the nitrogen sink strength of grains (AZODYN), critical curve as a function of development (and not of biomass) and action of stresses (CERES).

CROPGRO has constant maximum and minimum thresholds of nitrogen concentration for each vegetative tissue type during vegetative growth, with a temperature-dependent nitrogen mobilization rate from vegetative tissues that is slow during vegetative growth and accelerates during reproductive growth after grain growth begins. It similarly computes nitrogen stress effects on growth when nitrogen uptake and nitrogen fixation are both insufficient.

When absorption is calculated (in AZODYN, it is assumed that all the mineral nitrogen present in the root zone is absorbed), it takes account of two processes: the diffusion–convection of nitrogen in the soil (empirical approach in STICS and CROPSYST and mechanistic approach in CERES-EGC), and absorption by the roots (empirical approach in CROPSYST, flux-gradient approach in CERES-EGC and the Michaelian approach in STICS).

Models that simulate legumes (CROPGRO, CROPSYST, STICS) have an explicit simulation of nodule fixation functioning.

2.6.2. Soil nitrogen

Mineral nitrogen (N-NO_3 and N-NH_4) available to the plant arises from fertilization and the mineralization of organic matter. The latter is represented by compartments (or pools) characterized by their rate of mineral nitrogen production. The three most frequently identified pools are (Fig. 13): humus (the most stable compartment of organic matter), living soil biomass (micro-organisms active on mineralization) and inputs of crop residues into the soil by the farmer. Each pool is characterized solely by the C/N ratio of organic matter. The soil acts in these transformations *via* its permanent characteristics (clay and limestone) which set a mineralization potential, and its physical status (temperature and water content) which reduces it. The mineral nitrogen produced feeds a single pool from which the plant draws its requirements. The five reference models use this representation, with two pools of organic matter (humus and residues) in AZODYN and CROPSYST and three pools in CERES-EGC, CROPGRO and STICS. AZODYN considers only nitrate nitrogen, while the other models take account of both forms of mineral nitrogen.

3. User environment

In most cases, the choice of a model is based on pragmatic considerations concerning the environment of the user. The aim of this section is to provide some information on this environment, with respect to our reference models. These concern input variables,

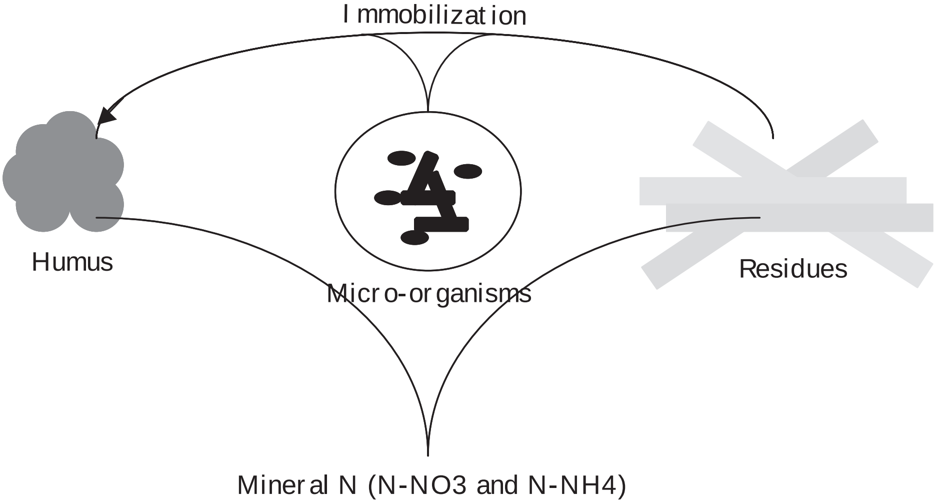


Figure 13. Schema of the principles of mineralization of organic matter; three pools of organic matter are represented here: stable humus, crop residues and micro-organisms.

parameters and the tools available to facilitate use of the model or interpretation of its results.

3.1. Input variables

All models require initialization of the water and mineral nitrogen profiles in the soil. In AZODYN, plant status at the beginning of the simulation (because this model starts during a crop cycle) must also be entered. It is also possible to initiate STICS during a crop cycle.

Climatic variables are input on a daily time step: minimum and maximum temperatures, global radiation, rainfall and potential evapotranspiration (or wind and air humidity, which are optional in STICS for the energy balance). CROPSYST and STICS are linked to climate generators (CLIMGEN: Richardson and Wright, 1984 and LARS: Semenov and Porter, 1995) which enable the simulation of long climatic series based on a shorter, observed series (a minimum of 5–10 years).

As for the hydrodynamic properties of the soil, CROPSYST and CERES-EGC use retention curves (water content–potential relationships), while STICS, CROPSYST and CROPGRO apply characteristic water content parameters (field capacity and permanent wilting point) which, in theory, constitute two specific points on retention curves. CROPSYST includes a tool which allows it to use pedotransfer functions to estimate these parameters from information on soil texture. The other soil parameters required in all models concern nitrogen mineralization (organic nitrogen, clay and limestone contents).

Management inputs are other types of explanatory variables (inputs) that must be entered, dealing with dates and amounts of irrigation or fertilizer. If not an explicit input, some models allow an automatic “decision” as to the management technique (date of sowing, irrigation or fertilization) as a function of system status and a decision rule fixed by the user, which associates the system status and the triggering and/or intensity of the action (this is possible with CROPSYST, STICS and CROPGRO).

3.2. Species and cultivar parameters

Each of our reference models has a set of plant (species)- as well as cultivar-specific parameters, which basically determine which species and variety is simulated; they distinguish wheat *versus* pea, or a winter wheat *versus* a spring wheat cultivar. The complexity of these species and cultivar inputs varies with the different models, and depends to some extent on the complexity of the equations for predicting LAI development, biomass growth or crop development. A rough estimate of the number of parameters in the reference models is shown in Table 1. Because the AZODYN model is specific to bread wheat, there are no plant-specific (or variety-specific) parameters. The other parameters (those which are explicitly available in input files) are internal parameters for the model equations, which may or may not have a physical or biological sense. The relatively large number of plant and variety parameters for models such as CROPGRO reflects modeller attempts to remove any species-specific parameters from the source code, and place them in read-in files. The complexity is also related to the fact that the model has either more processes or responds to a greater range of conditions (i.e. freezing temperatures, variable carbon dioxide concentrations, soil saturated with water, nitrogen fixation, etc.).

3.3. State variables

Typical state variables in the reference crop models include total biomass, LAI and crop developmental stage. The total biomass can be partitioned into component state variables (leaf, stem, root and seed mass), and numbers of reproductive sites (seeds) can be a state variable. Likewise, total plant nitrogen is a state variable, often partitioned similarly to biomass. Crop developmental stage can be both vegetative or reproductive. Soil water content, either total in soil or by layer, constitutes another state variable, as does soil mineral nitrogen and soil organic matter.

Normally, these state variables are calculated in the model, but for some models, users can supply exogenous data during a simulation as input, which replace the calculated values of the corresponding state variables. See below.

Table 1. Number of parameters classified according to their species/variety character in the reference models. The ranges correspond to the different options.

Type of parameters	AZODYN	CROPSYST	CERES-EGC	STICS	CROPGRO
Plant (species)	–	10–20	20–30	25–40	34
Variety (within species)		–	10–15	5–10	15
Other	10–20	30–50	30–50	30–50	70–100

3.4. Tools

STICS and AZODYN allow the user to input values of state variables that replace calculated state variables. In AZODYN, plant development input is obligatory. In STICS one can input development stages and also the LAI (Ripoche et al., 2001). This approach is useful when estimating the model parameters or when analyzing experimental results. However, this must be used with care in simulations for prediction, because driving variables (and notably LAI) are usually highly sensitive to environmental conditions and to the physiological status of the plant.

In all cases, graphic outputs make it possible to visualize the course of state variables and to compare them with observations. As for the statistical environment, CROPSYST, STICS and CROPGRO propose the calculation of mathematical distances between observed and simulated data, and STICS also has a tool to estimate parameters using a Quasi-Newton method based on ordinary least squares.

CROPSYST is linked to a GIS and a hydrological model, thus facilitating its use in the context of a spatial problem.

3.5. Programming language

AZODYN is simple enough to be programmed in the form of an EXCEL sheet, while STICS, CERES-EGC and CROPGRO are programmed in FORTRAN but can be used in a Windows environment (C++ interface), while CROPSYST is programmed in C++.

4. Discussion

Crop models provide a robust conceptual representation of the soil–crop system for annual herbaceous crops. This robustness has been tested for 25 years in numerous species and under a wide variety of pedo-climatic conditions. However, this context can become rigid and restrictive (average plant, notions of stress) when used under certain conditions. Other types of model have thus been developed by agronomists for these uses (e.g. Lescourret et al., 1998; Colbach et al., 2001), but they cannot be considered as crop models as these are defined in this presentation.

The extension of these crop models to perennial and vegetable crops (as is the case with STICS or CROPGRO) or to environmental concerns regarding the soil and atmosphere (as in CERES-EGC) is no doubt possible. If crop models are to assist in the planning of crop systems in the years to come, it is essential that they should include biotic stresses (diseases, pests, weeds) and the effects of soil cultivation operations. However, in these respects, the mechanisms are still poorly understood and research has been insufficiently oriented towards modeling, thus making it necessary for agronomists to adopt empirical approaches within crop models (e.g. weeds in STICS, Affholder, 2001). CROPGRO, for example, allows input of pest damage (from scouting) and subsequent coupling to yield reductions (Batchelor et al., 1993). Additional research in this area is essential during the coming years.

The current trend at an international level is to use functional models designed to target specific uses (APSIM, CROPSYST, STICS), or even models linked to agronomic diagnostic tools or major databases (such as AZODYN), which are less developed in terms of the processes described but use observed data to ensure model robustness.

First and foremost, crop models are tools upon which agronomic planning can be based: an aid to experimental analysis by the calculation of non-measured variables, an aid to diagnosis in a network of farm plots or a test of management techniques prior to experimentation (Boote et al., 1996). They can also provide information of value for decision-making: they can be used to simulate strategies as an aid to farm advisors (Levrault and Ruget, 2002) or they can be used to define the area of validity (soil, climate) of a strategy or a decision rule. For institutional decision makers, they may be of value by predicting the impact of regulatory decisions, state aid or a foreseeable change (climatic or economic).

Crop models also constitute training tools, either through their use or through the conceptualization of the system they provide. Use of a model with farmers or consultants (experience with APSIM, Meinke et al., 2001) helps them to visualize the impact of management choices and the hidden results of cropping systems (e.g. leaching of nitrates). Constructing a model with students allows the summarization and collection of basic knowledge reviewed during lectures (Wery and Lecoeur, 2000).

Research work can be carried out "with", "for" or "on" crop models, depending on whether that model is a tool, finality or subject of research:

- "With": the model is a research tool in the same way as a routine test or an analysis of variance. In particular, it allows one to explore complex combinations of techniques and environments before they are tested experimentally.
- "For": the development and use of crop models generate research into the functioning of a compartment (soil, plant, biological pests), the final objective being to improve the crop models.
- "On": crop models are themselves the subject of research for statisticians who can then develop methods to improve their use (parameterization, sensitivity analysis, comparison of models, etc.). This book largely concentrates on these issues.

References

Affholder, F., 2001. Modélisation de culture et diagnostic agronomique régional. Mise au point d'une méthode d'application au cas du maïs chez les petits producteurs du Brésil Central. Ph.D. Thesis, INA P-G, Paris, pp. 231.

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration. Guidelines for computing crop water requirements, FAO Irrigation and Drainage Paper 56. FAO, Rome.

Baker, D.N., 1980. Simulation for research and crop management. In: Corbin, F.T. (Ed), World Soybean Research Conference II, pp. 533–546.

Baker, D.N., Meyer, R.E., 1966. Influence of stand geometry on light interception and net photosynthesis in cotton. Crop Science 6, 15–19.

Batchelor, W.D., Jones, J.W., Hoogenboom, G., 1993. Extending the use of crop models to study pest damage. Transactions of the American Society of Agricultural Engineers 36, 551–558.

Bonhomme, R., Derieux, M., Edmeades, G.O., 1994. Flowering of diverse maize cultivars in relation to temperature and photoperiod in multilocation field trials. Crop Science 34, 156–164.

Boote, K.J., Pickering, N.P., 1994. Modeling photosynthesis of row crop canopies. HortScience 29, 1423–1434.

Boote, K.J., Jones, J.W., Pickering, N.B., 1996. Potential uses and limitations of crop models. Agronomy Journal 88, 704–716.

Boote, K.J., Jones, J.W., Hoogenboom, G., 1998. Simulation of crop growth: CROPGRO Model, Chapter 18. In: Peart, R.M., Curry, R.B. (Eds), Agricultural Systems Modeling and Simulation. Marcel Dekker, Inc, New York, pp. 651–692.

Brisson, N., 1998. An analytical solution for the estimation of the critical soil water fraction for the water balance under growing crops. Hydrological Earth System Science 2, 221–231.

Brisson, N., Delècolle, R., 1991. Dèveloppement et modèles de simulation de culture. Agronomie 12, 253–263.

Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M.H., Ruget, F., Gate, P., Devienne-Barret, F., Antonioletti, R., Durr, C., Nicoullaud, B., Richard, G., Beaudoin, N., Recous, S., Tayot, X., Plenet, D., Cellier, P., Machet, J.M., Meynard, J.M., Delécolle, R., 1998. STICS: a generic model for the simulation of crops and their water and nitrogen balance. I. Theory and parametrization applied to wheat and corn. Agronomie 18, 311–346.

Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche, D., Zimmer, D., Sierra, J., Bertuzzi, P., Burger, P., Bussière, F., Cabidoche, Y.M., Cellier, P., Debaeke, P., Gaudillère, J.P., Maraux, F., Seguin, B., Sinoquet, H., 2003. An overview of the crop model STICS. European Journal of Agronomy 18, 309–332.

Brisson, N., Bussière, F., Ozier-Lafontaine, H., Sinoquet, H., Tournebize, R., 2004. Adaptation of the crop model STICS to intercropping. Theoretical basis and parameterisation. Agronomie 24, 409–421.

Colbach, N., Clermont-Dauphin, C., Meynard, J.M., 2001. GENESYS: a model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. II. Genetic exchanges among volunteer and cropped populations in a small region. Agriculture, Ecosystems & Environment 83, 255–270.

de Wit, C.T., 1978. Simulation of assimilation respiration and transpiration of crops, Simulation Monographs. Pudoc, Wageningen.

de Wit, C.T., Brouwer, R., Penning de Vries, F.W.T., 1970. The simulation of photosynthetic systems. In: Setlik, I. (Ed), Prediction and Measurement of Photosynthetic Productivity. Proceeding IBP/PP Technical Meeting Trebon, Pudoc, Wageningen, The Netherlands, 1969, pp. 47–50.

Duncan, W.G., 1971. Leaf angles, leaf area, and canopy photosynthesis. Crop Science 11, 482–485.

Durand, R., 1967. Action de la température et du rayonnement sur la croissance. Annales de Physiologie. Végétales 9, 5–27.

Farquhar, G.D., Von Caemmerer, S., Berry, J.A., 1980. A biochemical model of photosynthetic CO_2 assimilation in leaves of C3 species. Planta 149, 78–90.

Gabrielle, B., Menasser, S., Houot, S., 1995. Analysis and field evaluation of the ceres models water balance component. Soil Science Society of America Journal 59, 1403–1412.

Gary, C., Daudet, F.A., Cruiziat, P., Brisson, N., Ozier-Lafontaine, H., Breda, N., 1996. Le concept de facteur limitant. In: Cruiziat, P., Lagouarde, J.P. (Eds), INRA Tome 1: de la plante au couvert végétal. Ecole chercheurs INRA en bioclimatologie, Le Croisic, April 3–7, 1995, pp. 121–128.

Hansen, S., Jensen, H.E., Nielsen, N.E., Swenden, H., 1990. DAISY – soil plant atmosphere system model NP0 research in the NAEP report, Nr A10, The Royal Veterinary and Agricultural University, pp. 272.

Itier, B., Brisson, N., Doussan, C., Tournebize, R., 1997. Bilan hydrique en agrométéorologie. In: Cruiziat, P., Lagouarde, J.P. (Eds), INRA, Tome 2: du couvert végétal à la petite région agricolel. Ecole chercheurs INRA en bioclimatologie, Le Croisic, April 3–7, 1996, pp. 383–398.

Jeuffroy, M.H., Recous, S., 1999. AZODYN: a simple model simulating the date of nitrogen deficiency for decision support in wheat fertilisation. European Journal of Agronomy 10, 129–144.

Jeuffroy, M.H., Barre, C., Bouchard, C., Demotes-Mainard, S., Devienne-Barret, F., Girard, M.L., Recous, S., 2000. Fonctionnement d'un peuplement de blé en conditions de nutrition azotée sub-optimale. In: INRA, Paris (Ed), Fonctionnement des peuplements végétaux sous contraintes environnementales, Paris, France, January 20–21, 1998, pp. 289–304.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. European Journal of Agronomy 18, 235–265.

Lacape, M.J., Wery, J., Annerose, D.J.M., 1998. Relationships between plant and soil water status in five field-grown cotton (*Gossypium hirsutum* L.) cultivars. Field Crop Research 57, 29–43.

Lafolie, F., 1991. The PASTIS model. Fertiliser Research 27, 215–231.

Lemaire, G., Gastal, F., 1997. N uptake and distribution in plant canopies. In: Lemaire, G. (Ed), Diagnosis of the Nitrogen Status in Crops. Springer-Verlag, Berlin pp. 3–44.

Lescourret, F., Habib, R., Génard, M., Agostini, D., Chadoeuf, J., 1998. Pollination and fruit growth models for studying the management of kiwifruit orchard. I. Models description. Agricultural Systems 56, 67–89.

Levrault, F., Ruget, F., 2002. COGITO. Cogito, un modèle pour l'irrigation du maïs. In: Jaeger (Ed), Modélisation des agro-écosystémes et aide à la décision, CIRAD, Collection Repères, Malézieux, Trébuil, pp. 281–300.

McCown, R.L., Hammer, G.L., Hargreaves, J.N.G., Holtzworth, D.P., Freebairn, D.M., 1996. APSIM: a novel software system for model development, model testing and simulation in agricultural systems research. Agricultural Systems 50, 255–271.

McKinion, J.M., Baker, D.N., Whisler, F.D., Lambert, J.R., 1988. Application of the GOSSYM/COMAX system to cotton crop management, ASAE Paper No. 88-7532. St. Joseph, MI.

Meinke, H., Baethgen, W.E., Carberry, P.S., Donatelli, M., Hammer, G.L., Selvaraju, R., Stöckle, C.O., 2001. Increasing profits and reducing risks in crop production using participatory systems simulation approaches. Agricultural Systems 70, 493–513.

Monteith, J.L., 1972. Solar radiation and productivity in tropical ecosystems. Journal of Applied Ecology 9, 747–766.

Olioso, A., Chauki, H., Courault, D., Wigneron, J.P., 1999. Estimation of evapotranspiration and photosynthesis by assimilation of remote sensing data into SVAT models. Remote Sensing Environment 68, 341–356.

Pellegrino, A., Wery, J., Lebon, E., 2002. Crop management adaptation to water-limited environments, 2002. Proceedings of the VII ESA Congress, July 15–18, Cordoba, Spain, pp. 313–314.

Penning de Vries, F.W.T., van Laar, H.H., 1982. Simulation of growth processes and the model BACROS. In: Penning de Vries, F.W.T., van Laar, H.H. (Eds), Simulation of Plant Growth and Crop Production. Center for Agricultural Publishing and Documentation, Wageningen, pp. 114–135.

Priestley, C.H.B., Taylor, R.J., 1972. On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly Weather Review 100, 81–92.

Richardson, C.W., Wright, D.A., 1984. WGEN: a model for generating daily weather variables, ARS-8. US Department of Agriculture, Agricultural Research Service, pp. 83.

Ripoche, D., Weiss, M., Prevot, L., 2001. Driving the STICS crop model by exogenous values of leaf area index. Application to remote sensing. Proceedings of the Second International Symposium on Modelling Cropping System, Florence, Italy, July 16–18, pp. 169–170.

Ritchie, J.T., 1985. A user-oriented model of the soil water balance in wheat. In: Fry, E., Atkins, T.K. (Eds), Wheat Growth and Modeling. NATO-ASI Series, Plenum Press, New York, pp. 293–305.

Ritchie, J.T., Otter, S., 1984. Description and performance of CERES-Wheat a user-oriented wheat yield model. USDA-ARS-SR Grassland Soil and Water Research Laboratory Temple RX, pp. 159–175.

Sau, F., Boote, K.J., Bostick, W.M., Jones J.W., Mínguez, M.I., 2004. Testing and improving evapotranspiration and soil water balance of the DSSAT crop models. Agronomy Journal 96, 1243–1257.

Semenov, M.A., Porter, J.R., 1995. Climatic variability and the modelling of crop yields. Agriculture for Meteorology 73, 265–283.

Sinclair, T.R., 1986. Water and nitrogen limitations in soybean grain production. I - Model development. Field Crop Research 15, 125–141.

Spaeth, S.C., Sinclair, T.R., 1985. Linear increase in soybean harvest index during seed-filling. Agronomy Journal 77, 207–211.

Spitters, C.J.T., 1986. Separating the diffuse and direct component of global radiation and its implications for modeling canopy photosynthesis. Part II. Calculation of canopy photosynthesis. Agriculture for Meteorology 38, 231–242.

Spitters, C.J.T., Toussaint, H.A.J.M., Goudriaan J., 1986. Separating the diffuse and direct component of global radiation and its implications for modeling canopy photosynthesis. Part I. Components of incoming radiation. Agriculture for Meteorology 38, 217–229.

Stöckle, C.O., Martin, S., Campbell, G.S., 1994. CropSyst, a cropping systems model: water/nitrogen budgets and crop yield. Agricultural Systems 46, 335–359.

Stöckle, C.O., Donatelli, M., Nelson, R., 2003. CROPSYST, a cropping systems simulation model. European Journal of Agronomy 18, 289–307.

van Ittersum, M.K., Leffelaar, P.A., van Keulen, H., Kropff, M.J., Bastiaans, L., Goudriaan, J., 2003. On approaches and applications of the Wageningen crop models. European Journal of Agronomy 18, 201–234.

Warren-Wilson, J., 1972. Control of crop processes. In: Rees, A.R., Cockshull, K.E., Hand, D.W., Hurd, R.G. (Eds), Crop Processes in Controlled Environment. Academic Press, London, pp. 7–30.

Watson, D.J., 1947. Comparative physiological studies on the growth of field crops. I. Variation in net assimilation rate and leaf area between species and varieties, and within and between years. Annals of Botany NS 11, 41–46.

Weir, A.H., Bragg, P.L., Porter, J.R., Rayner, J.H., 1984. A winter wheat crop simulation model without water or nutrient limitations. Journal of Agriculture Science, Cambridge 102, 371–382.

Wery, J., Lecoeur, J., 2000. Learning crop physiology from the development of a crop simulation model. Journal of Natural Resources and Life Sciences Education 29, 1–7.

Whisler, J.R., Acock, B., Baker, D.N., Fye, R.E., Hodges, H.F., Lambert, J.R., Lemmon, H.E., McKinion, J.M., Reddy, V.R., 1986. Crop simulation models in agronomic systems. Advances in Agronomy 40, 141–208.

Williams, J.R., Jones, C.A., Dyke, P.T., 1984. A modeling approach to determining the relationship between erosion and soil productivity. Transactions of American Society of Agricultural Engineers 27, 129–144.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 10

Crop models with genotype parameters

M.-H. Jeuffroy, A. Barbottin, J.W. Jones and J. Lecoeur

1. Introduction **281**
2. Uses of genotype-adapted crop models **282**
3. Issues related to genotype parameter estimation **284**
4. Methods for estimating genotype parameters **291**
5. Examples **294**
6. Conclusions **303**
References **304**

1. Introduction

The use of genotype specific parameters in models has two distinct objectives. The first is to determine model parameters with significantly different values for different genotypes. These differences may then indicate functional differences between varieties that might be useful during breeding, if they prove to be of value. The second aim is to improve model predictive quality and, as a consequence, the management recommendations that may be derived from the model. In this case, the parameter values themselves are not of interest. Instead, model performance is important, and compensation of errors in parameter values is acceptable if system performance is satisfactorily simulated.

There is the implicit assumption here that the model structure can represent genotypic variability through variations in parameter values. However, most models have been designed to have a limited number of parameters (with the objective of ensuring their generality), which simplifies the representation of the plant but restricts the ability of the model to simulate genotypes. In fact, few models are capable of reproducing variability between genotypes, or may do so for only a limited number of genotypes (often with a similar genetic origin), but prove unsuitable when a broader genotype pool is considered.

For those models which are potentially capable of describing genotype variability, the problem of estimating genotypic model parameters can be broken down as follows:

(1) identification of model parameters which are genotypic;
(2) estimation of parameter values for the genotypes targeted.

The methods that provide an answer to these two questions are not necessarily identical, although it is sometimes possible to answer both questions at the same time. Chapter 4 reviews rigorous statistical methods used to estimate model parameters. These methods could be applied to estimate genotypic parameters in a model, but probably would require additional information concerning the parameters. In some models, genotypic parameters are identified by model developers based on their knowledge of the crop rather than by statistical methods (e.g. the DSSAT models, Jones et al., 2003). In those models, parameters are separated into those that are assumed to be constant among genotypes (species parameters) and those that vary with the genotype.

Our aim here is to identify the problems confronted while estimating genotypic parameters in a model and also while using those models. These problems are linked primarily to one's objective and then to the data available for estimating these parameters. Hence we broaden our discussion to the need and possibility of making models genotype-specific for various objectives.

2. Uses of genotype-adapted crop models

2.1. Predicting differences in responses over a range of soil and climate conditions

The objective here is that the model with genotypic parameters be able to simulate important differences among varieties. These differences may be linked to genotype-specific characteristics or to an interaction between those characteristics and soil, climate, or management conditions in which the genotype is grown. Thus, the model is used as a tool to analyze genotype × environment (G × E) interactions.

Numerous statistical methods have been proposed for partitioning the G × E interaction, including multiplicative models, biadditive factorial regression, or linear factorial regression (van Eeuwijk, 1995; Brancourt-Hulmel et al., 1997). Currently the best developed methods for analyzing and predicting G × E interactions are based on static statistical relationships whose parameters are easy to estimate using experimental data from plant breeding trials, coupled with a diagnosis of limiting factors (Hulmel et al., 1999). The results of the diagnosis give variables describing the environmental conditions encountered by the varieties in test. Then these variables, and variables describing genotypic characteristics, are taken as covariables in the statistical model aiming at analyzing and partitioning G × E interaction (Denis, 1988; Brancourt-Hulmel et al., 2000; Brancourt-Hulmel and Lecomte, 2003).

The use of mechanistic dynamic crop models might improve the analysis of G × E interactions. Such models are often tested for their ability to predict yield in a broad range of environments (Travasso and Magrin, 1998; Mavromatis et al., 2001), so that they should be capable of correctly taking a range of conditions into account. For example, they could provide new variables to explain observed variations in yield, or they could provide simulated production potentials for the experimental environments, or they could help diagnose limiting factors (Hulmel, 1999; Boote et al., 1996). However, the use of such models raises numerous methodological problems, and the usefulness of these tools when compared with existing methods is currently under investigation.

2.2. Adapting crop management to specific cultivars

Adapting crop management to a cultivar signifies the choice of farming techniques best suited to the characteristics of the variety. If the aim is to identify, as soon as the variety is registered, the crop management that will best exploit its potential, one solution may be to classify the new genotype in an existing typology, based on known characteristics of that genotype. This typing represents the specific functioning of certain varieties and is based on genotypic characteristics, which can be represented by the genotypic parameters of crop models. The management factors that have been studied using this approach are limited, but are highly important in some environments, namely planting date, irrigation, and nutrient management (Keating et al., 1991; Kropff et al., 1995; Bindraban et al., 1997; Jones et al., 2000; Royce et al., 2002). The genotypic differences among varieties that are most frequently taken into account in these applications are those that control phenological development and crop duration. Differences in photoperiod response and basic phase durations translate into differences in water and nutrient requirements as well as in potential yield. That allows these types of models to be used to optimize these three management factors for a variety with known genotypic parameters and for a specific location.

In most cases, genotypic parameters in crop models describe the phenological differences between varieties (e.g. Liu et al., 1989; Hammer et al., 1995; Goyne et al., 1996). However, these differences are sometimes of limited importance compared to other factors that affect crop performance. For example, the usefulness of multi-resistant wheat varieties under low-input crop management (Loyce et al., 2001) is mainly linked to a combination of their high productivity and their multiple resistances to diseases. However, crop simulation models rarely simulate damage caused by diseases. These phenomena are complex and the reactions of varieties may vary considerably. This requires a large number of genotypic parameters and a high cost of parameterization. For this reason, crop models are not capable of predicting the differences in behavior among varieties for all possible crop management strategies or of identifying those strategies that are best suited to all characteristics of each variety. The BETHA model (Loyce et al., 2002) was specifically developed to choose the best variety and crop management for a given objective (Chapter 20), and was applied to the case of wheat for ethanol production. However, this is a static rather than dynamic crop model and does not simulate the processes of plant growth and development. The introduction of additional genotypic effects into crop models (although it might increase the number of parameters) will likely improve model performance (predictive quality and quality of model-guided decisions), but this remains to be proven.

2.3. Selecting varieties for specific environments

One aim is to determine the variety best suited to a given environment and management system. Crop models respond to seasonal patterns of rainfall, temperature, and daylength, primarily due to differences in genotypic parameters that control phenological development. It is therefore necessary to identify varietal characteristics (or genotypic parameters) that will predict the behavior of the variety under consideration as a function of different environments. Crop models have been used to select varieties with genotypic parameters that result in optimal yield (in terms of amount and annual variability) for a number of crops (e.g. sorghum (Chapman et al., 2000), rice (Aggarwal et al., 1997), soybean (Boote et al., 2003; Messina et al., 2005), and wheat (Hammer et al., 1999a)). However, there may be genotype by environment interactions that are key determinants in variety choice but not taken into account by the models. The results from model-based studies must be judged in the light of interactions not included in the models. For example, in the case of wheat, the varietal characteristics of tillering capacity, an important determinant of the behavior of a variety at a low density (Meynard et al., 1988), nitrogen absorption capacity (Le Gouis and Pluchard, 1996), sensitivity to low radiation at meiosis (inducing marked grain abortion, Demotes-Mainard et al., 1996), and tolerance to post-flowering water stress (Meynard et al., 1988) are rarely taken into account in dynamic crop simulation models.

2.4. Breeding varieties for specific environments

A priori evaluation of the possibilities of variety selection to improve yield or some other variable of interest can be assessed using crop simulation models (Agüera et al., 1997; Hammer et al., 1999b; Fargue, 2002). In this case, genetic variability is introduced into the model by varying the genotypic parameters, and the consequences on the variables of interest (yield, quality, gene flux, etc.) are estimated by simulation.

A prerequisite is to know how the characteristics represented by the genotypic parameters affect the variables of interest. It is also necessary to evaluate the range of variability in the existing genetic material. This requires experimental measurements for a large number of lines. For example, Barbottin (2004) showed that mean weight per grain, the ability of the variety to produce grain relative to intercepted radiation and maximum grain yield, earliness at heading and at the beginning of stem elongation and crop biomass at the end of winter were cultivar characteristics that could be included as genotypic parameters in a model, and that these parameters could help to understand and predict cultivar responses (in terms of grain yield and grain protein content) to different environments. Thus, such characteristics should be systematically measured in variety trials as they could help breeders to create cultivars adapted to specific environments (Barbottin and Jeuffroy, 2004). One could then test a reasonable range of values for the genotypic parameters in order to identify interesting new ideotypes. In this case, it is important to first analyze the genetic variability that exists for the different characteristics that are to be varied (for cultivated and/or wild genotypes). Then, the model can be used to analyze the behavior of virtual genotypes defined as original combinations of those parameter values that present genetic variability. Studies of this type include Hammer et al. (1996), Bindraban (1997), Fargue (2002), Boote et al. (2003) and Messina et al. (2006).

3. Issues related to genotype parameter estimation

3.1. Available data

Crop models are usually developed for a few varieties using databases that include numerous intermediate variables for testing model algorithms. Classically, model adjustment for a small or larger number of genotypes is pursued in a second step. Often, only a small number of measurements are available for these other genotypes. Nonetheless, they sometimes have been obtained from yield trials performed over a broad range of environments

(Mavromatis et al., 2001). This second step is usually carried out without reassessing the model's structure and relationships chosen before genotypic adaptation.

Numerous variety yield trials exist, performed by plant breeders during several years of candidacy for registration or by advisor services in order to prepare their cultivar choice for their specific range of environments. The variables measured in these trials are often restricted to the dates of major developmental stages and yield, perhaps also final biomass and yield components. They rarely include intermediate data, which would require additional sampling during the crop cycle. However, such trials are of considerable value for estimating genotype parameters because of their number. Experiments specifically designed for varietal parameter estimation would be costly and difficult to organize for a large number of genotypes (Reymond, 2001). Piper et al. (1998) and Mavromatis et al. (2001) estimated soybean genotype parameters using yield trial data with only grain yield, final biomass, maturity date and flowering date (in some cases) measured. Usually, it is possible to obtain the necessary weather data from historical archives for the site or a site close by. However, detailed data on soil characteristics may not be available for some or all sites in a set of yield trials and, if the trials were irrigated, detailed records of dates and amounts or irrigation may not be recorded. Thus, differences between cultivar responses among sites in such trials involve both unknown genotypic and site characteristics. The work by Mavromatis et al. (2001, 2002) attempted to overcome this problem by estimating both cultivar and site characteristics. In their case, the site characteristics included a soil water holding parameter as well as an empirical fertility parameter, but those parameters were not themselves of interest. They were simply used to characterize the sites in order to allow estimation of the genotypic parameters, which was the objective of their study.

In addition, attempts to adjust a model to different varieties based on data available from previous yield trial varietal studies may influence model functions themselves. For example, the equations used to characterize grain biomass requirements in the Azodyn model have evolved as a result of varietal adjustment (Jeuffroy et al., 2000). Another example is the use of soybean yield trial data in the USA to adjust temperature functions in the CROPGRO-Soybean model (Piper et al., 1996; du Toit, unpublished manuscript). A final example is the "yield loss due to disease" parameter, included in BETHA (Loyce et al., 2002), which is directly based on the disease resistance scores produced by the Groupe d'Evaluation des Variétés et des Semences (GEVES), the institution that controls registration of new varieties in France. This cultivar parameter is immediately available for any newly registered variety in France, allowing the model to be used for a new variety as soon as the variety is registered, without requiring any specific trial for estimating the genotypic parameters. Thus the type of data available may lead to a change in model formulation, in order to permit rapid evaluation of genotypic parameters. This approach can be justified because the ideal database providing all intermediate variables for the model chosen initially will never be available.

3.2. Keeping up with new variety releases

Varieties have increasingly short life-spans; a large number of new varieties are registered each year. It is therefore impossible to identify those that are likely to be widely used in subsequent years. For this reason, modeling tools need to be adjusted rapidly in order to estimate genotypic parameters each year for a large number of varieties. This requirement represents a major challenge.

The example of multi-resistant wheat varieties illustrates this need. These varieties are of considerable economic (better gross margins when average wheat price is low) and environmental (reduction in nitrogen and pesticide use on the crop) importance, but this is only expressed when associated with low-input management (Loyce et al., 2001; Félix et al., 2002). It is therefore necessary to determine, as soon as it is registered, the type of the new variety in order to recommend appropriate management to ensure its optimum exploitation. Easily measurable characteristics are therefore necessary to enable a rapid assessment of its capacities, particularly since comparative trials of varieties are usually poorly suited to identifying this type of variety (generally used in intensive farming). Wilkerson et al. (2001) developed a software to automate the approach published by Mavromatis et al. (2001, 2002), which uses yield trial data to estimate soybean genotype parameters for the CROPGRO-Soybean model (Boote et al., 1998). This software assembles appropriate soil, weather, and yield trial data, then searches for the best combination of parameters for each of the varieties in the trial.

The context of varietal development may be significantly modified if GMO varieties are introduced on a major scale. In this setting, crop model designers will no longer be confronted solely with a slow development of parameter values for new varieties, built up from a small number of genotypes forming an elite group, but with abrupt changes in the varietal ideotypes obtained. In *Arabidopsis thaliana*, the modification of a single transcription gene has led to considerable modifications in plant behavior in terms of its phenology (vegetative and reproductive development), response to light, morphology, and mode of leaf growth (Franck, 2001). Crop model developers are aware of the potential for using information on molecular DNA of a variety for characterizing its development and growth processes (White and Hoogenboom, 1996; Hammer et al., 1996; Boote et al., 2003; Messina et al., 2005). Crop modelers need to interact with molecular geneticists and physiologists to facilitate the translation of genetic knowledge to modes of action, and finally to integrate field performance under multiple environments in a way that accounts for limiting resources. Modelers' views of genetic coefficients will change to accommodate increasing availability of genetic information. While the modeled genetic coefficients will remain mathematical constructs to model phenotypic outcomes, in future they will be more closely linked to actual genes/DNA sequences (Boote et al., 2003; also see Chapter 11).

3.3. Need for precision of simulated variables

Genotypic variability of a characteristic or function is small, often smaller than the experimental error. Moreover, varieties usually differ with respect to several functions simultaneously. In order to reveal any difference in functioning and production between varieties, it is necessary for the model and the data used to estimate parameters to be accurate. There are many examples for this difficulty in managing experimental variabilities when comparing genotypes (e.g. Sinclair and Muchow, 1999). A first example could be biomass, which is frequently measured by varietal trial networks. The coefficients of variation of biomass measurements are usually between 10 and 20%, while the differences between varieties rarely exceed 5%. Access to a statistically very powerful experimental design is therefore necessary, in order to identify the occurrence of a varietal effect and estimate different parameters for each genotype. Other examples can also be cited, such as the grain protein content of wheat: the difference in protein contents between genotypes of common wheat varieties, with a similar yield, are usually about 0.5–1% (Bernicot, 2002). However, a 1% variation in protein content, around the mean value of 11.5%, corresponds to a 9% difference in the nitrogen uptake of the crop. This difference is usually lower than the experimental error on this variable which ranges from 10 to 15%. Finally, the difference in nitrogen absorption between wheat genotypes is about 20–30 kg ha^{-1} (Le Gouis and Pluchard, 1996), or approximately 10% of the mean value observed in a non-nitrogen limiting situation, which is also close to the experimental error. This small variability among genotypes makes it even more difficult to demonstrate using models.

One way to circumvent this constraint would be to simulate relative deviations between two genotypes or rankings for a group of genotypes, rather than results in terms of absolute values. In this case, rank tests might be more appropriate for evaluating model performance than mean squared error or mean squared error of prediction.

Another way of circumventing the problem of accuracy may be to examine the dynamics of crop growth rather than just cumulative values. An example is the radiation use efficiency in pea. Due to experimental errors in biomass measurements, it is difficult to demonstrate cultivar differences in this parameter. In a study by Lecoeur and Ney (2003), the use of cumulative dry matter measurements led to respective values of 2.90 and 2.84 g dry matter MJ^{-1}m^{-2} for the two cultivars studied. Nevertheless, differences in biomass were observed and related to cultivar differences in the variability over time of the radiation use efficiency. Thus, taking these differences into account made it possible to demonstrate varietal differences in the course of this parameter during the growth cycle (Lecoeur and Ney, 2003).

3.4. Need for physiologically significant parameters

It is generally assumed that genotypic parameters are robust, i.e. they can be used to describe the performance of a cultivar in environments other than those used for estimation. However, this is rarely confirmed. Thus the parameters may be biased and may provide acceptable results only for the site or region in which data were obtained for estimating them. If this is the case, those parameters may perform well for different sites and years within the region but fail when they are carried to other regions.

This issue was addressed using the CROPGRO-Soybean (Boote et al., 1998) model. The aim of the work by Mavromatis et al. (2002) was to evaluate the robustness of soybean genotypic parameters for this model. They used data from two different state yield trials (Georgia and North Carolina in the USA), for different sets of years, to determine (1) whether parameters estimated from two different regions had the same value and (2) whether prediction of development and yield in one state was degraded if genotypic parameters were estimated using data from another state. They found that the parameters that affect development (i.e. photoperiod response and duration of crop growth phases) were stable regardless of data origin. Figure 1 shows a graph of the photoperiod sensitivity parameter (CSDL) estimated in North Carolina plotted against the same parameter

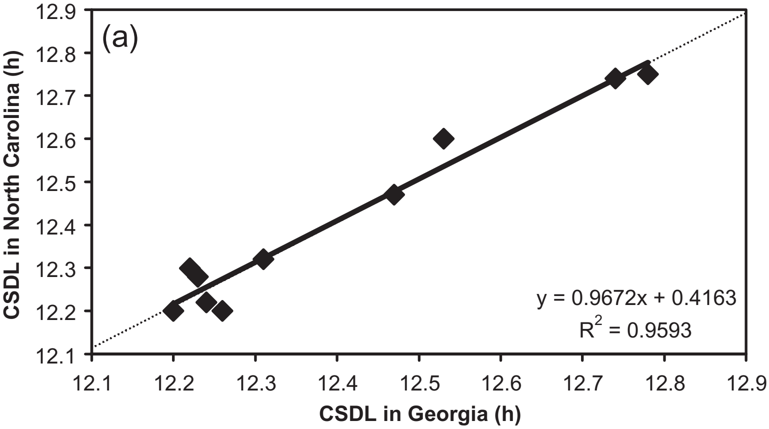


Figure 1. Comparison of critical day lengths (CSDL) estimated for ten soybean cultivars in Georgia and North Carolina. The 1:1 line (doted) and the regression line (solid) between the data are also shown.

estimated using data from Georgia. On the other hand, parameters that were estimated using yield data, such as seed filling duration and maximum leaf photosynthesis, varied between states. However, yield prediction for one state using parameters estimated from the other state was estimated with little loss in accuracy; there was compensation in parameter values such that yield predictions using parameters estimated from data in another state were degraded very little for this model. Figure 2 shows predicted maturity date and yield for North Carolina using parameters estimated from data in Georgia using all cultivars in the Mavromatis et al. (2002) study. This study demonstrates the fact that some parameters in crop models may be more robust than others and also shows the importance of evaluating the use of genotypic parameters for environments not included in the data used for estimation.

If certain parameters are critical in explaining variability in simulated output data, they may constitute good candidates for breeding criteria (Agüera et al., 1997; Fargue, 2002). In this case, breeders may attempt to measure the parameters directly on lines during the breeding process. Similarly, if a varietal characteristic is important in understanding and predicting variety behavior under given environments and management, it will play a major role in the choice of variety by the farmer (Loyce, 1998). Such characteristics should be made available to producers as soon as the variety is released. To attain these objectives, it is essential that the genotypic parameters should remain stable in a broad range of environments. Unfortunately, this quality is rarely verified by authors.

With this in mind, it may not be desirable to reduce the quality of a parameter by treating it as a statistical variable and estimating it by minimizing model output errors. The assessment of its quality should also include its biological significance (Sinclair and Seligman, 2000). One may even accept a less satisfactory parameter estimate in order to preserve its physiological significance, and thus attach less importance to the predictive value of the model than to the physiological significance of its parameters. However, this depends on one's objective. One should realize, however, that estimates obtained

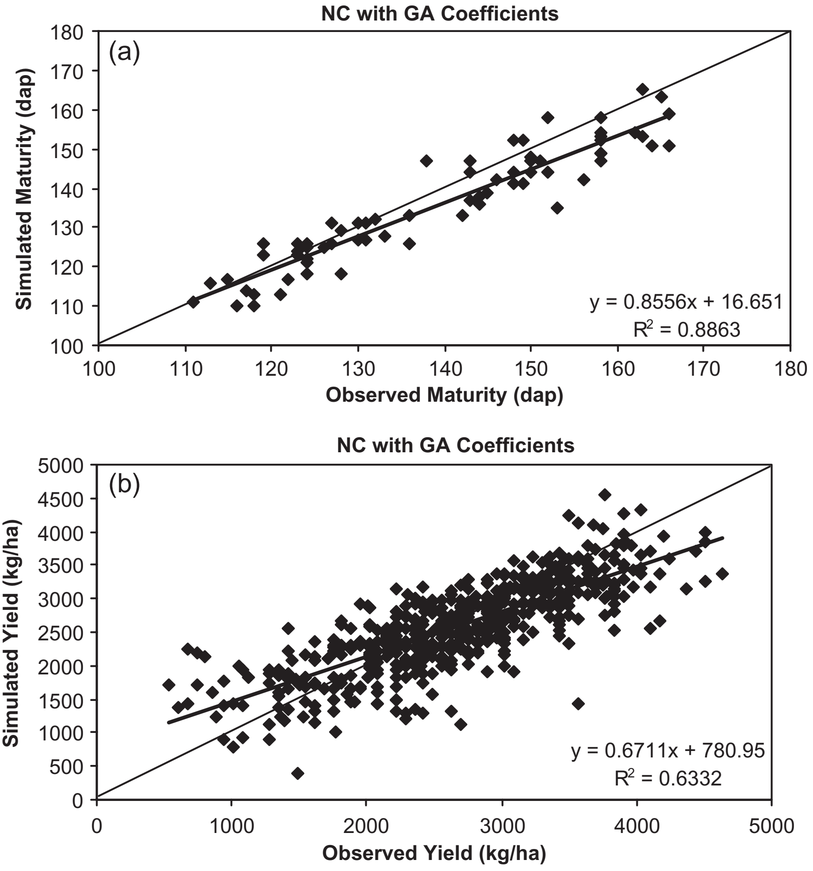


Figure 2. Comparison of simulated *versus* observed harvest maturity (a) and seed yield (b) for ten soybean cultivars in North Carolina with the coefficients developed in Georgia. The 1:1 line (dot) and the regression line (solid) between the data are also shown.

in an optimization approach may reflect errors in the model and may not be robust. It is important to evaluate the parameters using independent data before using them in environments different from those used in estimation.

3.5. Choice of parameters to be estimated

Identification of the genotypic parameters of a model consists in determining the parameters that differ between varieties and also strongly influence simulated output data. It is therefore necessary to determine the genotypic variability of parameters and also to analyze the model's sensitivity to those parameters.

In fact, there are few data in the literature that describe existing genotypic variability of different characteristics of interest. Furthermore, the explicit parameters of the model are rarely studied. For example, in the case of wheat, several authors (Van Sanford and MacKown, 1986; Le Gouis and Pluchard, 1996) demonstrated the existence of genotypic variability regarding nitrogen absorption capacity of plants. However, the references do not allow direct integration of these results into models, as no indication is given as to which parameters should be adjusted and to what extent it is possible to link results with model parameters. For this reason, experimental research into the variability of model parameters becomes the responsibility of those who adapt the model to include new genotypes. This raises three problems:

- The number of crop model parameters is often much too large for such a study to be rigorous if all of them are considered. Published literature or expert knowledge may help sort out those functions where, in principle, variability between varieties exists. Unfortunately, this does not apply in every case. Reference may be made to soil parameters which are often included in crop models and which do not vary between genotypes; to dilution curve parameters (Justes et al., 1994) or even to parameters of a climatic type, such as the ratio between photosynthetically active radiation and global radiation (PAR/GR). Mention could also be made of intercepted radiation using Beer's Law. LAI development parameters differ between varieties, as do radiation extinction parameters in the canopy, although maximum interception can be considered to remain constant within a species (Jones, 1992). Inversely, some parameters are certain to differ between varieties: individual leaf surface parameters, in the AFRC-wheat model (Weir et al., 1984), or parameters linked to yield components (Agüera et al., 1997; Travasso and Magrin, 1998). But what about parameters affecting crop responses to water stress or nitrogen limitation? Another way to achieve this prior discrimination is to analyze model sensitivity to the parameters (Makowski et al., 2005). If model output data are affected little by the value introduced for the parameter, then there is no need to search for the existence of variability. However, results of sensitivity analyses depend on the range tested for the parameter under study. In principle, this range must reflect variability that exists in nature, and hence the existing genotypic variability. We therefore find ourselves in a vicious circle, where sensitivity analyses are necessary to set up experiments aimed at estimating genotypic parameters, but experimental results are necessary to know the range of parameters to be tested in the sensitivity analysis. The pragmatic solution consists in adopting a step-by-step approach to choosing genotypic parameters based on expert knowledge or testing of previously defined ranges.
- Sensitivity to parameters depends not only upon the value of the parameter under study but also on all the values of other parameters or input variables used (see, for example, Girard, 1997). A trivial example might be the search for variability in parameters concerning a reduction in the number of grains linked to nitrogen deficiency, in situations where there is no effective nitrogen deficiency. However, the situation is more complicated for processes resulting from multiple interactions. Because the number of analyses is generally limited, it is therefore essential to make appropriate choices.
- When experiments are necessary to study the existence of genotypic variability in a parameter, the number of genotypes tested is often small. The choice of parameters is then crucial, as the aim is to extrapolate the conclusions of the study to a much broader range of genotypes. It is then useful to work on contrasting genotypes and not be restricted to those most recently registered, which are often very similar in terms

of performance or biology. For species that have been studied for several decades, one option would be to review those varieties that have made their mark at different times. The advantage is that for these varieties, considerable reference data are often available and are sufficiently contrasted because of advances in breeding and the development of management techniques (Vear et al., 2002). We are optimistic that, in the future, molecular genetics techniques will be useful in estimating parameters for each variety that is released (Boote et al., 2003).

Experimental research on genotypic variability and the analysis of model sensitivity to genotypic parameters can produce contrasting results. Indeed, three situations may be observed:

- Case 1: Literature reports the existence of genotypic variability in a function, the model is sensitive to the parameter of this function, but experimental results show that the variability of the parameter chosen to reflect this function is small and does not vary among varieties. The case of nitrogen remobilization from vegetative organs to wheat grains is an example, since Van Sanford and MacKown (1987) reported the existence of variability among varieties. Barbottin et al. (2005) showed that the Azodyn model's remobilization parameter was stable between genotypes and environments. This apparent contradiction between two series of experimental findings arises from the fact that nitrogen remobilization was not measured in the same way in both the cases, even though the same function was under study, and the range of environments tested was not the same.
- Case 2: Literature and experimental findings show the existence of genotypic variability in some parameters (e.g. EBMAX, D and VMAX parameters in Azodyn), but the model is not sensitive to them, at least in terms of outputs of interest (grain protein content output for Azodyn).
- Case 3: Variability exists and the model is sensitive to it. An example is the maximum weight per grain in Azodyn (Champeil, 2001; Philibert, 2001).

In conclusion, experimental research concerning genotypic variability in parameters, and analysis of model sensitivity to parameters which are, in principle, genotypic, constitute two complementary approaches. It is preferable to combine these two approaches to more reliably identify those model parameters that should be estimated for any new variety.

4. Methods for estimating genotype parameters

4.1. Direct measurement of parameters

Although direct measurement may appear to be the best approach for estimating genotypic parameters, it is uncommon in practice. It enables direct access to the desired parameter via experimental measurements. If the parameter is genotypic in nature and significantly affects crop performance, the breeder could measure it directly on lines under development in experiments in order to predict the expected effects. However, this method often requires specific trials and measurements, which may therefore be complicated, costly and even impossible to implement for a high number of genotypes (Reymond, 2001). Routine measurement of these parameters for a large number of varieties may pose a problem, particularly when measurements require special equipment and controlled condition experiments: e.g. parameters for the response of maize leaf growth to temperature, radiation or vapor pressure deficit (Reymond, 2001).

One good example to illustrate this approach is the thermal time to flowering in crops. In this example, dates of emergence and first flower are recorded for each variety in each trial. These dates, along with daily maximum and minimum temperatures recorded at the sites, allow one to compute directly, for each combination of variety and location, the degree days required for flowering. Examples of this approach can be found for peanut and maize (e.g. see Boote et al., 1998, 2003). Another example is the genotypic parameters "maximum leaf size" and "specific leaf area at the end of the vegetative stage" in the CROPGRO models in DSSAT (Boote et al., 1998). These variables can be measured directly in experiments that are grown under ideal conditions. Although such conditions may not exist in yield trials, one could obtain an estimate of these parameters by comparing data from a wide range of conditions in the trials. In some cases, it may be possible to estimate some important parameters indirectly from indicators that are routinely measured in yield trial networks. For example, in sunflower variety trial networks, one routinely measured variable is grain moisture at harvest. This variable is strongly dependent upon climatic conditions at the end of the growth cycle. However, classification of genotypes based on relative values calculated from reference genotypes exhibits satisfactory stability. Results from a series of closely-studied genotypes show that correlations exist between the relative value of this indicator and the length of the flowering-maturity stage. Thus, based on grain moisture at harvest, it is possible to estimate a value for the "duration of flowering – maturity stage" for new genotypes, without measuring it directly.

In most cases, however, it is not possible to directly measure some parameters that are used in crop models and it is impractical to measure others. Thus, direct measurement, while more appealing, will not be suitable in all cases. It may be possible to measure some parameters directly, whereas indirect methods will be required to estimate others.

4.2. Estimating parameters by minimizing errors in model outputs

This indirect method (see Chapter 4) consists of estimating one or more parameters by minimizing the differences between measured and calculated values. This optimization is generally performed on model output variables (usually yield), and sometimes on intermediate variables (dates of key development stages, for example, Liu et al., 1989; Grimm et al., 1993; Mavromatis et al., 2001). Particular attention must be paid to correlations existing between parameter estimators, which may produce parameter values which are satisfactory for prediction under a limited range of conditions. Such parameters may not have physiological significance and therefore may not be applicable for conditions other than those used in estimation (Jamieson et al., 1998). One way to counteract this problem is to estimate each parameter by minimizing the errors of intermediate simulated variables of the model, and not of the outputs of the model. Moreover, the error of parameter estimation is reduced if the estimation is based on situations which have a direct effect on the parameter (Dorsainvil, 2002). Furthermore, under this optimization procedure, the further the parameter to be estimated is situated from the variable to be optimized within the model structure, the greater are the possibilities that its value will depend on that of other parameters estimated simultaneously and thus not have any biological meaning.

In many practical studies, various measurements are available for estimating parameters, yet not all are used to estimate each parameter. Knowledge of the parameter and its role in affecting an output is used to target parameters to observable traits. Examples are photoperiod sensitivity being estimated from observations on time to first flower. Additional knowledge-based rules may be applied to help ensure that parameter estimates are reliable. Hunt et al. (1993) developed this approach in the GENCALC software used to estimate genotype parameters in the DSSAT suite of crop models. A data file is used to input rules for each crop linking parameters to specific traits measured. The GENCALC software varies the parameters, runs the model for those parameters, evaluates simulated response relative to one or more measured variables, then selects those parameters that minimize error between simulated and observed variable.

One advantage of this error minimization method is that it is inexpensive in terms of the data necessary to estimate parameters. It thus enables the exploitation of databases acquired for other purposes (e.g. the numerous variety yield trials performed by different organizations) and it is not necessary to repeat specific trials (Mavromatis et al., 2001). On the other hand, it has one major drawback: the variables upon which the optimization is based are often only slightly correlated with the parameters estimated, so that the physiological significance of the parameter itself may be compromised. This risk is greater if several parameters are estimated simultaneously. This is not necessarily serious if the aim is to predict the production of a variety within the range of environments in the estimation dataset but not as a criterion for breeding. One way to combine direct and indirect parameter estimation would be to use a Bayesian approach (see Chapter 4), which allows one to take account of prior knowledge about the value of a parameter in the estimation procedure. In terms of reflecting the various performances of several varieties, the usefulness of this method is unquestionable and does not pose any particular problems.

4.3. Discussion

The link between the available database and the method chosen to estimate parameters is not exclusive. However, the ease with which genotypic parameters are estimated often depends on the structure of the model and its functional relationships. To illustrate this, the case of BETHA can be used (Loyce et al., 2002), although this model is static. The parameters of the agronomic submodel simulating yield losses in the presence of disease are the GEVES varietal resistance scores. Thus in this case, the parameters are measured systematically during the registration phase of any new variety, and then provided to the public *via* an official catalogue of varieties. In this example, the choice of model to reflect the effect of disease on yield losses was guided by the availability of parameters, updated systematically for all new varieties. Thus the model is constantly being adjusted to allow for new genotypes, without any additional specific experiments. An example of a very different approach is provided by a model of maize leaf growth with genotypic parameters which require complicated and specific experiments for their estimation.

Finally, the method used to estimate parameters has to be chosen not only according to the objective (need for access to the biological significance of parameters, or for satisfactory predictive quality of the model) but also according to the ease with which it can be extrapolated for new varieties. In some cases, it may be more desirable to apply a clearly defined measurement protocol (e.g. disease resistance score protocol) for direct measurement of the parameter, whereas in other cases it may be necessary to use optimization methods, which allow for estimation of several parameters at the same time. One possibility may be to estimate parameters using intermediate variables of the model which are more closely related to the parameter than are final outputs. Costs must also be taken into account when choosing an estimation method, especially when a goal is to routinely estimate parameters of the model for newly registered varieties. If a new experimental network is to be set up for parameter estimation trials, one must ensure that its sites are representative of the region in which the model is to be used. If the BETHA model is compared with the model proposed by Bastiaans (1993) in terms of simulating the effect of disease on yield, it is undeniable that the disease resistance scores (BETHA parameters) will be more easily and rapidly available than Bastiaans' Betha parameter, which requires measurement of photosynthesis of healthy and diseased plants under controlled conditions. Another example also illustrates the flexibility of model adaptation methods to genotypes. The vertical distribution of grain numbers in peas, proposed by Roche and Jeuffroy (2000), can be transposed to numerous varieties by measuring the mean weight per grain from each variety, a variable that is normally recorded by registration and development organizations.

The choice of estimation method has also to be considered as a function of the ultimate objective:

- If the model is to be used in plant breeding applications, parameters that are rapidly and individually measurable are essential, and they must have physiological meaning.
- If the objective is the *a posteriori* analysis of variations in behavior, parameters can either be optimized or directly measured, so that newly registered varieties can be typed more easily without waiting for several years of field experimentation.

5. Examples

5.1. Direct measurement of parameters vs *minimum error approach*

In order to assess the consequences of parameter estimation methods on parameter values, the quality of model adjustment to data, and the predictive quality of the model, we estimated the three parameters of the "potential crop function" module in Azodyn (see details in Chapter 4). These three parameters represent biomass radiation use efficiency (EBMAX), the ratio between the leaf area index and the critical nitrogen level of the crop (D), and the maximum nitrogen absorption rate by the crop (VMAX).

The data were obtained from the 1999 trial carried out at Grignon containing five varieties: Soissons, Baltimore, Cockpit, Florence-Aurore, and Trémie. The measurements taken during the trial consisted of weekly monitoring of leaf area index, above-ground biomass, and the quantity of nitrogen accumulated in above-ground parts of the crop, between the end of January and flowering (early June). Based on results obtained by Akkal (1998), we assumed stability of parameters between genotypes for the equation linking radiation use efficiency and leaf area index. On this basis, we estimated biomass conversion efficiency by linking the biomass produced at different dates with the radiation intercepted by the crop on the same dates. Then, to estimate the maximum nitrogen absorption rate, we calculated, between two successive sampling dates, the ratio between the quantity of nitrogen accumulated in the crop between these two dates and the sum of degree-days between the two dates. The maximum value of this ratio, over the entire period considered, was taken as the VMAX parameter. A second method of parameter estimation was also done. Here weighted least squares were used to adjust all the parameters to the three observed variables. The values of the three parameters, obtained using each method, are shown for the five varieties in Table 1.

The quality of model adjustment was estimated by comparing measured and simulated values on the non-limiting treatment of the trial. The quality of model prediction was estimated on experimental treatments of the same trial, during periods of nitrogen deficiency. The results (see Table 2) show that the quality of adjustment was often better in the case of optimization, but predictive quality was similar using the two methods, with either of them being better on different occasions. In this example, the two methods produced equivalent results.

A second illustration concerns the demonstration of a genotypic difference between parameters. Using the least squares method to estimate the three parameters in the "potential function" module of Azodyn, it was found that numerous triplets of values led to similarly good fits to the data for the three variables concerned (Fig. 3). For example, both the triplets EBMAX $= 3.76$, D $= 0.028$, VMAX $= 0.32$ and EBMAX $= 2.67$, D $= 0.04$, VMAX $= 0.35$, led to $R^2 = 0.983$ for the "nitrogen uptake" variable. Using this approach, it was not possible to determine differences in the parameters between different varieties. In contrast, the direct measurement method of estimation produced differences in the parameters between genotypes (Table 1).

Finally, it should be pointed out that parameter estimation by optimization based on a variable which is only indirectly linked to the parameter, leads to a greater risk of differences between parameter values estimated using different datasets than would be the case for directly measured parameter values. If we return to the above example, we can see that for Soissons, the range of EBMAX values producing a good simulation of above-ground biomass was relatively small and centered around the directly measured value, compared to the range of values obtained when the parameter was adjusted to two other variables (LAI and QN less directly linked to EBMAX (Fig. 4).

5.2. Characterizing genotype variability

Analysis of the behavior of several genotypes using the structure of a model may make it possible to identify those modules where genotypic variability is most often observed. Most crop simulation models are based on a series of modules: a phenological module, describing the dates of onset of development stages (e.g. for leaves: initiation, appearance, end of expansion, yellowing, senescence) or the duration of stages (vegetative

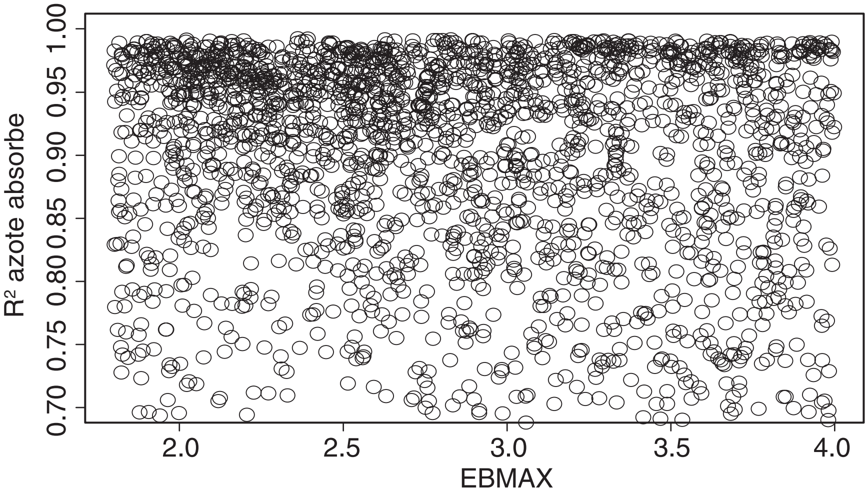
Table 1. Values for 3 parameters in the "potential function" module of Azodyn, obtained for 5 varieties using two different methods, direct measurement (mes.) and least squares optimization (opt.). See model equations for the precise definition of the parameters: EBMAX is the maximum radiation use efficiency (without any stress), D is the ratio LAI/critical nitrogen uptake, VMAX is the maximum value of the nitrogen uptake rate (kg/ha/degree-day).

	Baltimore		Cockpit		Florence-Aurore		Soissons		Trémie	
	Mes.	Opt.	Mes.	Opt.	Mes.	Opt.	Mes.	Opt.	Mes.	Opt.
EBMAX	3.01	3.00	2.80	2.98	3.11	3.08	3.25	3.22	3.20	3.11
	(0.06)	(0.12)	(0.03)	(0.11)	(0.08)	(0.13)	(0.09)	(0.13)	(0.08)	(0.20)
D	0.039	0.035	0.036	0.034	0.024	0.026	0.026	0.027	0.028	0.030
	(0.002)	(0.002)	(0.002)	(0.0015)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
VMAX	0.55	0.41	0.42	0.34	0.49	0.37	0.39	0.43	0.39	0.41
		(0.05)		(0.04)		(0.04)		(0.01)		(0.03)

Table 2. Comparison of adjustment and prediction errors of the "potential function" module of Azodyn, for the two parameter estimation methods of Table 1. The 3 variables considered are: above-ground crop biomass (DM in kg/ha), leaf area index (LAI) and nitrogen levels in above-ground parts (N in kg/ha). RMSE and RMSEP are respectively root mean squared error and root mean squared error of prediction.

	Baltimore		Cockpit		Florence-Aurore		Soissons		Trémie	
	Mes.	Opt.	Mes.	Opt.	Mes.	Opt.	Mes.	Opt.	Mes.	Opt.
DM										
RMSE	290	246	225	89	373	258	352	325	385	376
RMSEP	967	790	855	875	679	586	651	680	481	484
LAI										
RMSE	0.85	0.85	0.68	0.65	0.29	0.25	0.45	0.36	0.42	0.33
RMSEP	0.80	0.71	0.68	0.67	0.39	0.33	0.50	0.55	0.56	0.61
N										
RMSE	29.8	36.6	15.7	19.8	7.4	13.0	15.9	14.2	18.0	17.1
RMSEP	18.2	16.6	18.0	16.3	17.0	16.0	15.2	16.0	12.4	12.6

Soissons 1999:



Cockpit 1999:

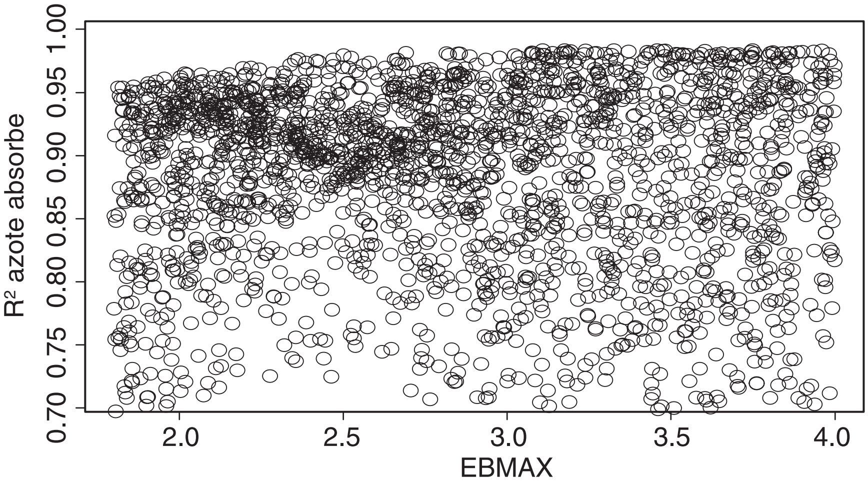


Figure 3. Relationship between the EBMAX parameter values in Azodyn and the coefficient of determination between measured and simulated values for the quantity of nitrogen accumulated in above-ground parts, for two varieties, Soissons and Cockpit (calculations performed on data from the Grignon trial, 1999).

or reproductive); a module describing the evolution of surfaces available for exchanges (leaves, roots); a module for biomass production (often based on radiation use efficiency); a module for biomass distribution; one or several modules simulating the effects of abiotic stresses (water, nitrogen) and finally, physical modules estimating the state of the environment (availability of water or nitrogen, etc.) (e.g. see Jones et al., 2002).

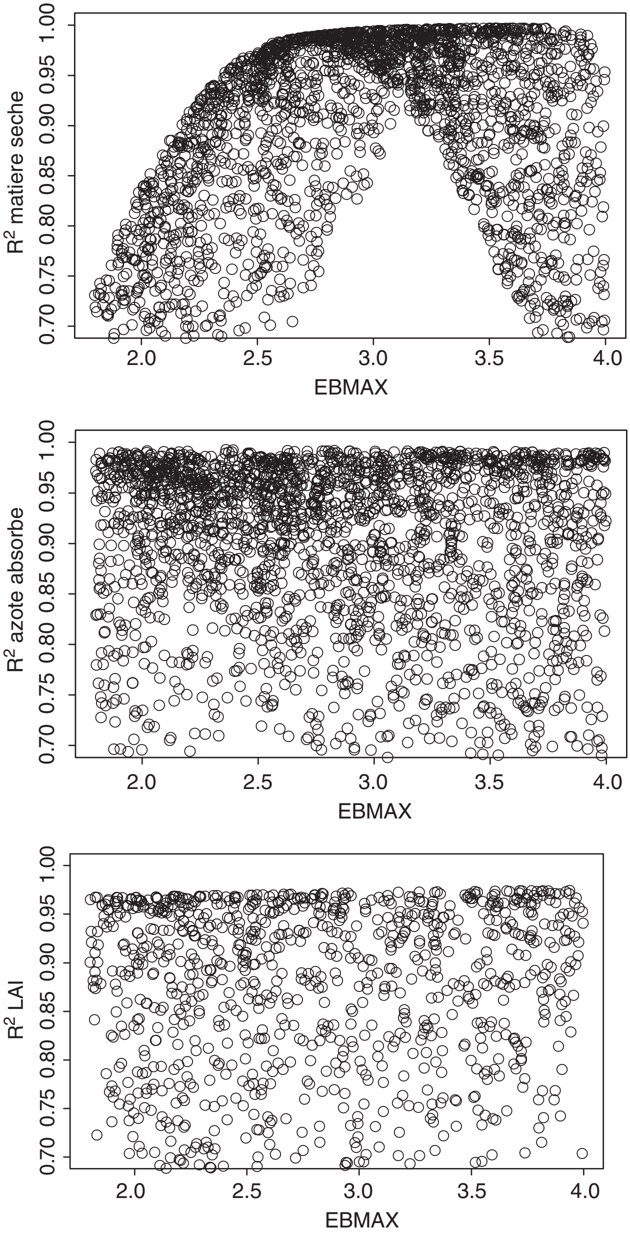


Figure 4. Relationship between EBMAX parameter values in Azodyn and the coefficient of determination between measured and simulated values for the quantity of nitrogen accumulated in above-ground parts, for the Soissons variety (calculations performed on data from the Grignon trial, 1999).

This type of model was used to analyze the behavior of five sunflower genotypes, representative of three time periods: the 1970s for Mirasol, the 1980s for Albena and the 1990s for Heliasol, Melody and Prodisol (Lecoeur, 2002). These varieties exhibit reproducible differences in yield up to a maximum of 30%, which are correlated with the date of registration (Vear et al., 2002), suggesting that considerable genetic advances were achieved over a period of 30 years. Figure 5 summarizes the results, using Monteith's biomass production formalization (1977). The analysis showed that the varieties studied did not exhibit genetic variability in radiation use efficiency (RUE) or in the distribution of biomass between different parts of the plant, whether dynamically or at harvest. In contrast, the course of radiation interception efficiency (RIE_i) presented differences which, when cumulated over the growth cycle, produced differences in radiation used of up to 10%. These differences arose from the architecture of above-ground plant parts and, in particular, leaf surface profiles. It is noted that the most recent varieties produced a lower total number of leaves than older varieties, and the larger leaves were closer to the top of the plant. This plant architecture confers an advantage in radiation interception, which can be quantified by calculating the extinction coefficient (k) using Beer's law, linking the leaf area index of the crop with radiation interception efficiency. More recent varieties exhibited a higher k coefficient. The final difference identified was more classic, since it concerned phenology. Different varieties presented different life-spans for their foliage, which was indicated by the duration of the flowering – maturity stage. All these differences were sufficient to explain the yield differences observed.

This example shows that, in different sunflower varieties, the main differences in behavior concerned the architecture of above-ground parts and phenology, while biomass production and distribution did not demonstrate any variability. Phenological aspects are easier to integrate into standard models. However, differences in the evolution of surfaces for exchanges and their performance result from smaller differences in plant structure that are often more difficult to integrate. This example also demonstrates the usefulness of basing the analysis of genotypic variability on a crop model (and not only on analytical experiments) in order to identify the origins of differences in varietal behavior.

Other crop models have also been used to evaluate genotypic differences among old *vs* new varieties to determine mechanisms that led to yield improvement. For example, in the study presented by Boote et al. (2003), two new and one old soybean varieties were planted at two different locations in Iowa and Illinois for two years. The new varieties yielded about 17% higher than the old variety. Phenology and growth measurements were made periodically during the season in each trial and used to estimate genotypic parameters for the CROPGRO-Soybean model. Boote et al. (2003) found that a combination of parameters differed between new and old varieties and could explain differences in yield. They found that new cultivars had faster pod addition (about 30%), longer seed filling phase (about 10%), higher leaf photosynthesis (about 9%) and slower leaf nitrogen loss with age (about 10%). They concluded that the potential yield differences between these old and new soybean varieties could be explained by the parameters in the model that control the above processes. Much of the yield gain was related to partitioning of photosynthesis and to a longer duration of seed filling.

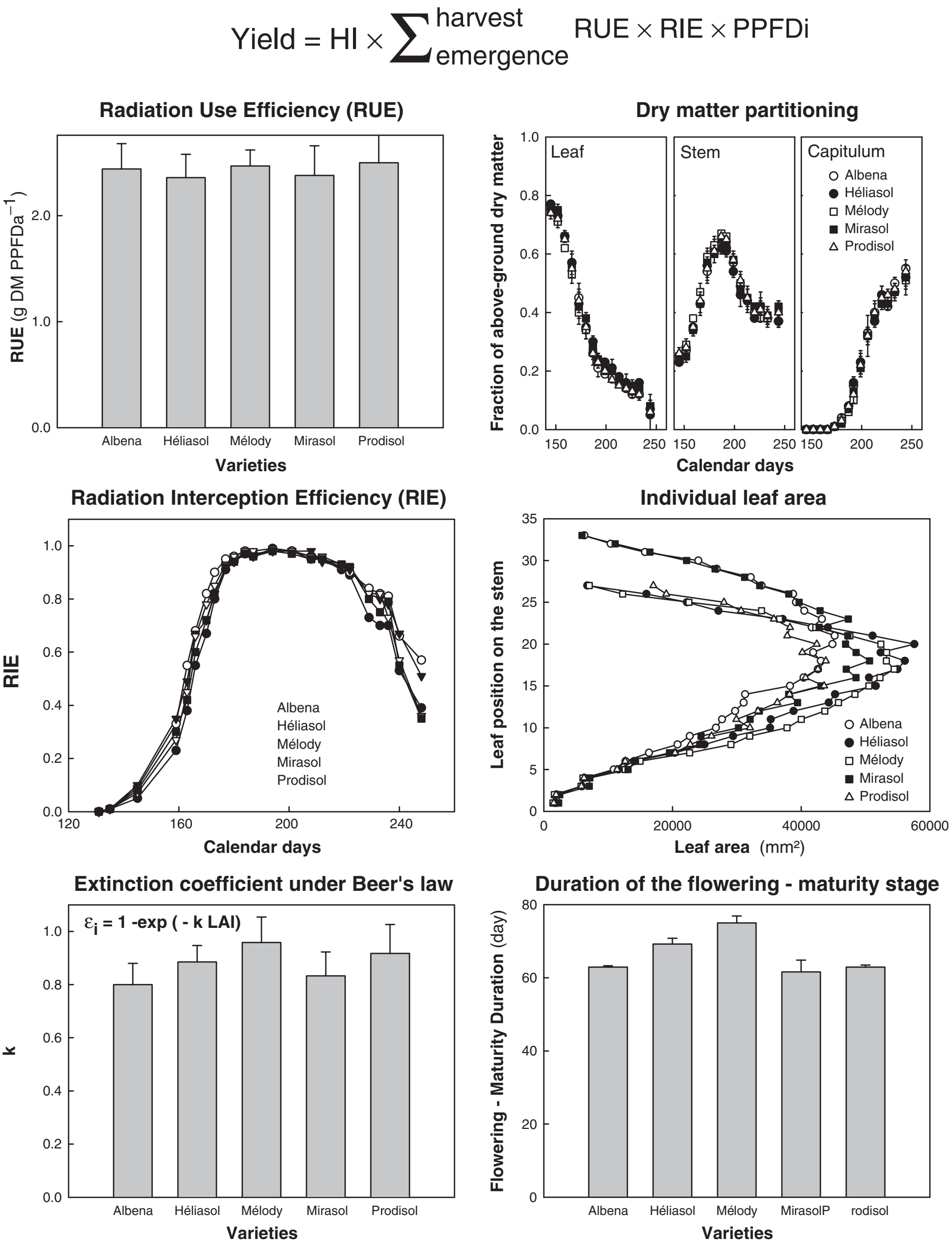


Figure 5. Analysis of the behavior of 5 sunflower varieties using Monteith's (1977) energy approach formalism and identification of parameters presenting genetic variability (Lecoeur, 2002). HI, harvest index of above-ground dry matter; PPFD, photosynthetic photon flux density; RIE, radiation interception efficiency; RUE, radiation use efficiency. Mirasol was registered in the 1970s, Albena in the 1980s and Heliasol, Mélody and Prodisol in the 1990s.

5.3. Breeding: ideotype design for target environments

Strong genotype × environment × management interactions in most agricultural systems make it necessary to use genotypes that are adapted to the specific agro-environments. Shorter et al. (1991) suggested that simple biological models could help integrate plant breeding and crop physiology to evaluate adaptation of genotypes to target environments. Although crop models lack the ability to describe all of the complexities of genotype response, they contain powerful relationships, based on physiologically sound mechanisms, and can be used as a tool to identify genotypes suited for particular climatic zones (Hammer et al., 1996).

Several studies have been reported that demonstrate this approach. Kropff et al. (1995) used models to optimize the performance of specific genotypes of potato in target environments using the LINTUL-POTATO model. The key to maximizing potential yield was the proper timing of tuber initiation. They determined the cultivar characteristics with respect to temperature and daylength response that give the highest yield in a particular environment. Kropff et al. (1995) and in more detail, Aggarwal et al. (1997) determined the importance of various traits for maximum yield potential in the tropical dry season and the wet season using ORYZA1 (Kropff et al., 1994). The critical model parameter values varied with respect to those of the variety IR72 to simulate the effect of a change in specific leaf area, spikelet growth factor, potential grain weight, maximum leaf N concentration, and crop development rates during juvenile phase and grain filling period. Results showed that no trait individually or in combination could provide more than 5% yield increase for usual N management practices. These genotypes were not able to express these traits in these somewhat limited-N environments. With improved N management, yield potential was predicted to be significantly increased (>30%) by an increased sink capacity, maintenance of high leaf N content and a longer grain filling duration (Aggarwal et al., 1997).

Boote et al. (2001) used CROPGRO-Soybean and CERES-Maize models to determine the optimal sets of genotype parameters for soybean and maize, respectively, growing in two environments that differ considerably in soil and climate (Ames, Iowa, Latitude 42.00°N, Longitude 93.77°W; *vs* Gainesville, Florida, Latitude 29.63°N, Longitude 82.37°W). Variations in parameters resulted in widely varying crop season lengths within a location, and these durations differed considerably between sites. For example, a genotype with fairly low photoperiod sensitivity (i.e. one belonging to maturity group II) matured in 84 and 121 days at Gainesville and Ames, respectively, and yielded 834 *vs* 1936 kg ha^{-1}. In contrast, a genotype that was very sensitive to photoperiod (i.e. one in maturity group VII) matured in 145 days at Gainesville and yielded 2741 kg ha^{-1} whereas that genotype did not mature and gave no yield at all at Ames.

There is potential for use of this ideotype design application in plant breeding programs. However, this approach has not yet been adopted by plant breeders to our knowledge. Reasons for this lack of adoption appear to be due neither to a lack of interest on the part of breeders, since they are indeed looking for tools to help them be more effective in breeding plants for target environments, nor to a lack of interest on the part of crop modelers, who have proposed such uses and demonstrated the potential value of the approach. Reasons seem to be complex. However, as plant breeding invests more efforts in molecular genetics techniques and as crop modelers learn to link control of physiological processes to specific combinations of genes, powerful crop model-based tools will be developed to help plant breeders design ideotypes for target environments. (See Chapter 11 for a discussion of research linking crop model parameters with molecular markers.)

6. Conclusions

Without trying to answer the many questions raised by the estimation of genotypic parameters in crop models, we have raised a certain number of problems of current importance which are still the subject of considerable debate, in a relatively recent research area.

In the introduction, we presented the two main objectives targeted while introducing varietal effects into a crop simulation model. If the objective is to identify and analyze differences in functional parameters between genotypes and their relationships, a method that adjusts parameters may be misleading, because parameter values estimated by minimizing the difference between measured and calculated values are difficult to interpret in terms of varietal effects. Factors other than variety can determine parameter values, such as model structure or the values of other parameters, etc. On the other hand, if the objective is to improve model performance (improvement in predictive quality and management recommendations), less attention will be paid to the actual values of varietal parameters than model performance. In this case, adjustment methods are useful. However, before introducing new parameters, and thus increasing their total number, it is necessary to consider whether this might degrade predictive quality, as has already been demonstrated in other examples (see Chapter 4).

Two points should be emphasized. First, the desire to adapt models to a range of genotypes affects the traditional relationship between experimentation and modeling. Subsequent model adaptation to new genotypes must be taken into account when the model is initially designed and adapted to a small number of genotypes.

In this chapter, we demonstrated the use of indirect methods to estimate genotype parameters. It is important to keep in mind the fact that prediction errors can be much more than the residual error obtained in the parameter estimation exercise. One must evaluate the parameters using independent data to estimate prediction errors. If data are limited, and this is typically the case, one should at least use cross validation to make sure that parameters are not over-fit and that an estimation of prediction error is made.

In addition, this desire to adapt models to genotypes also calls into question the choice of model structure. Thus, recent advances in physiology may modify the representation of a plant in future crop models. It may not be possible to efficiently integrate genetic variability into classical canopy function models. Some of their modules may require considerable modification. There is then a risk that these models will involve too many parameters, difficult to estimate for a large number of genotypes. A different adaptation method may be to use static models (such as BETHA), which would allow the inclusion of considerable data and expertise on varieties, and the exploitation of varietal characteristics in the form of parameters. The choice of approach will mainly depend upon the objective for which the model is being used. One could also consider a combination of the two approaches. By estimating as many parameters as possible using direct measurements, one will have more confidence in those as well as in other parameters that have been estimated using a minimum error criterion.

References

Aggarwal, P.K., Kropff, M.J., Cassman, K.G., Ten Berge, H.F.M., 1997. Simulating genotypic strategies for increasing rice yield potential in irrigated, tropical environments. Field Crops Research 51, 5–17.

Agüera, F., Villalobos, F.J., Orgaz, F., 1997. Evaluation of sunflower (*Helianthus annuus*, L.) genotypes differing in early vigour using a simulation model. European Journal of Agriculture 7, 109–118.

Akkal, N., 1998. Pilotage de la fertilisation azotée du blé d'hiver sur la base d'une évaluation précoce de la réflectance radiométrique ou du taux de couverture du sol, en vue d'une application à l'agriculture de précision. Thèse de Doctorat, INA P-G, p. 198.

Barbottin, et al., 2002. The genotypic variability of nitrogen remobilisation efficiency. Proceedings of the VIIth E.S.A. Congress, Cordoba.

Barbottin, A., 2004. Utilisation d'un modèle de culture pour évaluer le comportement des génotypes: Pertinence de l'utilisation d'Azodyn pour analyser la variabilité du rendement et de la teneur en protéines du blé tendre. Thèse de doctorat, INA P-G, p. 178.

Barbottin, A., Jeuffroy, M.H., 2004. The use of a crop model simulating the grain protein content of winter wheat to define breg targets. Workshop on Modelling Quality Traits and their Genetic Variability for Wheat, Clermont-Ferrand (FRA), July 18–21.

Barbottin, A., Lecomte, C., Bouchard, C., Jeuffroy, M.H., 2005. Nitrogen remobilisation during grain filling in wheat: genotypic and environmental effects. Crop Science, sous presse.

Bastiaans, L., 1993. Understanding yield reduction in rice due to leaf blast. Wageningen Agricultural University.

Bernicot, M.H., 2002. Qualité: il faut augmenter le taux de proteines. Grandes Cultures Infos, July 08.

Bindraban, P.S., 1997. Bridging the gap between plant physiology and breeding: identifying traits to increase wheat yield potential using systems approaches. Ph.D. Thesis, Wageningen Agricultural University, Wageningen, The Netherlands.

Boote, K.J., Jones, J.W., Pickering, N.B., 1996. Potential uses and limitations of crop models. Agronomy Journal 88(5), 704–716.

Boote, K.J., Jones, J.W., Hoogenboom, G., 1998. CROPGRO model for grain legumes. In: Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds), Understanding Options for Agricultural Production. Kluwer Academic Press, Boston, pp. 99–128.

Boote, K.J., Kropff, M.J., Bindraban, P.S., 2001. Physiology and modelling of traits in crop plants: Implications for genetic improvement. Agricultural Systems 70, 395–420.

Boote, K.J., Jones, J.W., Batchelor, W.D., Nofziger, E.D., Myers, O., 2003. Genetic coefficients in the CROPGRO-soybean model: links to field performance and genomics. Agronomy Journal 95, 32–51.

Brancourt-Hulmel, M., Lecomte, C., 2003. Effect of environmental varieties on genotype × environment interaction of winter wheat: a comparison of biadditive factorial regression to AMMI. Crop Science 43, 608–617.

Brancourt-Hulmel, M., Biarnès-Dumoulin, V., Denis, J.B., 1997. Points de repère dans l'analyse de la stabilité et de l'interaction genotype-milieu en amelioration des plantes. Agronomie 17, k219–246.

Brancourt-Hulmel, M., Denis, J.B., Lecomte, C., 2000. Determining environmental covariates which explain genotype × environment interaction in winter wheat through probe genotypes and biadditive factorial regression. Theoretical and Applied Genetics 100, 285–298.

Champeil, A., 2001. Vers une prévision opérationnelle de la teneur en protéines des grains de blé tendre d'hiver: estimation des variables d'entrée et analyse de l'effet variétal dans le modèle Problé. Mémoire de DEA "Adaptation des Plantes Cultivées", INA PG, Paris Sud, p. 20.

Chapman, S.C., Cooper, M., Hammer, G.L., Butler, D., 2000. Genotype by environment interactions affecting grain sorghum. II. Frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. Australian Journal of Agricultural Research 50, 209–222.

Demotes-Mainard, S., Doussinault, G., Meynard, J.M., 1996. Abnormalities in the male developmental programme of winter wheat induced by climatic stress at meiosis. Agronomie 16, 505–515.

Denis, J.B., 1988. Two way analysis using covariates. Statistics 19, 123–132.

Dorsainvil, F., 2002. Evaluation, par modélisation, de l'impact environnemental des modes de conduite des cultures intermédiaires sur les bilans d'eau et d'azote dans les systèmes de culture. Thèse de Doctorat, INAP-G, p. 122.

Fargue, A., 2002. Maîtrise des flux de gènes chez le colza: Etude *ex-ante* de l'impact de différentes innovations variétales, Thèse de doctorat, INA P-G, p. 168.

Felix, I., Loyce, C., Bouchard, C., Meynard, J.M., Bernicot, M.H., Rolland, B., Haslé, H., 2002. Asssocier des variétés rustiques à des niveaux d'intrants réduits. Intérêts économiques et perspectives agronomiques. Perspectives Agricoles 279, 30–35.

Franck, N., 2001. Utilisation de la modélisation 3D pour analyser l'effet du rayonnement incident sur la mise en place de la surface foliaire de trois génotypes d'Arabidopsis thaliana. DEA Développement et Adaptation des Plantes, Université Montpellier 2, Université de Perpignan, ENSAM, Montpellier, p. 22.

Girard, M.L., 1997. Modélisation de l'accumulation de biomasse et d'azote dans les grains de blé tendre d'hiver:simulation de la teneur en protéines à la récolte. Thèse de doctorat, INA-PG, p. 96.

Goyne, P.J., Meinke, H., Milroy, S.P., Hammer, G.L., Hare, J.M., 1996. Development and use of a barley crop simulation model to evaluate production management strategies in north-eastern Australia. Australian Journal of Agricultural Research 47, 997–1015.

Grimm, S.S., Jones, J.W., Boote, K.J., Hesketh, J.D., 1993. Parameter estimation for predicting flowering date of soybean cultivars. Crop Science 33, 137–144.

Hammer, G.L., Sinclair, T.R., Boote, K.J., Wright, G.C., Meinke, H., Bell, M.J., 1995. A peanut simulation model: I. model development and testing. Agronomy Journal 87, 1085–1093.

Hammer, G.L., Butler, D., Muchow, R.C., Meinke, H., 1996. Integrating physiological understanding and plant breeding via crop modelling and optimisation. In: Cooper, M., Hammer, G.L. (Eds), Plant Adaptation and Crop Improvement. CAB International, Wallingford, UK, pp. 419–441.

Hammer, G.L., Chapman, S.C., Snell, P., 1999a. Crop simulation modelling to improve selection efficiency in plant breeding programs. Proceedings of the Ninth Assemble Wheat Breeding Society of Australia, Towomba, September 1999, pp. 79–85.

Hammer, G.L., Keating, B., Meinke, H., Carberry, P., Freebairn, D., Probert, M., Holzworth, D., 1999b. An integrated systems approach to improving agricultural systems using the agricultural production systems simulator APSIM. In: Donatelli, M., Stockle, C., Villalobos, F., Villar Mir, J.M. (Eds), Proceedings of the International Symposium Modelling Cropping Systems, Lleida, Spain, June 21–23, pp. 31−37.

Hulmel, M., 1999. Expliquer l'interaction génotype × milieu par des génotypes révélateurs chez le blé tendre d'hiver. Thèse de Doctorat, ENSAR, p. 152.

Hunt, L.A., Pararhasingham, S., Jones, J.W., Hoogenboom, G., Immamura, D.T., Ogoshi, R.M., 1993. GENCALC: software to facilitate the use of crop models for analyzing field experiments. Agronomy Journal 85, 1090–1094.

Jamieson, P.D., Porter, J.R., Goudriaan, J., Ritchie, J.T., van Keulen, H., Stol, W., 1998. A comparison of the models AFRCWHEAT2, CERES-Wheat, Sirius, SUCROS2 and SWHEAT with measurements from wheat grown under drought. Field Crops Research 55, 23–44.

Jeuffroy, M.H., Girard, M.L., Barré, C., 2000. Qualité du blé tendre: comprendre et prévoir la teneur en protéines des grains. Perspectives in Agriculture 261, 24–31.

Jones, H.G., 1992. Plants and Microclimate: A Quantitative Approach to Environmental Plant Physiology. Cambridge University Press, UK, p. 428.

Jones, J.W., Hansen, J.W., Royce, F.S., Messina, C.D., 2000. Potential benefits of climate forecasting to agriculture. Agriculture, Ecosystems and Environment 82, 169–184.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2002. The DSSAT cropping system model. European Journal of Agronomy 18(3–4), 235–265.

Justes, E., Mary, B., Meynard, J.-M., Machet, J.-M., Thelier-Huche, L., 1994. Determination of a critical nitrogen dilution curve for winter wheat crops. Annals of Botany 74, 397–407.

Keating, B.A., Godwin, D.C., Watiki, J.M., 1991. Optimising nitrogen inputs in response to climate risk. In: Muchow, R.C., Bellamy, J.A. (Eds), Climatic Risk in Crop Production: Models and Management for the Semiarid Tropics and Subtropics. CAB International, Wallingford, Oxford, UK, pp. 329–358.

Kropff, M.J., Van Laar, H.H., Matthews, R.B. (Eds), 1994. ORYZA1: An ecophysiological model for irrigated rice production. SARP Research Proceedings, AB-DLO, Wageningen and IRRI, Los Baños, Philippines, p. 110.

Kropff, M.J., Haverkort, A.J., Aggarwal, P.K., Kooman, P.L., 1995. Using systems approaches to design and evaluate ideotypes for specific environments. In: Bouma, J., Kuyvenhoven, A., Bouman, B.A.M., Luyten, J.C., Zandstra, H.G. (Eds), Eco-regional Approaches for Sustainable Land Use and Food Production. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 417–435.

Le Gouis, J., Pluchard, P., 1996. Genetic variation for nitrogen use efficiency in winter wheat (*Triticum aestivum* L.). Euphytica 92, 221–224.

Lecoeur, J., 2002. Rapport d'étape du projet "Productivité du tournesol" INRA – Promosol, dans Actes du "Carrefour du tournesol 2002", Clermont-Ferrand, France, 30 janvier 2002.

Lecoeur, J., Ney, B., 2003. Change time with in potentiel radiation-use efficiency in field pea. European Journal of Agriculture sous press.

Liu, W.T.H., Botner, D.M., Sakamoto, C.M., 1989. Applications of CERES-maize model to yield prediction of a brazilian maize hybrid. Agriculture and Forest Meteorology 45, 299–312.

Loyce, C., 1998. Mise au point d'itinéraires techniques pour un cahier des charges multicritère: le cas de la production de blé éthanol en Champagne crayeuse. Thèse de Doctorat, INA P-G, p. 196.

Loyce, C., Rolland, B., Bouchard, C., Doussinault, G., Haslé, H., Meynard, J.M., 2001. Les variétés de blé tolérantes aux maladies: une innovation majeure à valoriser par des itinéraires techniques économes. Perspectives Agricoles 268, 50–56.

Loyce, C., Rellier, J.P., Meynard, J.M., 2002. Management planning for winter wheat with multiple objectives (1): the BETHA system. Agricultural Systems 72, 9–31.

Makowski, D., Naud, C., Monod, H., Jeuffroy, M.H., Barbottin, A., 2005. Global sensitivity analysis for calculating the contribution of genetic parameters to the variance of crop model prediction. Reliability Engineering and System Safety, in press.

Mavromatis, T., Boote, K.J., Jones, J.W., Irmak, A., Shinde, D., Hoogenboom, G., 2001. Developing genetic coefficients for crop simulation models with data from crop performance trials. Crop Science 41, 40–51.

Mavromatis, T., Boote, K.J., Jones, J.W., Wilkerson, G.G., Hoogenboom, G., 2002. Repeatability of model genetic coefficients derived from soybean performance trials across different states. Crop Science 42, 76–89.

Messina, C.D., Jones, J.W., Boote, K.J., Vallejos, C.E., 2006. A gene-based model to simulate soybean development and yield responses to environment. Crop Science 46, 456–466.

Meynard, J.M., Ribeyre, C., Boudon, O., Laurent, E., 1988. Pour mieux connaître les variétés de blé: analyser l'élaboration de leur rendement. Perspectives Agricoles 131, 17–23.

Monteith, J.L., 1977. Climate and the efficiency of crop production in Britain. Philosophical Transactions of Royal Society of London, Ser. B 281, 277–294.

Philibert, M.C., 2001. Ajuster la fertilisation azotée du blé à la variété avec le modèle Azodyn. Mémoire de fin d'études, ENITA, Clermont-Ferrand, p. 40.

Piper, E.L., Boote, K.J., Jones, J.W., Grimm, S.S., 1996. Comparison of two phenology models for predicting flowering and maturity date of soybean. Crop Science 36, 1606–1614.

Piper, E.L., Boote, K.J., Jones, J.W., 1998. Evaluation and improvement of crop models using regional cultivar trial data. Applied Engineering in Agriculture 14(4), 435–446.

Reymond, M., 2001. Variabilité Génétique des réponses de la croissance foliaire du maïs à la température et au déficit hydrique. Combinaison d'un modèle écophysiologique et d'une analyse QTL. Thèse de l'Ecole Nationale Supérieure Agronomique de Montpellier, Montpellier, France, p. 70.

Roche, R., Jeuffroy, M.H., 2000. A model to calculate the vertical distribution of grain number in pea. Agronomy Journal 92, 663–671.

Royce, F., Jones, J.J., Hansen, J.W., 2002. Model-based optimization of crop management for climate forecast applications. Transactions of the ASAE 44, 1319–1327.

Shorter, R., Lawn, R.J., Hammer, G.L., 1991. Improving genotypic adaptation in crops: a role for breeders, physiologists and modellers. Experimental Agriculture 27, 155–175.

Sinclair, T.R., Muchow, R.C., 1999. Radiation use efficiency, Advances in Agronomy 65, 215–265.

Sinclair, T.R., Seligman, N., 2000. Criteria for publishing papers on crop modeling. Field Crops Research 68, 165–172.

Travasso, M.I., Magrin, G.O., 1998. Utility of CERES-barley under Argentine conditions. Field Crops Research 57, 329–333.

van Eeuwijk, F.A., 1995. Linear and bilinear models for the analysis of multi-environment trials: I. an inventory of models. Euphytica 84, 1–7.

Van Sanford, D.A., MacKown, C.T., 1986. Variation in nitrogen use efficiency among soft red winter wheat genotypes. Theoretical and Applied Genetics 72, 158–163.

Van Sanford, D.A., MacKown, C.T., 1987. Cultivar differences in nitrogen remobilization during grain fill in soft red winter wheat. Crop Science 27, 295–300.

Vear, F., Bony, H., Joubert, G., Tourvieille de Labrouhe, D., Pauchet, I., Pinochet, X., 2002. The results of 30 years of sunflower breeding for France. OCL sous presse.

Weir, A.H., Bragg, P.L., Porter, J.R., Rayner, J.H., 1984. A winter wheat crop simulation model without water or nutrient limitations. Journal of Agricultural Science, Cambridge 102, 371–382.

White, J.W., Hoogenboom, G. 1996. Simulating effects of genes for physiological traits in a process-oriented crop model. Agronomy Journal 88, 416–422.

Wilkerson, G.G., Buol, G.S., Fallick, J., Boote, K.J., Jones, J.W., Porter, C., 2001. SoyCCE: soybean cultivar coefficient estimator version 1.0 user's manual, Research Report 190. Crop Science Department, NCSU, Raleigh, NC.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 11

Model-assisted genetic improvement of crops

C.D. Messina, K.J. Boote, C. Löffler, J.W. Jones and C.E. Vallejos

1. Introduction	**309**
2. Crop model use for environmental characterization	**311**
3. Crop model use for analysis of past genetic improvement	**314**
4. Crop model use to design new cultivars	**317**
5. Crop models that incorporate gene information	**319**
6. Use of gene-based crop models in plant breeding applications	**325**
7. Discussion	**330**
References	**331**

1. Introduction

Modern plant breeding provides genetic solutions to improve plant productivity. The goal of a breeding program is to develop improved cultivars by manipulating available genetic variability to create new allelic combinations best adapted to target environments and applications. Plant breeding can be viewed as an optimization process during which the plant breeder guides the search for new cultivars by integrating knowledge from several scientific disciplines. A few of the traits manipulated by breeders are controlled by single genes, but most breeding efforts deal with traits controlled by several genes, such as organ size, days to maturity, photoperiod sensitivity and yield. In this regard, breeders have used the classical view of quantitative genetics in which the phenotype (P) is the result of the expression of the genotype (G), the environment (E) and the interactions between the genotype and the environment (GEI).

The advent of molecular marker technology has opened new opportunities in quantitative genetics by providing a way to separate complex traits. This is accomplished by using linkage information between molecular markers and quantitative trait loci (QTL) (Lander and Botstein, 1989; Lander and Schork, 1994). This approach has become feasible for a relatively large number of crops with extensive molecular marker-based linkage maps (Phillips and Vasil, 2001). With the inclusion of QTL locus, the phenotype response can be estimated by a linear model that sums the contribution of each allele at its corresponding QTL, the effect of the environment and the interactions between the alleles and the environment:

$$P = \mu + \sum b_i g_i + \sum (b_i g_i \times E) + E + e$$

where P is the phenotype response, μ is the overall experimental mean, b_i is the phenotypic effect of an allele at QTL i, g_i is the allele dosage at QTL i, E is the effect of the environment, and e denotes the error term.

However, many plant physiological effects are highly non-linear. In view of this complexity, estimating a phenotype response will require one to measure several traits and handle a large number of interactions arising from the large number of genes governing the behavior of complex traits. Selection indexes and statistical methods such as BLUPS (Henderson, 1975) were developed to handle multiple and complex traits in selection. However, the assumption of linearity of the relationships among factors used in statistical methodologies is still a major obstacle for their use in multiple trait selection. The widespread occurrence of gene by gene interactions or epistasis in complex traits limits the power of purely statistical approaches to develop informative selection indices. Despite the importance of epistasis in determining phenotypic response, it has been largely overlooked (Carlborg and Haley, 2004).

Crop models can assist plant breeding by integrating physiological and biochemical understanding, agronomic practices, the environment, and genetic information. Shorter et al. (1991) and Cooper and Hammer (1996) proposed four main ways in which crop models can be used to assist genetic improvement:

(1) Environmental characterization for testing genotypes.
(2) Assessment of specific putative traits for designing improved plant types.
(3) Analysis of responses of probe genotypes for improved interpretation of multi-environment trials (METs).
(4) Optimizing combinations of genotype and management for target environments.

Implicit in uses 1 through 3 is that crop models can help our understanding of the physiological basis of genotype by environmental interactions, hence, increase the sensitivity of the analysis of variance of trial data. Boote et al. (2001) expanded crop model applications in plant breeding to the analysis of the physiological mechanisms involved in past genetic yield improvement, and to understand additive, epistatic and pleiotropic effects (effects observed when one gene affects more than one trait), of traits that affect yield. Recent applications of crop models proposed by Cooper (1999) were to help understand the differences in adaptation among genotypes, quantify the contribution of adaptation components that improve performance of the breeding program, and analyze the breeding process. This latter application used crop models in combination with quantitative genetic simulation models (e.g. QU–GENE, Podlich and Cooper, 1998) to predict phenotypic response to selection over multiple generations. A key component in this application was the availability of models that map genes-to-phenotypes for complex traits. An emerging application of crop models has been the integration of molecular marker information with crop models (Cooper et al., 2002; Boote et al., 2003; Chapman et al., 2003; White and Hoogenboom, 2003; Messina et al., 2006).

Impacts of crop models in plant improvement have not yet met expectations despite their potential (White, 1998; Boote et al., 2001). Some reasons that may hamper the use of models in plant breeding are: (1) the reliance on field trials to fit model coefficients and the lack of methods for their estimation (White, 1998), (2) inadequate links between genes and plant traits (White and Hoogenboom, 1996), and (3) inadequate representation of epistatic and pleiotropic effects (Messina et al., 2004). Plant breeding efforts could benefit from a new generation of crop models that include information on the genetic control of physiological and morphological traits. One such approach is to associate genes and genetic coefficients, which are paradoxically phenotypic in nature, in current crop models.

Alternative approaches seek to simulate gene networks at the molecular level and scale up these processes to predict crop phenotypes. The predictability and stability of such gene-based models across environments and across genetic backgrounds pose a major challenge.

This chapter illustrates three major applications of crop models to plant breeding. Section 2 illustrates a recent novel application of a maize crop model using principal component analysis (PCA) and GIS for characterizing environments and improving selection through better understanding of genotype by environment interactions (GEIs). Section 3 describes the use of crop models to integrate plant physiology and breeding. Two applications are presented describing the use of crop models to (a) understand past genetic improvement and (b) to estimate the value of putative traits. Section 4 discusses a current paradigm for gene-to-phenotype modeling, describes a gene-based model for soybean and an application to plant breeding, and reviews an approach to simulate breeding strategies. This section also introduces the use of global optimization to search for best combination of traits for a given environment.

2. Crop model use for environmental characterization

Multi-environmental trials (METs) are traditionally used to assess cultivar adaptation within a target population of environments. Variable environments and genotype by environment interactions hamper the development of better-adapted genotypes as the need for adequate sampling of the environmental space increases in order to obtain gains from selection. Crop models can be used for characterizing the target population of environments (TPEs) and their environmental challenges (variables in the environment that affect crop performance) (Chapman et al., 2002). Thus, they can assist the plant breeder to make informed decisions about the environments in which to conduct field trials, to evaluate the extent to which the trial environments are an unbiased sample of the TPE, to determine weighting factors to correct bias due to inadequate sampling, and to interpret field results in the context of the type of environment sampled by that particular trial. This ultimately can increase gains from selection of genotypes in breeding programs.

The need to characterize the environments used for plant breeding and variety characterization trials has been widely documented (e.g. Comstock, 1977; Cooper et al., 1993; Chapman et al., 2000; Hartkamp et al., 2000; Löffler et al., 2005). Attempts to characterize maize environments largely fall into three categories:

(1) Classifications based on climatic and soils data (e.g. Runge, 1968; Pollak and Corbett, 1993; Edmeades et al., 2000). While useful to describe environmental variables affecting crop productivity over long periods of time, these efforts did not attempt to identify what environmental variables were most important in influencing GEI and thus the genetic correlations for genotype performance among testing sites, a key factor in determining the efficiency and efficacy of a cultivar evaluation system.

(2) Classifications based on the statistical analysis of variety performance data. For example, Eberhardt and Russell (1966) characterized environments using an "environment index" computed as the mean yield of all the varieties grown at a

given location. While this approach is appealing for its simplicity and it has been widely used, it does not provide a measure of the environment that is independent of the variety performance measured at that environment. Since a wide array of environmental conditions may result in similar crop yields, unless these conditions are described, this methodology offers limited predictive value. Cooper et al. (1993) compared the relative merit of four strategies for classifying wheat (*Triticum aestivum* L.) environments and favored classifications based on the standardized and rank transformations. The value of these classifications for predicting cultivar performance is enhanced by knowledge of the underlying causes of the observed GEI, and of whether the classification adequately depicts long-term pattern.

(3) Classifications using crop models to integrate weather, soil and management information. Model outputs can be used to produce categorical variables that describe environments in terms of levels of stress that impact crop productivity (Chapman et al., 2000). Löffler et al. (2005) used the CERES-Maize model (Ritchie, 1986) and GIS technology to classify each US corn-belt township for 1952–2002, and for each of 266 trial locations in 2000–2002. The crop modeling approach combined with GIS facilitated the generation of maps that depict different views of the environmental conditions for crop production allowing a first characterization of the environmental challenge. The Figure 1a shows the spatial pattern of simulated water stress at the stages of crop development for 2002. A similar map could be generated for any period of time that characterizes the TPE. Knowledge of drought patterns at the TPE level is essential in breeding for crop tolerance to drought (Edmeades et al., 2000).

Löffler et al. (2005) also used crop model outputs to develop a classification methodology to enhance the sensitivity of the analysis of variance to detect yield differences among hybrids. Model outputs included simulated water stress, mean temperatures and photoperiod at three developmental stages, vegetative (from emergence to V7), flowering (from V7 to R2) and grain filling (R2 to physiological maturity). Significant hybrid by environment class interaction variance was identified for six main environmental classes (temperate, temperate humid, temperate dry, high latitude, European Corn Borer and subtropical). Figure 1b shows the spatial distribution of these classes in 2002.

However, the relative frequencies of these environmental classes varied greatly from year to year raising the question of how well a MET samples the TPE for a given year. The aggregated view of classification at the township level over a period of years (51 in this case) provided a description of the TPE. Therefore, this long-term historical frequency distribution was used as a reference to compare the distribution of environments in a given year and in any given MET. Figure 2 shows an example for 2002. In this year, temperate and temperate dry environments occurred with higher frequencies than in the long-term, thus these environments were overrepresented in the MET ($\chi^2 = 36.64$) introducing a bias in the selection process.

The stratification of locations by environmental class enabled interpretations of observed variety responses to growing conditions. For example, the GGE biplot (Cooper and DeLacy, 1994) can be used to infer hybrid adaptation to the specific environmental classes. This technique allows displaying both genotypes and environments simultaneously, thus helping identify systematic patterns in G, E and GEI. Singular value

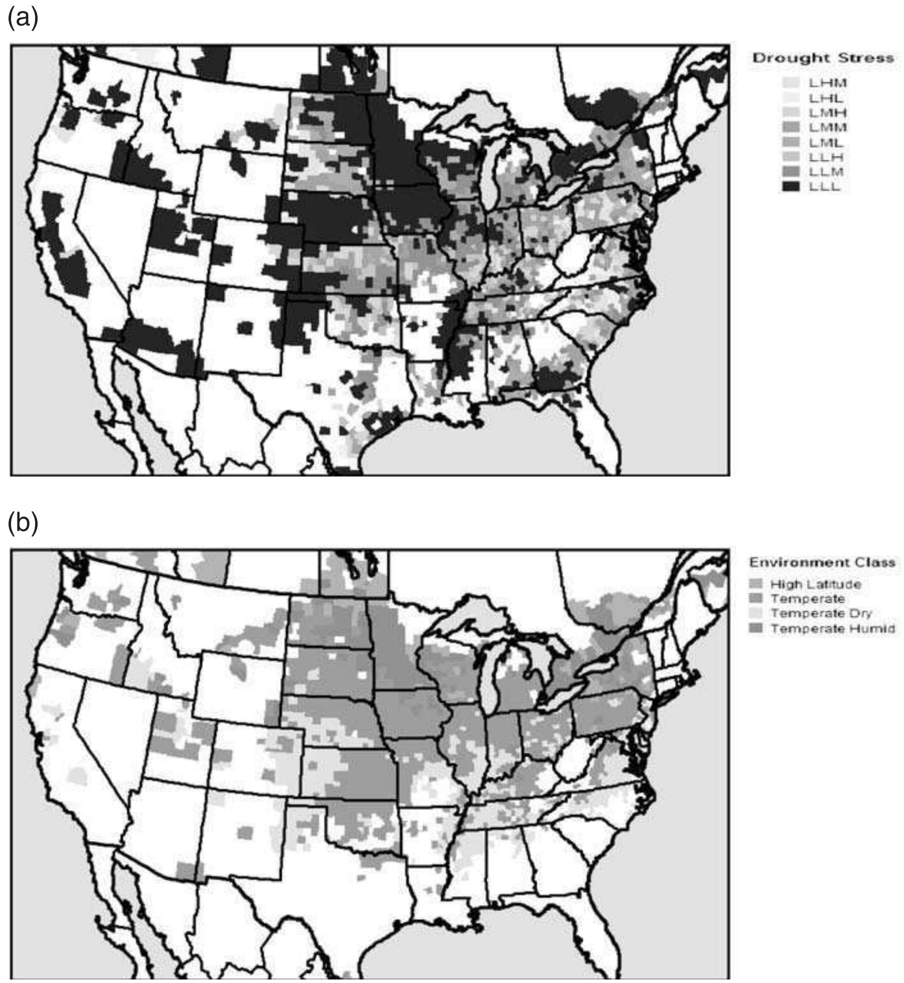


Figure 1. Geographic distribution of five abiotic environment classes (a) and drought stress (b) for maize production in the United States in 2002. Water stress levels: L – Low, M – Medium, H – High. Developmental stages are represented by the order of letters. First: V1–V7, second: V7–R2, third: R2–R5.

decomposition was used to find the best representation of the two-way data matrix **X**, containing yield information for G genotypes in E environments, in a low-dimensional space. Figure 3 presents the results of applying this technique to a matrix of 18 hybrids and 266 trial locations. Yield data were centered on environment by removing the environment mean from each observation. The hybrid H9 outperformed all other hybrids included in this study in the temperate, temperate dry and European Corn Borer (ECB) environments, which are the most frequent environments found in the central Corn Belt. H9 is an example of a hybrid with broad adaptation; the projection of the vector with coordinates (0, H9) onto any of these environments was greater than for any other hybrid. GEIs are

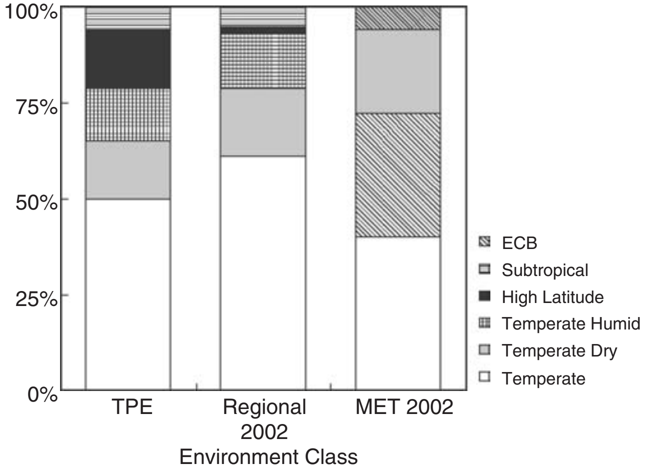


Figure 2. Frequency in percent of total hectares of Corn Relative Maturity (CRM) 110 maize environments in the USA in 1952–2002, compared to frequencies in 2002, both at the regional (TPE) and site-specific (MET) levels. Adapted from Löffler et al. (2005). H_0: TPE = MET. $\chi^2_{\mathrm{MET}} = 36.64^*$.

evident in this graph, however. When comparing H9 with five other hybrids (H1–H5) in high latitude environments (project the vector (0, H9) onto the high latitude vector), the clear yield advantage of H9 vanished and its yield was comparable to any of the hybrids from H1 to H5.

3. Crop model use for analysis of past genetic improvement

Increased understanding of the underlying physiological causes of past genetic improvement can increase the efficiency of plant breeding programs by helping define selection criteria and breeding objectives. Crop physiological changes associated with cycles of plant breeding were described for maize (Echarte et al., 2000), soybean (Boote et al., 2001; Kumudini et al., 2001, 2002), sunflower (Lopez Pereira et al., 1999a,b, 2000) and peanut (Duncan et al., 1978). These studies illustrate that crop models can assist in the analysis of past genetic improvement in formal and informal ways by providing a theoretical, physiological framework. Echarte et al. (2000) measured increased kernel number for a given plant growth rate in newly released maize hybrids relative to old ones. Their analysis is analogous to comparing the genetic coefficient maximum kernel number in the CERES-Maize crop model. Another example is shown by Lopez-Pereira et al. (1999a,b, 2000) in sunflower. A series of open pollinated cultivars and hybrids released to the market over six decades were compared for a series of physiological traits. Traits that regulate sunflower phenology were genetic coefficients in controlling plant development in OILCROPSUN (Villalobos et al., 1996). Using this framework they identified a negative trend in total crop cycle duration associated with a reduction in the thermal time to flowering, but no reduction for seed fill duration.

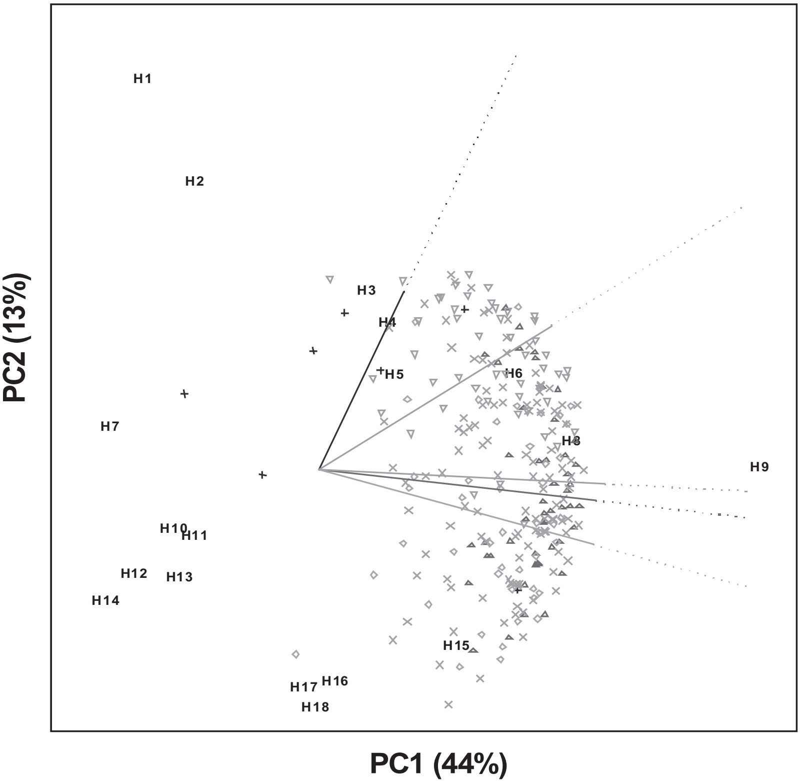


Figure 3. Environment-standardized GGE biplot of grain yield of 18 maize hybrids (H1–H18) grown in 266 environments over three years, stratified by environment class: High Latitude (+), Temperate Humid (▽), Temperate (×), Temperate Dry (◊) and European Corn Borer (ECB) (Δ). Percent of the total GGE variation explained by the main two principal components in parentheses.

A realistic use of crop models is to generate hypotheses and investigate causal relationship for yield improvement. Boote et al. (2001) used CROPGRO-Soybean to help understand the physiological causes of past genetic improvement in soybean. This inverse engineering use of crop models requires detailed data on development and growth measured in cultivars released at different times in a breeding program. Boote et al. (2001) collected detailed growth analyses data for one old cultivar (Williams 82) and two modern cultivars (Kruger 2828 and Stine 3660) in the absence of biotic stresses. Soil water measurements were used to verify the proper simulation of the water balance. Average dry matter production was used to estimate a site fertility factor that helps avoid confounding site and cultivar trait merits. Genetic coefficients were estimated for each cultivar following the procedure outlined by Boote (1999) and Boote et al. (2001). This procedure led to an adequate simulation of crop and seed growth (Fig. 4) for all cultivars, thus helping to identify the causes of yield gains. Simulated yield for Kruger 2828 and Stine 3660 were 15–19% higher than Williams 82. Simulated and observed traits leading to yield improvement of modern cultivars included earlier pod set, earlier seed set, longer time from seed to maturity and higher harvest index. Increased yield due to differences in harvest index, biomass accumulation and leaf area duration around first seed were also reported (Kumudini et al., 2001). However, the analyses of the genetic coefficients helped dissect these traits and identify the physiological causes of yield gains. Modern cultivars had 30% faster pod addition, 9% longer SD-PM, 8% higher SFDUR, 9% higher leaf photosynthesis, and 10% lower leaf nitrogen remobilization. The increased dry matter accumulation at first seed, harvest index and leaf area duration (Boote et al., 2001; Kumudini et al., 2001) were modeled outcomes associated with changes in genetic coefficients.

Another application of crop models to understand past genetic improvement was illustrated by Duncan et al. (1978), using a crop modeling analyses of peanut cultivars released over a period of 40 years of genetic improvement. They documented that yield improvement was associated with increased intensity of partitioning to pods and more rapid transition to a full pod load. The oldest peanut cultivar from the 1940s continued vegetative growth all the way to maturity and had only 40% partitioning intensity to pods, as compared to 92–98% partitioning intensity for two recently released high-yielding cultivars. Their perspective toward the future was that peanut improvement had nearly reached the limit of improvement toward this particular trait. Soybean cultivars, by comparison, have nearly 100% partitioning intensity during seed fill.

These types of analysis of traits using crop models can point the way for future genetic improvement. Crop models not only provide better insight of past genetic improvement, but also can provide a view on whether further changes in particular traits associated

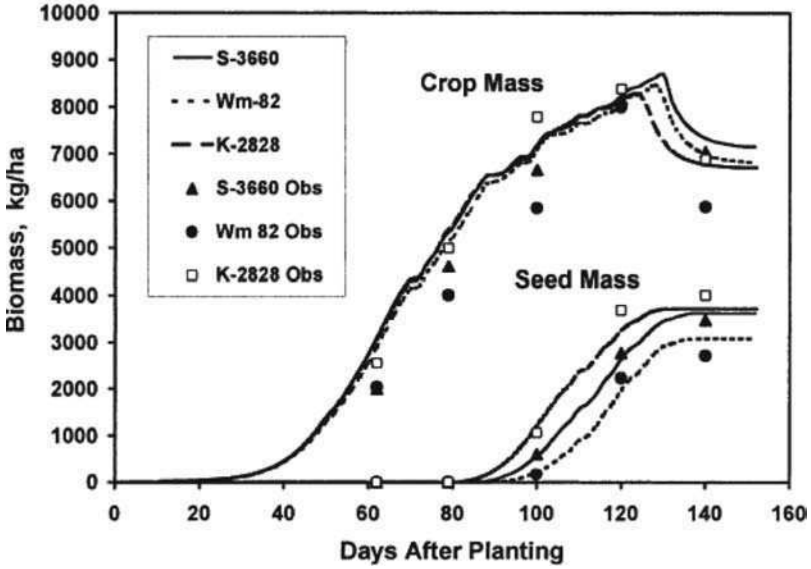


Figure 4. Growth dynamics for seed and total crop mass of new soybean cultivars Kruger 2828 and Stine 3660 compared to old cultivar Williams 82 at Lewis, Iowa in 1997. Simulated values are shown in lines. (Reprinted from Boote et al., 2001. Agricultural Systems 70, 395.)

with yield gains are likely to continue. Finally, one major value of using crop models to understand past genetic improvement is the change in paradigm that emphasizes thinking in terms of processes rather than states.

4. Crop model use to design new cultivars

One application of models in plant breeding is through the design of ideotypes for a target population of environments (Boote and Tollenaar, 1994; Kropff et al., 1995; Hammer et al., 1996). Plant ideotypes are those with the combination of plant traits that maximize yield in a target environment. Crop models have been used to identify ideotypes for soybean (Boote and Tollenaar, 1994; Boote et al., 2001), maize (Boote and Tollenaar, 1994; Boote et al., 2001; Sinclair and Muchow, 2001), wheat (Aggarwal et al., 1997), peanut (Boote and Jones, 1986) and rice (Kropff et al., 1995) among other crops. Individual or combinations of model parameters within known genotypic ranges were varied to study their impact on yield. These simulation studies identified traits contributing to increases in potential yield (e.g. leaf photosynthesis, stay green, higher synchronism of fruit addition, longer seed fill duration) and yield under drought (e.g. deeper root systems, osmotic adjustment). Some of these findings are being corroborated by experimental data generated in managed environments and through the comparison of hybrids and cultivars released throughout the decades, as described in the previous section for soybean.

At the same time, simulation studies suggest the need for varying multiple traits to attain significant, albeit modest, increases in yield. In reality, breeder's direct selection for higher yield would naturally be due to multiple traits. For example, to simulate a genetic gain of 10% in soybean, Boote et al. (2001) had to simulate genetic modifications in growth habit, seed fill duration and leaf maximum photosynthesis. The need for multiple traits led researchers to seek alternative methods to design ideotypes. Early work by Aggarwal et al. (1997) used Monte Carlo simulation to generate a relatively large number of cultivar-specific parameter combinations in rice in order to search for those combinations that maximized potential yield in irrigated tropical environments. Aggarwal et al. (1997) generated 500 hypothetical genotypes by resampling combinations of parameters that varied 20% around the value estimated for a reference cultivar IR72. In this study, as in other similar research (Paruelo and Sala, 1993; Kropff et al., 1995; Hammer et al., 1996; Sinclair and Muchow, 2001), parameter controlling crop traits were considered stochastic and independent. This research found that no trait individually or in combination offered more than a 5% increase in yield in the simulated breeding program. Similarly, Boote et al. (2003) showed more response to certain traits, when placed in good management (high population and narrow row spacing) and high carbon dioxide environment, than in poor management or environment. Ideotype design must hence include not only information about the target population of environments, but also about the management of the crop to account for the genotype × management interactions. Hammer et al. (1996) acknowledged these problems and proposed a method for ideotype design that linked a sunflower model with a simplex algorithm to optimize crop traits and its management for a given environment. This numerical optimization method searched effectively the genetic coefficient space. In this particular case, the response surface was smooth enough for local search algorithms to provide solutions close to expected global optima. However, these conditions may not be met (Royce et al., 2001) due to the complex interactions between physiological processes, crop management and the environment.

There are often non-linear correlations between genetic traits, or considered in terms of physiology, there are connections between morphological traits and process traits. For example, a thicker leaf causes higher leaf photosynthesis, but lower leaf area. Ignoring correlations between genetic coefficients and designing ideotypes by selecting combinations of physiological model parameters under the assumption of independence can lead to (a) infeasible combinations of traits, and (b) local maximum yield. Little attention has been paid to this important problem, however. Boote et al. (2001) illustrated the implications of ignoring correlations between genetic coefficients (pleiotropic effects). Early research showed a strong correlation between maximum photosynthesis and leaf-specific leaf weight (SLW) (Dornhoff and Shibles, 1970). Boote et al. (1994, 2001, 2003) conducted a sensitivity analyses in which grain yield variations were recorded for variations in the genetic coefficients LFMAX and SLAVAR when these coefficients were varied independently and when considering the appropriate coupling between parameters. In the presence of coupling between increased SLW and leaf photosynthesis, the yield response to increasing leaf photosynthesis (*via* SLW) rapidly reaches an asymptote (Fig. 5). This response is caused by the negative feedback between SLW and LAI, and therefore light interception. In contrast, in the absence of coupling, seed yield continues to increase with increasing maximum leaf photosynthesis. However, trait "interaction" with environment and management (narrow row spacing, high sowing density, high fertility or elevated carbon dioxide), could overcome the LAI-light interception limitation, and allow response to increased photosynthesis even if linked to SLW, as suggested by Boote et al. (2003).

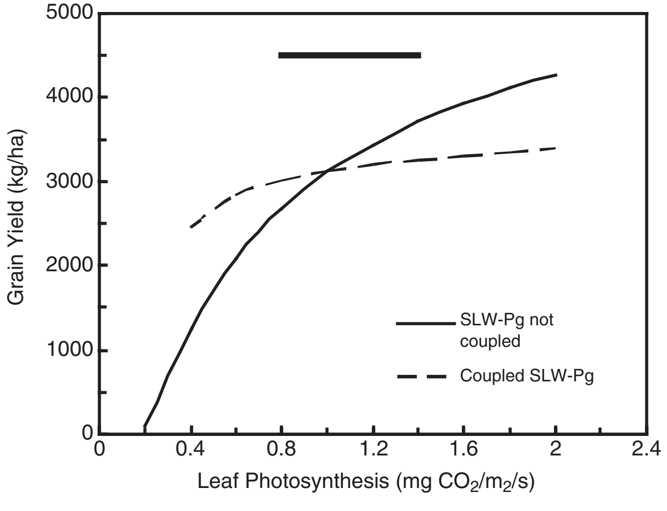


Figure 5. Simulated soybean yield as a function of variation in leaf P_{max}, attributed to inherent rate (no change in SLW), or attributed (coupled) only to SLW, over 17 rainfed seasons at Ames, IA. Horizontal bar represents feasible genetic range for P_{max} about the mean of reported literature values. (Adapted from Boote et al., 2003. Agronomy Journal 95, 32–51.)

5. Crop models that incorporate gene information

A major challenge in plant breeding is making inferences about how a given genetic material, a commercial hybrid, an F_1 topcross or an ideotype we are interested in creating, will perform over the TPE. Crop models could allow us to make informed quantitative predictions over yield trials and TPE based on local experimental results and sensitivity analyses (Fig. 6a). Achieving this capability depends on a number of factors, including how well physiological mechanisms controlling the phenotypic variation are represented in the model and the accuracy with which model parameters that capture the genetic variation are estimated. Considering a crop model developed using a sound physiological basis, parameter estimation becomes a limiting factor for model application to plant breeding. Most genetic coefficients are seldom measured directly. Expensive experimentation may be required to measure some genetic coefficients, and that makes this practice difficult to use for large numbers of genotypes. One way to make crop models more useful for plant breeding applications is to include more basic information on the genetic makeup of specific cultivars.

Advances in agricultural genomics can facilitate the inclusion of genetic information in crop models. New technologies developed in agricultural genomics allow us to perturb one of many functions in an organism (Valenzuela et al., 2003), monitor whole-genome gene expression (Brown and Botstein, 1999), protein concentrations (Ghaemmaghami et al., 2003), protein modification dynamics (Raghothama and Pandey, 2003) and large numbers of metabolite concentrations (Weckwerth, 2003). Gene mapping technologies have also

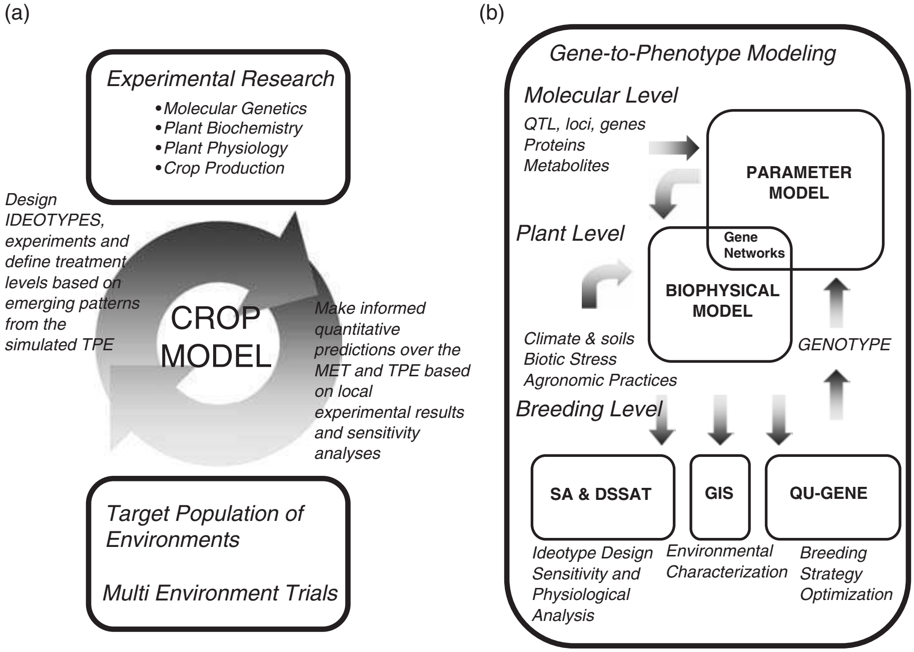

(a)
Experimental Research
•Molecular Genetics
•Plant Biochemistry
•Plant Physiology
•Crop Production
Design IDEOTYPES, experiments and define treatment levels based on emerging patterns from the simulated TPE
CROP MODEL
Make informed quantitative predictions over the MET and TPE based on local experimental results and sensitivity analyses
Target Population of Environments
Multi Environment Trials
(b)
Gene-to-Phenotype Modeling
Molecular Level
QTL, loci, genes
Proteins
Metabolites
PARAMETER MODEL
Plant Level
Gene Networks
BIOPHYSICAL MODEL
Climate & soils
Biotic Stress
Agronomic Practices
GENOTYPE
Breeding Level
SA & DSSAT
GIS
QU-GENE
Ideotype Design Sensitivity and Physiological Analysis
Environmental Characterization
Breeding Strategy Optimization

Figure 6. Gene-based modeling applications in plant breeding. (a) Crop model as a tool to integrate knowledge and make quantitative predictions of performance. (b) Gene-to-phenotype modeling.

evolved allowing us to have more precision in identifying genomic locations associated with plant traits (Syvänen, 2001; Darvasi and Pisanté-Shalom, 2002). It is now possible to make use of such information to help parameterize crop models and make them more useful for plant breeding purposes.

Figure 6b shows how this approach can operate. At the molecular level, genes and QTL-associated markers are identified for each cultivar through laboratory techniques. Relationships between gene combinations and crop model genetic coefficients are developed through field research (see below). These relationships are the core of the parameter model, which is used to estimate coefficients in the crop model. Some authors proposed terms such as meta-mechanisms (Hammer et al., 2002, 2004; Tardieu, 2003) and gene-based modeling (White and Hoogenboom, 1996) to refer to the parameter model and to acknowledge the empirical nature of the approach, yet highlight the causal link between the genetic controls and the physiological process. This approach follows the philosophy of modeling hormone effects without simulating hormone action (de Wit and Penning de Vries, 1984) and iterative model building. Then, these cultivars can be simulated for different target environments and management practices to predict performance, optimize cultivar choice, design ideotypes and simulate breeding strategies.

This procedure allows improving the links between genes and plant traits, thus better representing epistatic and pleiotropic effects, yet maintaining phenotypic predictability. By linking genetic information to processes mediating physiological responses to the environment we can simulate explicit genotype by environment interactions. White and Hoogenboom (1996) developed the first gene-based crop model, Genegro; a process-oriented model that incorporated effects of seven genes affecting phenology, growth habit and seed size of common bean (*Phaseolus vulgaris* L.). Genetic coefficients in Genegro were based on the allelic configuration of a set of loci, and a set of linear functions. Genegro accurately predicted dry bean phenological development but poorly explained yield variations between sites (Hoogenboom et al., 1997). Recent improvements of Genegro included the simulation of the effects of temperature on photoperiod sensitivity regulated by the gene *Tip*, and a new function to predict seed weight (Hoogenboom and White, 2003; Hoogenboom et al., 2004). Similar modeling approaches were used to incorporate quantitative trait loci (QTL) effects on leaf elongation rate in maize (*Zea mays* L.) (Reymond et al., 2003), plant height, pre-flowering duration, carbon partitioning to spike, spike number and radiation use efficiency in barley (*Hordeum vulgare* L.) (Yin et al., 2003) (Table 1).

Gene-based crop model development is an iterative process with a core three-step procedure (Fig. 7). The first step consists in measuring (e.g. Reymond et al., 2003) traits for known genotypes grown in experiments or yield trials. The second step is estimating the genetic coefficients of those data (e.g. Hoogenboom et al., 2004). The third step is finding a parameter model that estimates the genetic coefficients using marker, and other genetic information across all genotypes. When the parameter model uses molecular markers as inputs and mapping algorithms (Wang et al., 2003) to predict the genetic coefficients (e.g. composite interval mapping), the genetic coefficients are mapped in the genome and QTL are estimated as part of the same procedure. These relationships should be evaluated using independent data. This is particularly relevant for QTL-based models. Most QTL studies use bi-parental populations, therefore, genetic context dependencies are unknown. QTL models must be validated across genetic backgrounds as a condition for model application. Here we summarize the work done by Messina et al. (2006) and Messina (2003) for soybean to illustrate this approach.

Table 1. Gene-based approaches to simulate crop growth and development.

Reference	Crop	Genetic material	Genetic information	Observations
Tardieu (2003), Reymond et al. (2003)	Corn	RIL	QTL	Demonstrated predictability for new QTL combinations but within RIL population
White and Hoogenboom (1996), Hoogenboom et al. (2004)	Bean	Cultivars	Allelic information at given loci	Demonstrated predictability across cultivars
Welch et al. (2003)	*A. thaliana*	Mutants	Gene mutation	Not transferable to other crops
Yin et al. (2003)	Barley	RIL	QTL	Valid within RIL population
Messina (2003), Messina et al. (2004), Stewart et al. (2003)	Soybean	NIL	Allelic information at selected loci	Demonstrated predictability in independent genetic background

RIL – recombinant inbred line; QTL – quantitative trait loci; NIL – near-isogenic line.

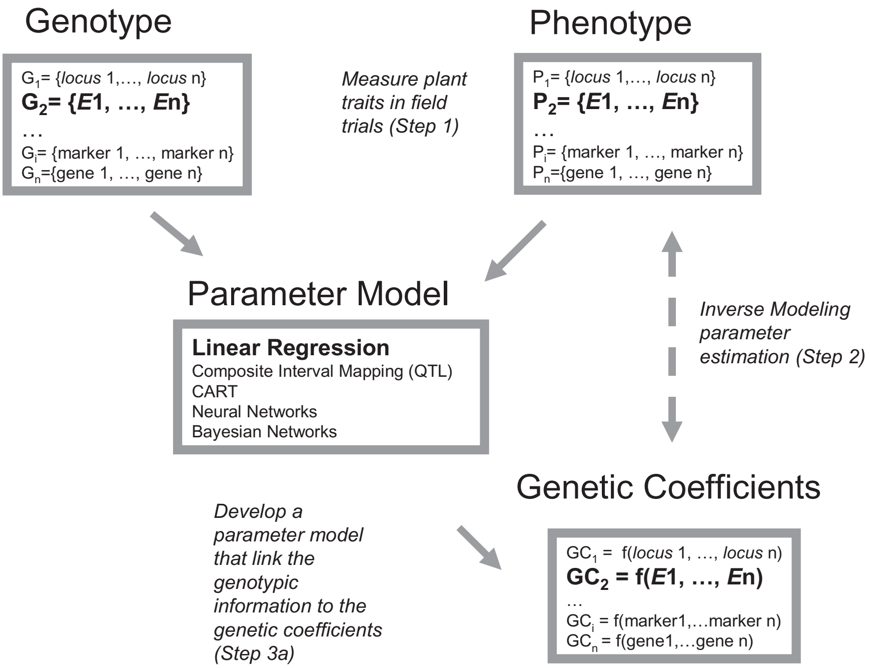

Genotype
G_1= {*locus* 1,…, *locus* n}
G_2= {*E*1, …, *E*n}
…
G_i= {marker 1, …, marker n}
G_n={gene 1, …, gene n}
Measure plant traits in field trials (Step 1)
Phenotype
P_1= {*locus* 1,…, *locus* n}
P_2= {*E*1, …, *E*n}
…
P_i= {marker 1, …, marker n}
P_n={gene 1, …, gene n}
Parameter Model
Linear Regression
Composite Interval Mapping (QTL)
CART
Neural Networks
Bayesian Networks
Inverse Modeling parameter estimation (Step 2)
Genetic Coefficients
Develop a parameter model that link the genotypic information to the genetic coefficients (Step 3a)
GC_1 = f(*locus* 1, …, *locus* n)
GC_2 = f(*E*1, …, *E*n)
…
GC_i = f(marker1,…marker n)
GC_n = f(gene1,…gene n)

Figure 7. Methodological approaches for gene-based modeling. In bold is indicated the specific methodologies used by Messina et al. (2002) and Messina (2003).

5.1. Experimental work and parameter estimation

Gene-based models can be developed using a number of genetic resources, including near isogenic lines (NILs), recombinant inbred lines (RILs), mutants, transgenics or simply plant populations for which one or many loci or pedigrees are known (Fig. 7). The important aspect of the selection of the plant material to use is the availability of information about their allelic makeup at loci with known effects on physiological processes. This can involve discovering new loci or assigning new functions to known loci. Near-isogenic lines are created by introgression of a locus or group of loci from a donor parent into a recurrent parent of interest, and subsequent backcrossing to a recurrent parent. This procedure leads to a set of plants with almost identical genetic background except for the loci of interest. NILs are particularly suitable to identify and test for epistatic and pleiotropic effects. A family of recombinant inbred lines (RILs) is created from an F_2 obtained between two parental inbred lines with contrasting genetic backgrounds. The F_2 progeny is then advanced to latter generations by a program of single seed descent. During this process, recombination is fixed and homozygosity is attained at practically every locus. Thus, each RIL in the family represents a unique combination of the parental alleles, and all members of a line are genetically identical to each other. Recombinant inbred lines are created by an initial crossing between two parental lines with contrasting genetic backgrounds followed by several cycles of self-pollination. The resulting offspring carry several homozygous combinations of the parental alleles at each locus. Recombinant inbred lines are commonly used for QTL studies (e.g. Reymond et al., 2003; Yin et al., 2003), but large populations and precise measurements are required for obtaining good estimates of the QTL effects. Recombinant inbred lines are most suitable for identifying additive effects and have less power to identify epistatic interactions.

Messina et al. (2002) used the physiological framework in CROPGRO-Soybean (Boote et al., 1998) to develop a gene-based approach to simulate growth and development of soybean (herein Genegro-Soybean). Field experiments were conducted using NILs with different allelic combinations for a set of six *E* loci (*E*1, *E*2, *E*3, *E*4, *E*5, *E*7) (step 1, Fig. 7). These loci were known to regulate time to flowering and maturity responses to photoperiod (Cober et al., 1996, 2001). The developmental phenotypes of the NILs were recorded under two contrasting daylength at the same site; this was accomplished with two sowing dates. Plant development was measured at several stages, emergence, first flower, first small pod, last small pod, first seed and physiological maturity. Genetic coefficients CSDL (critical short daylength), PPSEN (photoperiod sensitivity), EMFL (photothermal time to flowering), FL-SD (photothermal time to first seed), FL-PM (photothermal time from flowering to physiological maturity), VI-JU (photothermal time duration of the juvenile phase), and R1PRO (post-flowering reduction in CSDL) were estimated for each NIL using inverse modeling, as described by Mavromatis et al. (2001) (step 2 in Fig. 7). (See Chapter 4 for more information on parameter estimation.)

5.2. Parameter models from known gene combinations in genotypes

There are several statistical methods to model associations between genotypes and genetic coefficients. In Genegro-Soybean the parameter models are linear models estimated using regression. For example, in the case of a single gene with two alleles,

$$GC_i = a + b \cdot L$$

where *GC* is the genetic coefficient controlling the physiological process *i*. The variable *L* indicates which allele of the given loci gene is present. This variable can take a value of 1 or 0 for dominant and recessive alleles, respectively. Different degrees of dominance can be modeled by letting *L* vary continuously between 0 and 1, or between –1 and 1 as used in Falconer and MacKay (1996). Parameters *a* and *b* are estimated through linear regression. An example of this model in Genegro-Soybean is represented by the function describing the effects of *E*3 on R1PRO, the reduction in photoperiod sensitivity after flowering:

$$\text{R1PRO} = 0.24 + 0.13E3 \; (R^2 = 0.69)$$

Genes that affect more than one trait or physiological process, are said to have pleiotropic effects (MacKay, 2001). Additional equations are used to model pleiotropic effects, for example,

$$GC_i = a + b \cdot L$$

$$GC_{i+1} = c + d \cdot L$$

where GC_{i+1} is the genetic coefficient for a different trait, *c* and *d* are regression parameters and *L* is the same allele for both equations. Pleiotropy is common in physiological traits expressed at the crop level. In soybean, the gene *dt* controls the type of growth habit the soybean plant is going to have – determinate *vs*. indeterminate. The time between the appearance of the first flower and that of the last node in the main stem (FL-VS) or the end of leaf area expansion (FL-LF) is shorter in determinate $(Dt/\frac{1}{\mathrm{H}})$ than in indeterminate (*dt/dt*) soybeans. Boote et al. (2001) estimated that on average, the *Dt* allele shortens FL-VS from 26 to 6 physiological days, and FL-LF from 26 to 18 physiological days. It also increases the rate of pod addition (1/PODUR) by about 33%.

The vast majority of agronomic traits are complex and polygenic (Daniell and Dhingra, 2002; Stuber et al., 2003). In their simplest form, the effects of multiple genes or loci on a trait can be represented by additive effects,

$$GC_{i+2} = a + b \cdot L + c \cdot L_1$$

For example, in Genegro-Soybean, the photothermal time between the crop emergence and flowering is regulated by *E*1 and *E*3,

$$\text{EM-FL} = 20.77 + 2.1E1 + 1.8E3 \; (R^2 = 0.78)$$

Interactions among loci or between genes and environmental factors make a substantial contribution to variation in complex traits (Carlborg and Haley, 2004). When one or more genes influence the effects of other gene(s) we refer to this interaction as epistasis (MacKay, 2001). In linear models, epistasis can be represented as the product of interacting genes,

$$GC_{i+3} = a + b \cdot L + c \cdot L_1 + d \cdot L \cdot L_1$$

The genetic coefficient PPSEN in Genegro-Soybean is an example of epistasis between *E* loci. The *E*1 locus interacts with all other *E* loci as shown by the term $E1 \cdot NLOCI$ in the Genegro-Soybean model equation to estimate the genetic coefficient for photoperiod sensitivity (PPSEN):

$$\mathrm{PPSEN} = 0.11 + 0.063\,NLOCI + 0.58E1 - 0.13E1 \cdot NLOCI \ (R^2 = 0.70)$$

Epistasis becomes an important issue when modeling complex traits using QTL analysis, as the opportunity for identifying epistatic effects not only increases but also does the uncertainty associated with them. Methodological requirements often cause researchers to neglect epistasis in complex trait studies (MacKay, 2001; Carlborg and Haley, 2004). The power to detect epistasis between QTL in mapping populations is low for a number of reasons. The number of interactions increases with the number of genes controlling a trait. Often, after adjusting the significance threshold for the multiple statistical tests involved in searching for epistatic interactions, only extremely strong interactions remain significant (MacKay, 2001). Several statistical models were developed to address these problems in QTL analysis (e.g. Lander and Schork, 1994; Doerge et al., 1997; Carlborg and Haley, 2004). From an experimental point of view, this implies phenotyping an increasing number of individuals carrying many of all possible combinations of genes controlling the trait in different genetic backgrounds so that spurious correlations are minimized; even large mapping populations contain few individuals in the rarer two-locus genotype classes.

5.3. *From loci-based model to a molecular marker-based model*

Genegro-Soybean was developed using information about Mendelian loci. Cultivars carrying the gene *Fin* or *dt* can be easily recognized by their phenotype. The tagging of a locus controlling a trait of agricultural importance was first described by Sax (1923) who reported linkage between a phenotypic marker controlling pigmentation in seed coat of common bean and a locus controlling seed size. In fact, extensive linkage maps were constructed with phenotypic markers during most of the twentieth century. However, practical applications of these maps are very limited as almost all loci were identified by recessive mutant alleles. Another limitation of phenotypic markers is that they are not always reliable predictors of the genotype. For instance, different soybean cultivars can display the same phenotypic photoperiodic response with different allelic combinations at *E* loci. Genotypic characterization of these cultivars would only be possible by genetic analysis. Therefore, a reformulation of the model is needed to use information about the genotype that is relatively easy to obtain (step 3, Fig. 7). Molecular markers provide the means for cultivar genotyping and molecular maps allow us to select those markers closely linked to the gene of interest. Then, the parameter models will take the form,

$$GC_{i+3} = a + b \cdot MM_1 + c \cdot MM_2 + d \cdot MM_1 \cdot MM_2$$

where *MM* are molecular markers. Their values will depend upon the effect of the loci with which they are associated and the mode of gene action. For example, *MM* associated with $E1$ will take a value of 1, while the *MM* allele corresponding to $e1$ will take a value of 0. In Genegro-Soybean, Messina (2003) estimated EM-FL by replacing the alleles at $E1$ and $E3$ by the alleles for the microsatellites Satt357 and Satt229,

$$\text{EM-FL} = 20.77 + 2.1\,\text{Satt357} + 1.8\,\text{Satt229}$$

This approach was used to genotype commercial cultivars and predict growth and development under different conditions. To test Genegro-Soybean, cultivars were genotyped using simple sequence repeats (Akkaya et al., 1992; Ellegren, 2004) (microsatellites) at marker loci closely linked to *E* loci and QTL regulating soybean development. Genegro-Soybean predicted 75% of the variance in the time to maturity and 54% of the yield variance in variety trials conducted in Illinois (Messina, 2003; Messina et al., 2004, 2006). This result shows that gene-based approaches can reduce expensive and time-consuming experimentation for model parameterization. It also demonstrates that gene-based approaches have potential to design cultivars *in silico*, provided there is a careful validation of the model across genetic backgrounds.

Various molecular marker types include allozymes (Tanksley and Orton, 1983), RFLP (Restriction Fragment Length Polymorphism) (Botstein et al., 1980), RAPD (Random Amplification of Polymorphic DNA) (Williams et al., 1990), SSR (Simple Sequence Repeat), AFLP (Amplified Fragment Length Polymorphisms) (Vos et al., 1995) and SNP (Single Nucleotide Polymorphism) (Syvänen, 2001). The selection of molecular markers for use in gene-based models can restrict the portability of the model to other cultivars and genetic backgrounds for a species. Markers sensitive to ploidy level should be avoided in polyploidy species (e.g. wheat) or ancient polyploids (e.g. soybean) as the same marker cannot differentiate between members in a gene family. Codominant markers are necessary only in species of interest, as on hybrids (e.g. Maize). Locus specificity dictates the portability of a gene-based model. For example, a gene-based model that uses RAPD or AFLP markers as variables in the parameter models has a limited scope and domain of application. Inferences can only be made within the plant population used for developing the model (e.g. Yin et al., 2003).

6. Use of gene-based crop models in plant breeding applications

Availability of gene-based models can improve the efficiency of breeding programs. Plant breeders may not need to rely entirely on extensive field trials to fit model parameters for each cultivar, a major limitation for crop model application in plant breeding (White, 1998). But they can concentrate their efforts in conducting directed experimentation to develop the parameter model that adequately accounts for epistatic and pleiotropic effects. Gene-based crop models have the potential to predict the performance of a cultivar over the multiple environment trials and TPE (Fig. 6). In cases where the cultivar is already available, crop models can help identify those environments where it is best adapted. It can also be a synthetic cultivar generated through the combination of different alleles at a given set of loci. Synthetic cultivars could be created and evaluated through computer simulation for many possible or target environments. Pattern analyses (Cooper and DeLacy, 1994) can help identify positive GEI interactions and thus help optimize the structure of testing sites. Gene-based models can predict phenotypes for a given genotype generated at each stage of a simulated breeding program. In summary, crop models can help us maximize the efficiency of the breeding strategies through simulation and optimization (Chapman et al., 2003).

6.1. Using gene-based crop models for ideotype design

Gene-based crop models can account for some of the epistatic and pleitropic effects described above. Messina (2003) compared soybean ideotype design using the Genegro-Soybean model with that obtained using the original CROPGRO-Soybean model (Boote et al., 1998). He used a robust optimization method that has characteristics somewhat analogous to the plant breeding process (simulated annealing (SA)). This algorithm was developed for solving large complex functions in combinatorial optimization problems (Kirkpatrick et al., 1983). It evaluates many combinations of parameters and compares each new combination of parameters with all combinations in the set. One analogy with the plant breeding process is that many genotypes are retained in a breeding program and new breeding lines are compared with this set in target environments. The objective function in plant breeding is typically maximum yield over a number of target environments. In the simulated annealing method, optimization involves selecting a new combination of parameters that lead to a higher or more stable yield at each step in the algorithm. This new combination may be retained, depending on how many combinations are being kept in the set. At first, many combinations are retained in the search process, but as the value of the objective function converges to its highest value, the number of combinations retained is decreased. Plant breeding uses a selection process that searches for genotypes that produces the highest yield in a similar manner.

Simulated annealing allows search of the parameter space in continuous and discrete domains and is robust to discontinuities in the objective function (Goffe et al., 1994). It includes checks for global optima and bounds to restrict the search to a subset of the parameter space. In plant breeding, one seeks to maximize the objective function,

$$f = {}_{\max} \sum_{i} w_i E(Y)_i$$

where, w_i is the weight given to a certain environment class i, and $E(Y)_i$ is the expected yield in that environment. In the simplest case, f is a simple average across environments in the TPE. The SA algorithm starts by estimating the value for the objective function (e.g. average simulated crop yields over the TPE) at a given initial combination of parameters $\mathbf{X}$, an n-dimensional vector; in the application to plant breeding $\mathbf{X}$ is a vector of genetic coefficients of genes. A second evaluation f' is made at $\mathbf{X}'$ by varying the ith element,

$$x_i' = x_i + r v_i \tag{1}$$

where r is a uniformly distributed random number and v_i is the step length for parameter x_i. In maximization problems, if f' is greater than f, then $\mathbf{X}'$ replaces $\mathbf{X}$, and the algorithm moves uphill. If this combination of parameters produces the largest value of f, then both **X** and f are recorded as the best current value of the optimum. When f' is lower or equal to f, the Metropolis criterion (Eq. 2) is used to decide acceptance of **X**. The Metropolis criterion is based on a simplified Boltzman probability distribution,

$$p = \frac{\exp\left(f' - f\right)}{\Phi} \tag{2}$$

where probability p is compared with a uniformly distributed random number p_{r}. If p is greater than p_{r} then $\mathbf{X}'$ is accepted and the algorithm could temporarily move downhill. Both, the difference between function values and Φ, affects the probability of accepting downhill movements. At the beginning the user defines the parameter Φ high enough such that there is a wide sampling of the function. As the optimization progresses, Φ gradually decreases based on the function,

$$\Phi' = r_{\Phi}\Phi \tag{3}$$

where r_{Φ} [0,1] controls the rate at which the algorithm (a) increases the probability of rejecting non-optimal steps, and (b) narrows the search to the neighborhood of the current best solution. Φ' is the updated parameter after each time the function is computed. Low initial Φ and r_{Φ} can lead SA towards local optima. Adequate initial values for Φ are such that the parameter space is fully sampled at the beginning of the simulation process. Values of r_{Φ} greater than one will gradually increase Φ and the breadth of the sample space. Corana et al. (1987) showed that a value of 0.85 for r_{Φ} is adequate to avoid local optima in complex problems. The algorithm ends by comparing the last N_{ε} values for the largest function values, where ε denotes a subjective small difference.

Messina et al. (2004) used this methodology to design ideotypes for five target environments in Argentina. Environments differed in magnitude and timing of drought stress. For every environment, the coupled model found ideotypes yielding at least 40% more than actual varieties grown in the region and fourfold higher than estimates from previous studies that used local sensitivity analyses (Fig. 8). When the Genegro-Soybean model was used, Messina et al. (2004) found that ideotypes yielded less than the one optimized using the CROPGRO-Soybean model. By restricting combinations of genetic coefficients to those that could be obtained through known combinations of loci, yield gain was less than that obtained when all coefficients were varied independently (Fig. 8). Thus it is clear that ideotype design needs to account for correlations among genetic coefficients due to pleiotropic effects and epistasis. These effects are commonly ignored in ideotype design (e.g. Hammer et al., 1996; Aggarwal et al., 1997; Sinclair and Muchow, 2001), misrepresenting the ideotype capacity of adaptation to the environment. If these effects are ignored, the optimization process can lead to infeasible solutions. Gene-based models can increase the realism of the ideotype by selecting infeasible combinations of traits and overestimating genetic gains. Gene-based models however, are not available for all crops and all traits for a given crop. Application of traditional crop models in combination with global optimization may assist genetic improvement despite their limitations.

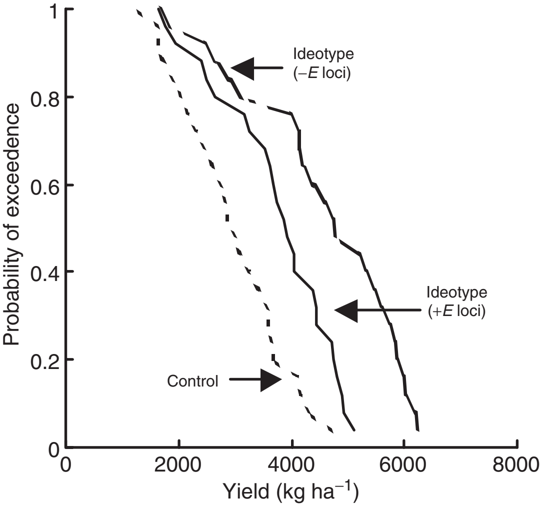


Figure 8. Probabilities of exceedence of simulated yields for a reference cultivar (control), an ideotype designed using SA and (a) CROPGRO (−*E* loci), and (b) Genegro-Soybean (+*E* loci). Simulated over 10 years in Argentina.

6.2. Other applications of crop models in plant breeding

Sensitivity analyses have been useful to evaluate all genotypes resulting from a relatively low number of genes of interest. Hoogenboom et al. (2004) used CSM-Genegro model to show how specific genes and gene combinations at seven loci can simulate yield and yield variability for four sites in major bean production areas in the USA under actual and possible global change scenarios. To facilitate examining multiple traits across genotypes and temperature regimes, simulation outputs were plotted as pseudo-maps using ArcMap 8.2 (Environmental Systems Research Institute, Inc., Redlands, California). Phenotypic traits or environments were plotted as rows and genotypes as columns in a matrix arrangement. This technique allows the rapid identification of best genotypes and the causes responsible for the GEI.

The biplot graphic display (Gabriel, 1971) is another methodology that can be used to visualize and identify GEI patterns (Cooper and Delacy, 1994) in a simulated multi environment trial. Figure 9 shows a biplot of simulated yield results for soybean. Soybean yields were simulated using Genegro-soybean for 32 combinations of five *E* loci in six locations in Illinois, USA, over five years. Locations were classified according to the latitude (high, mid- and low latitudes). Other outputs generated by the model could be used to improve this classification (e.g. water stress, temperature, Fig. 1). By looking at the orthogonal projection into the environment vectors, it can be shown that genotypes *e*1*e*2*E*3*e*4*E*5, *e*1*E*2*E*3*e*4*E*5 and *E*1*e*2*e*3*e*4*e*5 are best adapted to high, mid- and low latitudes, respectively. These genotypes represent a gradient of photoperiod sensitivity; *E*1*e*2*e*3*e*4*e*5 was the most sensitive in the set. Higher photoperiod sensitivity is required in a lower latitude environment for the genotype to fully explore the growing season, maximizing the capture of resources such as light, nitrogen and water. Genotypes too

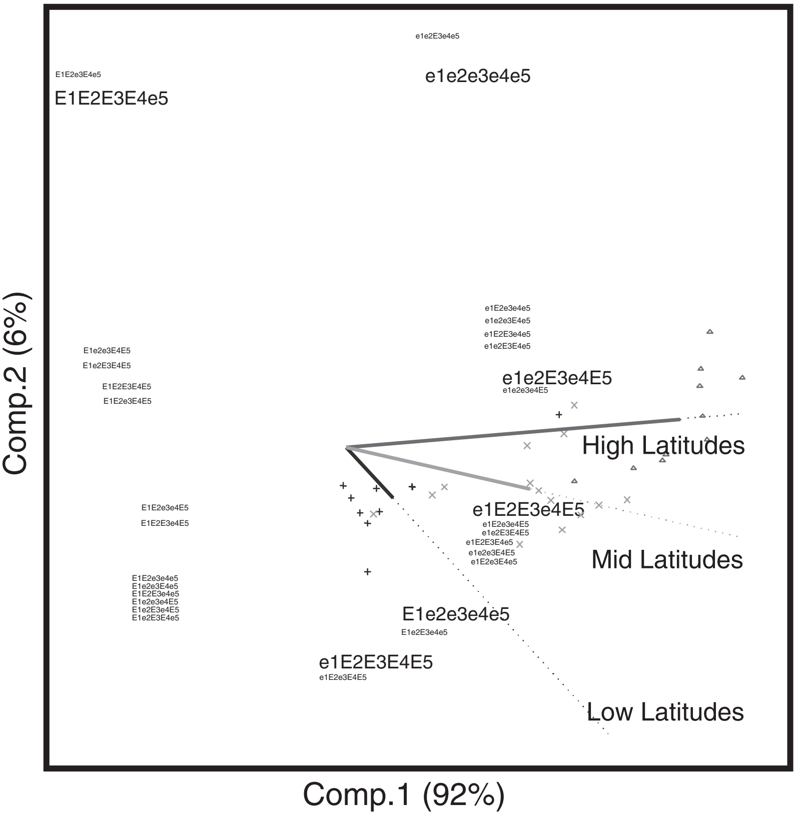


Figure 9. Environment-standardized GGE biplot of seed yield for a set of synthetic soybean isolines differing only in the *E* loci (*E*1 through *E*7) makeup. Simulations were conducted with Genegro-Soybean for six environments over five years in Illinois. Environments were stratified by latitude.

sensitive to photoperiod such as *E*1*E*2*E*3*E*4*e*5 have low yields due to delayed maturity and freeze damage. Note that the projection of this genotype in any of the environment vectors falls opposite to the projections of best-adapted genotypes. The Biplot can help to rapidly identify positive GEI for further study.

Either sensitivity analyses or optimization can be used to characterize ideotypes by their genotypes. However, a second optimization may be necessary to create an ideotype by gene pyramiding (Bertrand et al., 2004). Servin et al. (2004) proposed an algorithm to create ideotypes, by combining into a single genotype a series of target genes coming from different parents. In other words, the procedure searches for the succession of crosses over a number of generations that produces the optimal pedigree corresponding to the best gene-pyramiding scheme.

Gene-based models can assist in the optimization of breeding strategies *via* computer simulation. This application of crop models to plant breeding is now feasible due to the availability of high-speed computers and critical developments in modeling quantitative genetics and plant breeding programs, more specifically the development of QU-GENE (Podlich and Cooper, 1998). QU-GENE was developed on the basis of the

E(N:K) framework as a simulation platform for the study of genetic models. The *E(N:K)* framework allows the definition of a family of genetic models based on the frequencies of environments (*E*) occurring in the TPE, the number of genes (*N*) in the genetic network and the degree of gene interaction or epistasis (*K*). QU-GENE combines deterministic and stochastic features of both linear and adaptation landscape models. The stochastic components in QU-GENE allow the simulation of gene recombination, segregation of genes based on map distances and the search for genotypes with highest fitness in the adaptation landscape. Note the analogy between this component in QU-GENE and optimization algorithms. Linear models are the basis to predict phenotypes from genotypes using conventional quantitative genetic models. Thus, there is no biophysical connection between genotypes and phenotypes. Specific modules in QU-GENE allow the simulation of alternative breeding strategies: (a) mass selection, (b) pedigree and single-seed descent, (c) double haploid, (d) S1 recurrent selection and (e) half-sib reciprocal recurrent selection, by managing the creation, evaluation and selection of genotypes within the breeding program.

Dynamic simulation biophysical models of plant growth and development (Jones et al., 2003; Keating et al., 2003; van Ittersum et al., 2003) can be used to predict phenotypes for environments in the TPE or MET using genotypic information generated by QU-GENE, and parameter models of the form presented in section 5.2 (Fig. 6). Under this scheme, epistasis and GEI are emergent properties of the dynamics of the crop simulation model. Further statistical analysis using QU-GENE would determine the best-adapted genotypes based on simulated yield, and these would be carried forward during the simulation of the breeding program. Chapman et al. (2003) demonstrated this approach using Sorghum breeding for dry land environments in Australia as a case study. This particular implementation used APSIM-Sorg (Keating et al., 2003) to predict sorghum yields for six locations and 108 years. Their parameter model included four traits: transpiration efficiency, flowering time, osmotic adjustment and stay green. The parameter model was developed mainly on expert knowledge due to the limitations on the current understanding of genetic control of the traits. However, this study conducted to demonstrate the feasibility of the approach if such information was available, proved useful to demonstrate: (a) the potential for studying the dynamics of breeding programs *via* biophysical and quantitative genetic modeling and simulation, (b) that additive effects at the trait level can give rise to complex epistatic and GEI effects at the physiological/crop level, and (c) an approach to design ideotypes, which is the last genotype at the end of the simulation of the breeding program.

7. Discussion

Crop models can assist plant breeding by integrating physiological and biochemical understanding, agronomic practices, the environment and genetic information. Use of such models in plant breeding will increase for characterizing environments, for assessing the value of putative traits, for understading adaptation and for designing improved plant types. They will also be used for integrating knowledge and promoting discussion among researchers across disciplines and, as a means to make inferences from experimental plots to the TPE. Advances in plant genetics, genomics, biochemistry and allied fields offer opportunities for improving the representation of growth and development process.

We presented a simple methodology based on linear equations to link genes to crop model parameters. Future approaches could improve sub-models using more complex representations of gene action. Crop models, by representing plants as systems in a modular manner, will provide the necessary organized framework to incorporate the new representations, yet retaining the required predictability necessary for plant breeding applications.

References

Aggarwal, P.K., Kropff, M.J., Cassman, K.G., ten Berge, H.F.M., 1997. Simulating genotypic strategies for increasing rice yield potential in irrigated, tropical environments. Field Crops Research 51, 5–17.

Akkaya, M.S., Bhagwat, A.A., Cregan, P.B., 1992. Length polymorphisms of simple sequence repeat DNA in soybean. Genetics 132, 1131–1139.

Bertrand, S., Martin, O.C., Mézard, M., Hospital, F., 2004. Toward a theory of marker-assisted gene pyramiding. Genetics 168, 513–523.

Boote, K.J., 1999. Concepts for calibrating crop growth models. In: Hoogenboom, G., Wilkens, P.W., Tsuji, G.Y. (Eds), DSSAT Version 3. A Decision Support System for Agrotechnology Transfer, Vol. 4. University of Hawaii, Honolulu, HI, pp. 179–200.

Boote, K.J., Jones, J.W., 1986. Applications of, and limitations to, crop growth simulation models to fit crops and cropping systems to semi-arid environments. In: Bidinger, F.R., Johansen, C. (Eds), Drought Research Priorities for the Dryland Tropics. International Crops Research Institute for the Semi-Arid Tropics, Patancheru, A.P. 502 324, India, pp. 63–75.

Boote, K.J., Tollenaar, M., 1994. Modeling genetic yield potential. In: Boote, K.J. et al. (Eds), Physiology and Determination of Crop Yield. ASA, CSSA and SSSA, Madison, WI, pp. 533–561.

Boote, K.J., Jones, J.W., Hoogenboom, G.H., 1998. Simulation of crop growth: CROPGRO model. In: Peart, R.M., Curry, R.B. (Eds), Agricultural Systems Modeling and Simulation. Marcel Dekker, New York, pp. 651–693.

Boote, K.J., Kroft, M.J., Brindraban, P.S., 2001. Physiology and modeling of traits in crop plants: implications for genetic improvement. Agricultural Systems 70, 395–420.

Boote, K.J., Jones, J.W., Batchelor, W.D., Nazfiger, E.D., Myers, O., 2003. Genetic coefficients in the CROPGRO-Soybean model: links to field performance and genomics. Agronomy Journal 95, 32–51.

Botstein D., White, R.L., Skolnick, M., Davis, R.W., 1980. Construction of genetic map of man using restriction fragment length polymorphisms. American Journal of Human Genetics 32, 314–331.

Brown, P.O., Botstein, D., 1999. Exploring the new world of the genome with DNA microarrays. Nature Genetics 21, 33–37.

Carlborg, O., Haley, C.S., 2004. Epistasis: too often neglected in complex traits studies? Nature Reviews Genetics 5, 618–625.

Chapman, S.C., Cooper, M., Hammer, G.L., Butler, D., 2000. Genotype by environment interactions affecting grain sorghum. II. Frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. Australian Journal of Agricultural Research 50, 209–222.

Chapman, S.C., Cooper, M., Hammer, G.L., 2002. Using crop simulation to generate genotype by environment interaction effects for sorghum in water-limited environments. Australian Journal of Agricultural Research 53, 379–389.

Chapman, S., Cooper, M., Podlich, D., Hammer, G., 2003. Evaluating plant breeding strategies by simulating gene action in dryland environment effects. Agronomy Journal 95, 99–113.

Cober, E.R., Tanner, J.W., Voldeng, H.D., 1996. Genetic control of photoperiod response in early-maturing, near-isogenic soybean lines. Crop Science 36, 601–605.

Comstock, R.E., 1977. Quantitative genetics and the design of breeding programs. Proceedings of the International Conference on Quantitative Genetics, August 16–21, 1976. Iowa State University Press, Ames, USA, pp. 705–718.

Cooper, M., 1999. Concepts and strategies for plant adaptation research in rainfed lowland rice. Field Crops Research 64, 13–34.

Cooper, M., DeLacy, I.H., 1994. Relationship among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theoretical and Applied Genetics 88, 561–572.

Cooper, M., Hammer, G.L. (Eds), 1996. Plant Adaptation and Crop Improvement. CAB International, Wallingford, UK, p. 636.

Cooper, M., Byth, D.E., DeLacy, I.H., 1993. A procedure to assess the relative merit of classification strategies for grouping environments to assist selection in plant breeding regional evaluation trials. Field Crops Research 35, 63–74.

Cooper, M., Chapman, S.C., Podlich, D.W., Hammer, G.L., 2002. The GP problem: quantifying gene-to-phenotype relationships. In: Silico Biology (Available online at http://www.bioinfo.de/isb/; verified October 10, 2003).

Corana, A., Marchesi, M., Martini, C., Ridella, S., 1987. Minimizing multimodal functions of continuous variables with the simulated annealing algorithm. ACM Transactions on Mathematical Software 13, 262–280.

Daniell, H., Dhingra, A., 2002. Multigene engineering: dawn of an exciting new era in biotechnology. Current Opinion in Biotechnology 13, 136–141.

Darvasi, A., Pisanté-Shalom, A., 2002. Complexities in the genetic dissection of quantitative trait loci. Trends in Genetics 18, 489–491.

de Wit, C.T., Penning de Vries, F.W.T., 1983. Crop growth models without hormones. Netherlands Journal of Agricultural Research 31, 313–323.

Doerge, R.W., Zeng, Z-B., Weir, B.S., 1997. Statistical issues in the search for genes affecting quantitative traits in experimental populations. Statistical Science 12, 195–219.

Dornhoff, G.M., Shibles, R.M., 1970. Varietal differences in net photosynthesis of soybean leaves. Crop Science 10, 42–45.

Duncan, W.G., McCloud, D.E., McGraw, R.L., Boote, K.J., 1978. Physiological aspects of peanut yield improvement. Crop Science 18, 1015–1020.

Eberhardt, S.A., Russell, W.A., 1966. Stability parameters for comparing varieties. Crop Science 6, 36–40.

Echarte, L., Luque, S., Andrade, F.H., Sadras, V.O., Cirilo, A., Otegui, M.E., Vega, C.R.C., 2000. Response of maize kernel number to plant density in Argentinean hybrids released between 1965 and 1993. Field Crops Research 68, 1–8.

Edmeades, G.O., Baeziger, M., Ribaut, J., 2000. Chapter 6. Maize improvement for drought-limited environments. In: Otegui, M.E., Slafer, G.A. (Eds), Physiological Basis for Maize Improvement. Food Products Press, New York, pp. 75–112.

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics 5, 435–445.

Falconer, D.S., Mackay, T.F.C., 2001. Introduction to Quantitative Genetics, 4th Edition. Prentice Hall, England, pp. 464.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principle component analysis. Biometrika 58, 453–467.

Ghaemmaghami, S., Huh, W., Bower, K., Howson, R., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S., 2003. Global analysis of protein expression in yeast. Nature (London) 425, 737–741.

Goffe, W.L., Ferrier, G.D., Rogers, J., 1994. Global optimization of statistical functions with simulated annealing. Journal of Econometrics 60, 65–99.

Hammer, G.L., Butler, D.G., Muchow, R.C., Meinke, H., 1996. Integrating physiological understanding and plant breeding via crop modeling and optimization. In: Cooper, M., Hammer, G.L. (Eds), Plant Adaptation and Genetic Improvement. CAB International, New York, pp. 419–442.

Hammer, G.L., Kropff, M.J., Sinclair, T.R., Porter, J.R., 2002. Future contributions of crop modeling from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. European Journal of Agronomy 18, 15–31.

Hammer, G.L., Sinclair, T.R., Chapman, S.C., van Oosterom, E., 2004. On systems thinking, systems biology, and the in silico plant. Plant Physiology 134, 909–911.

Hartkamp, A.D., White, J.W., Rodriguez Aguilar, A., Bänziger, M., Srinivasan, G., Granados, G., Crossa, J., 2000. Maize Production Environments Revisited: A GIS-based Approach. CIMMYT, Mexico, D.F.

Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31, 423–447.

Hoogenboom, G., White, J.W., 2003. Improving physiological assumptions of simulation models by using gene-based approaches. Agronomy Journal 95, 82–89.

Hoogenboom, G., White, J.W., Acosta-Gallegos, J., Gaudiel, R.G., Myers, J.R., Silbernagel, M.J., 1997. Evaluation of a crop simulation model that incorporates gene action. Agronomy Journal 89, 613–620.

Hoogenboom, G., White, J.W., Messina, C.D., 2004. From genome to crop: integration through simulation modeling. Field Crops Research 90, 45–163.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. European Journal of Agronomy 18, 235–265.

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Brown, S., Bristow, K.L., Asseng, S., Chapman, S., McCown, R.L., Freebairn, D.M., Smith, C.J., 2003. An overview of APSIM, a model designed for farming systems simulation. European Journal of Agronomy 18, 267–288.

Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science (Washington DC) 220, 671–680.

Kropff, M.J., Haverkort, A.J., Aggarwal, P.K., Kooman, P.L., 1995. Using systems approaches to design and evaluate ideotypes for specific environments. In: Bouma, J. et al. (Eds), Eco-regional Approaches for Sustainable Land Use. Kluwer Academic Publishers, Dordrecht, pp. 417–435.

Kumudini S., Hume, D.J., Chu, G., 2001. Genetic improvement in short season soybeans: I. dry matter accumulation, partitioning, and leaf area duration. Crop Science 41, 391–398.

Kumudini, S., Hume, D.J., Chu, G., 2002. Genetic improvement in short-season soybeans: II. nitrogen accumulation, remobilization, and partitioning. Crop Science 42, 141–145.

Lander, E.S., Botstein, D., 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121, 185–199.

Lander, E.S., Schork, N.J., 1994. Genetic dissection of complex traits. Science 265, 2037–2048.

Löffler, C.M., Wei, J., Fast, T., Gogerty, J., Langton, S., Bergman, M., Merrill, R., Cooper, M., 2005. Classification of maize environments using crop simulation and geographic information systems. Crop Science 45, 1708–1716.

Lopez Pereira, M., Sadras, V.O., Trapani, N., 1999a. Genetic improvement of sunflower in Argentina between 1930 and 1995. I. Yield and its components. Field Crops Research 62, 157–166.

Lopez Pereira, M., Trapani, N., Sadras, V.O., 1999b. Genetic improvement of sunflower in Argentina between 1930 and 1995. II. Phenological development, growth and source-sink relationship. Field Crops Research 63, 247–254.

Lopez Pereira, M., Trapani, N., Sadras, V.O., 2000. Genetic improvement of sunflower in Argentina between 1930 and 1995. Part III. Dry matter partitioning and grain composition. Field Crops Research 67, 215–221.

Mackay, T.F.C., 2001. The genetic architecture of quantitative traits. Annual Review Genetics 35, 303–339.

Mavromatis, T., Boote, K.J., Jones, J.W., Irmak, A., Shinde, D., Hoogenboom, G., 2001. Developing genetic coefficients for crop simulation models with data from crop performance trials. Crop Science 41, 40–51.

Messina, C.D., 2003. Gene-based systems approach to simulate soybean growth and development and application to ideotype design in target environments. Ph.D. Thesis, University of Florida, Gainesville, Florida.

Messina, C.D., Boote, K.J., Jones, J.W., 2002. Basis for modeling genetic controls of pod addition duration in soybean. Annual Meetings Abstracts (CD-ROM). American Society of Agronomy, Madison, WI.

Messina, C.D., Jones, J.W., Boote, K.J., Vallejos, C.E., 2004. Linking biophysical models, plant genomics and optimization algorithms for plant ideotype design. Biological Systems Simulation Conference, March 8–10, Gainesville, FL, p. 53.

Messina, C.D., Jones, J.W., Boote, K.J., Vallejos, C.E., 2006. A gene-based model to simulate soybean development and yield responses to environment. Crop Science 46, 456–466.

Paruelo, J.M., Sala, O.E., 1993. Effect of global change on maize production in the Argentine Pampas. Climate Research 3, 161–167.

Phillips R.L., Vasil, I.K., 2001. DNA-based markers in plants, 2nd Edition. Kluwer Academic Publishers, Dordrecht, Holland, pp. 301–317.

Podlich, D.W., Cooper, M., 1998. QU-GENE: a simulation platform for quantitative analysis of genetic models. Bioinformatics 14, 632–653.

Pollak, L.M., Corbett, J.D., 1993. Using GIS datasets to classify maize-growing regions in Mexico and Central America. Agronomy Journal 85, 1133–1139.

Raghothama, C., Pandey, A., 2003. Absolute systems biology-measuring dynamics of protein modifications. Trends in Biotechnology 21, 467.

Reymond, M., Muller, B., Leonardi, A., Charcosset, A., Tardieu, F., 2003. Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiology 131, 664–675.

Ritchie, J.T., 1986. The CERES-Maize model. In: Jones C.A., Kiniry, J.R. (Eds), CERES Maize: A Simulation Model of Maize Growth and Development. Texas A&M University Press, College Station, TX, pp. 1–6.

Royce, F.S., Jones, J.W., Hansen, J.W., 2001. Model-based optimization of crop management for climate forecast application. Transactions of ASAE 44, 1319–1327.

Runge, E.C.A., 1968. Effects of rainfall and temperature interactions during the growing season on corn yield. Agronomy Journal 60, 503–507.

Sax, K., 1923. The association of size difference with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8, 552–560.

Servin, B., Martin, O.C., Mézard, M., Hospital, F., 2004. Toward a theory of marker-assisted gene pyramiding. Genetics 168, 513–523.

Shorter, R., Lawn, R.J., Hammer, G.L., 1991. Improving genotypic adaptation in crops – a role for breeders, physiologists and modelers. Experimental Agriculture 27, 155–175.

Sinclair, T.R., Muchow, R.C., 2001. System analysis of plant traits to increase grain yield on limited water supplies. Agronomy Journal 93, 263–270.

Stewart, D.W., Cober, E.R., Bernard, R.L., 2003. Modeling genetic effects on the photothermal response of soybean phenological development. Agronomy Journal 95, 65–70.

Stuber, C.W., Polacco, M., Senior, M.L., 2003. Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield potential. Crop Science 39, 1571–1583.

Syvänen, A., 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nature Reviews Genetics 2, 930–942.

Tanksley S.D., Orton, T.J., 1983. Isozymes in plant genetics and breeding. Elsevier, Amsterdam, Vol. A pp. 516, and Vol. B pp. 472.

Tardieu, F., 2003. Virtual plants: modelling as a tool for the genomics of tolerance to water deficit. Trends in Plant Sciences 8, 9–14.

Valenzuela, D.M., Murphy, A.J., Frendewey, D., Gale, N.W., Economides, A.N., Auerbach, W., Poueymirou, W.T., Adams, N.C., Rojas, J., Yasenchack, J., Chernomorsky, R., Boucher, M., Elsasser, A.L., Esau, L., Zheng, J., Griffiths, J.A., Wang, X., Su, H., Xue, Y., Dominguez, M.G., Noguera, I., Torres, R., Macdonald, L.E., Stewart, A.F., DeChiara, T.M., Yancopoulos, G., 2003. High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. Nature Biotechnology 21, 652–659.

van Ittersum, M.K., Leffelaar, P.A., van Keulen, H., Kropff, M.J., Bastiaans, L., Goudriaan, J., 2003. On approaches and applications of the Wageningen crop models. European Journal of Agronomy 18, 201–234.

Villalobos, F.J., Hall, A.J., Ritchie, J.T., Orgaz, F., 1996. OILCROP-SUN: a development, growth and yield model of the sunflower crop. Agronomy Journal 88, 403–415.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Vandelee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., Zabeau, M., 1995. AFLP a new technique for DNA fingerprinting. Nucleic Acids Research 23, 4407–4414.

Wang, S., Basten, C.J., Zeng, Z.B., 2003. Windows QTL Cartographer 2.0. Department of Statistics, North Carolina State University, Raleigh, NC. (Available online at http://statgen.ncsu.edu/qtlcart/WQTLCart.htm).

Weckwerth, W., 2003. Metabolomics in systems biology. Annual Review of Plant Biology 54, 669–689.

Welch, S.M., Roe, J.L., Zhanshan, D., 2003. A genetic neural network model of flowering time control in Arabidopsis thaliana. Agronomy Journal 95, 71–81.

White, J.W., 1998. Modeling and crop improvement. In: Tsuji, G.Y. et al. (Eds), Understanding options for Agricultural Production. Kluwer Academic Publishers, Dordrecht, pp. 179–188.

White, J.W., Hoogenboom, G., 1996. Simulating effects of genes for physiological traits in a process-oriented crop model. Agronomy Journal 88, 416–422.

White, J.W., Hoogenboom, G., 2003. Gene-based approaches to crop simulation: past experiences and future opportunities. Agronomy Journal 95, 52–64.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A., Tingey, S.V., 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research 18, 6531–6535.

Yin, X., Stam, P., Kropff, M.J., Schapendonk, A.H.C.M., 2003. Crop modeling, QTL mapping, and their complementary role in plant breeding. Agronomy Journal 95, 90–98.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 12

Parameterization and evaluation of a corn crop model

D. Wallach

1. Introduction	**337**
2. Model and data	**337**
3. Parameter estimation	**338**
4. Comparison with data	**340**
5. Conclusions	**347**
Reference	**348**

1. Introduction

A corn crop model was developed with the intention of coupling it with a decision model in order to evaluate irrigation strategies for corn. The decision model is described in Chapter 19, and the uses of the combined model are also presented therein. This chapter presents information about parameter estimation and evaluation of the crop model.

Section 2 describes the model very briefly and presents the data used for parameter estimation. Section 3 describes how the model parameters were estimated. Section 4 presents an evaluation of the parameterized model, based on comparisons with the data. Conclusions are presented in Section 5.

2. Model and data

The state variables in the model are thermal time, leaf area index, fraction senescent leaves, aboveground biomass, rooting depth, harvest index and soil water in 4 soil layers. Thermal time is used to calculate development stage and potential leaf area increase. Total and senescent leaf area index are used to calculate radiation interception which in turn is used to calculate potential biomass increase. Rooting depth defines the soil depth from which the plant can extract water. Soil moisture in the root zone is used to calculate the daily ratio of actual to potential transpiration, and from that ratio, stress factors are calculated which reduce the daily increases in leaf area index, biomass and harvest index. Nitrogen stress is not taken into account. Final aboveground biomass times final harvest index gives yield. The model equations can be found in Wallach et al. (2001).

Over the last few years, a fairly large number of experiments comparing irrigation treatments for corn have been carried out in southwest France. As many as possible of these data were collected into a database to serve as the basis for parameterization and evaluation of the model. The data span the period from 1986 to 1997 and come from

16 different locations. Overall there are 181 different situations (i.e. site-year-treatment combinations). Crop management in all these experiments was based on recommended practices except for irrigation, which was varied to cover a range of strategies.

Yield data were available for every experimental situation, and final biomass for most. For most situations there were also several measurements of leaf area index and biomass during the growing season. Unfortunately, there were very few soil water measurements available and so these data were not used.

3. Parameter estimation

Parameter estimation for this model is described in Wallach et al. (2001). There are 24 parameters in the model. No attempt was made to estimate all parameters from the available data. It was felt that this would not be a good idea. Not only is the number of parameters fairly large, but also certain parameters are only remotely related to the measured data and so would be very poorly estimated.

The parameter estimation procedure consisted of 4 steps.

1. An initial value was assigned to each parameter, based on information in the literature or on an informed guess.
2. A criterion MSE_{m} of model fit to the data was defined. The criterion takes into account model errors with respect to yield, aboveground biomass and leaf area index. The smaller the value of the criterion the better the fit of the model to the data.
3. A forward regression procedure was used to create an ordered list of parameters to be adjusted to the data. First, each parameter was adjusted individually to the data. The parameter that gave the smallest value of MSE_{m} was the first parameter in the list. Next all combinations of this first parameter and one other were adjusted to the data. The combination that gave the smallest value MSE_{m} gave the second parameter in the list. The procedure was continued until the best four parameters to adjust were identified. At the end of this step we had four candidate models, corresponding to the model with 1, 2, 3 or 4 adjusted parameters. We could have extended the list to include more parameters but this is not necessary as shown by the next step.
4. Cross validation was used to estimate prediction error for each candidate model. The final model chosen was that with the smallest estimated prediction error.

The criterion of model fit for the model with estimated parameter vector $\hat{\theta}_{\mathrm{m}}$ is

$$MSE_{\mathrm{m}} = \frac{w_{\mathrm{Y}}}{N}\sum_{i=1}^{N} MSE_{\mathrm{Y},i}(\hat{\theta}_{\mathrm{m}}) + \frac{w_{\mathrm{B}}}{N_{\mathrm{B}}}\sum_{i=1}^{N_{\mathrm{B}}} MSE_{\mathrm{B},i}(\hat{\theta}_{\mathrm{m}}) + \frac{w_{\mathrm{L}}}{N_{\mathrm{L}}}\sum_{i=1}^{N_{\mathrm{L}}} MSE_{\mathrm{L},i}(\hat{\theta}_{\mathrm{m}}) \tag{1}$$

where

$$MSE_{\mathrm{Y},i}(\hat{\theta}_{\mathrm{m}}) = [y_i - \hat{y}_i(\hat{\theta}_{\mathrm{m}})]^2$$

$$MSE_{\mathrm{B},i}(\hat{\theta}_{\mathrm{m}}) = \frac{1}{M_{\mathrm{B},i}} \sum_{j=1}^{M_{\mathrm{B},i}} [b_{ij} - \hat{b}_{ij}(\hat{\theta}_{\mathrm{m}})]^2$$

$$MSE_{\mathrm{L},i}(\hat{\theta}_{\mathrm{m}}) = \frac{1}{M_{\mathrm{L},i}} \sum_{j=1}^{M_{\mathrm{L},i}} [l_{ij} - \hat{l}_{ij}(\hat{\theta}_{\mathrm{m}})]^2$$

The first term in the criterion refers to yield, the second to biomass and the third to leaf area index. The weightings are $w_{\mathrm{Y}} = \sigma_{\mathrm{Y}}^{-2}$, $w_{\mathrm{B}} = \sigma_{\mathrm{B}}^{-2}$ and $w_{\mathrm{L}} = \sigma_{\mathrm{L}}^{-2}$, where σ_{Y}^2, σ_{B}^2 and σ_{L}^2 are the empirical measurement variances based on replicates, for yield, biomass and leaf area index, respectively. N is the total number of situations, and N_{B} and N_{L} are the number of situations with at least one biomass or leaf area index measurement. In $MSE_{\mathrm{Y},i}(\hat{\theta}_{\mathrm{m}})$, y_i and $\hat{y}_i$ are respectively observed and calculated yields for the ith situation. $MSE_{\mathrm{B},i}(\hat{\theta}_{\mathrm{m}})$ is the squared error for biomass estimation, averaged over all the $M_{\mathrm{B},i}$ biomass measurements for situation i. Here b_{ij} and $\hat{b}_{ij}$ are respectively observed and calculated biomass values for the jth biomass measurement date in the ith situation. The term for the average squared error for leaf area index estimation is analogous. $M_{\mathrm{L},i}$ is the number of leaf area index measurements in situation i and l_{ij} and $\hat{l}_{ij}$ are respectively observed and calculated leaf area index values for the jth measurement date in the ith situation. If there are no biomass or leaf area measurements for a situation, then the corresponding mean squared error is set to 0.

In generalized least squares, the criterion to be minimized includes the variance–covariance matrix between errors for different measurements. In the present case that would involve estimating the covariances between yield, biomass at various dates and leaf area index at various dates. Instead, we chose to use the simplified criterion of Eq. (1). This criterion weights each type of measurement by the inverse of the measurement error variance but ignores covariances between different types of measurement. Furthermore, no attempt is made to estimate covariances between different dates for the same type of measurement. Rather, the criterion uses the squared error averaged over all measurements of a given type for each situation.

The cross validation estimate of prediction error for the combined criterion of Eq. (1), noted $\widehat{MSEP}_{\mathrm{m}}$, is calculated as

$$\widehat{MSEP}_{\mathrm{m}} = \frac{w_{\mathrm{Y}}}{N} \sum_{i=1}^{N} MSE_{\mathrm{Y},i}(\hat{\theta}_{\mathrm{m},-i}) + \frac{w_{\mathrm{B}}}{N_{\mathrm{B}}} \sum_{i=1}^{N_{\mathrm{B}}} MSE_{\mathrm{B},i}(\hat{\theta}_{\mathrm{m},-i}) + \frac{w_{\mathrm{L}}}{N_{\mathrm{L}}} \sum_{i=1}^{N_{\mathrm{L}}} MSE_{\mathrm{L},i}(\hat{\theta}_{\mathrm{m},-i})$$

where $\hat{\theta}_{\mathrm{m},-i}$ is calculated in the same way as $\hat{\theta}_{\mathrm{m}}$ but excluding the data from situation i and from situations from the same site and/or year.

Table 1 shows the results of the forward regression procedure and the estimated mean squared error of prediction $\widehat{MSEP}_{\mathrm{m}}$ values. As expected, MSE_{m} decreases systematically as more parameters are adjusted to the data whereas $\widehat{MSEP}_{\mathrm{m}}$ has a minimum.

Table 1. The candidate models with different numbers of adjusted parameters. The model chosen is that with 3 adjusted parameters.

Number of parameters adjusted	0	1	2	3	4
Names of adjusted parameters	–	p2logi	p2logi, r2hi	p2logi, r2hi, himax	p2logi, r2hi, himax, p2evap
$\sqrt{MSE_{\mathrm{m}}}$	10.72	9.15	8.01	7.68	7.49
$\sqrt{\hat{MSEP}_{\mathrm{m}}}$	10.72	9.55	8.39	8.20	8.34
$\sqrt{\hat{MSEP}_{\mathrm{Y,m}}}$ (t/ha)	2.48	2.17	1.68	1.50	1.57
$\sqrt{\hat{MSEP}_{\mathrm{B,m}}}$ (t/ha)	2.31	2.35	2.36	2.30	2.38
$\sqrt{\hat{MSEP}_{\mathrm{L,m}}}$	1.17	1.0	0.95	1.01	0.99

The minimum occurs with 3 adjusted parameters and therefore this is chosen as the candidate model.

We are interested in prediction not only of the combined criterion but also individually of yield, biomass and LAI. The estimated mean squared errors of prediction for these quantities, noted respectively $MSEP_{\mathrm{Y,m}}$, $MSEP_{\mathrm{B,m}}$ and $MSEP_{\mathrm{L,m}}$ are also given in Table 1.

The use of cross-validation is meant to avoid over-parameterization, i.e. estimating more parameters than the number that minimizes prediction error. However, using estimated *MSEP* values to choose among candidate models creates a "selection bias". This is easily understood if we imagine selecting one model from among several, all with the same true value of *MSEP*. The estimated *MSEP* values will in general be different because of estimation error. By construction we will choose the model with the smallest *MSEP* value. That is, the selection procedure leads us to choose an *MSEP* value that is underestimated, though it is difficult to know by how much.

4. Comparison with data

We now have a model with its estimated parameters, as well as estimates of prediction error. We would however like to go further in the analysis of model quality, in order to try to determine the major causes of model error. This should help to guide future efforts toward model improvement. This leads us to examine in detail the comparison between calculated and measured values.

4.1. Contributions of errors in yield components to errors in yield

Yield can be written as the product of final biomass and final harvest index. The first question concerns the relative importance of errors in these two terms. Let $r_{y,i} = y_i - \hat{y}_i$ be the model error in predicting yield for situation i, $r_{f,i} = f_i - \hat{f}_i$ the error in predicting

Table 2. Average absolute error in yield prediction and for the individual terms in Eq. (2).

Term	Value
$\frac{1}{N}\sum_{i=1}^{N}\lvert r_{y,i}\rvert$	1.09
$\frac{1}{N}\sum_{i=1}^{N}\lvert h_i r_{f,i}\rvert$	0.84
$\frac{1}{N}\sum_{i=1}^{N}\lvert f_i r_{h,i}\rvert$	0.54
$\frac{1}{N}\sum_{i=1}^{N}\lvert r_{f,i} r_{h,i}\rvert$	0.09

final biomass for situation i, where f_i and $\hat{f}_i$ are respectively measured and calculated final biomass values for this situation, and $r_{h,i} = h_i - \hat{h}_i$ the error in predicting final harvest index, where h_i and $\hat{h}_i$ are respectively measured and calculated final harvest index values for situation i. Then it can easily be shown that

$$r_{y,i} = y_i - \hat{y}_i = (f_i h_i - \hat{f}_i \hat{h}_i) = h_i r_{f,i} + f_i r_{h,i} - r_{f,i} r_{h,i} \tag{2}$$

Table 2 shows the average absolute value of each of the three terms on the right hand side of the equation. The first term, proportional to the error in final biomass, is somewhat larger than the second term which depends on the error in final harvest index, but both terms are appreciable. The conclusion is that it is worthwhile to work on reducing both errors. The third term, the interaction term, is small.

4.2. First analysis of errors

Table 3 shows for each output the mean squared error (*MSE*), the three terms the sum of which is equal to *MSE* (squared bias, *SDSD* and *LCS*) and the modeling efficiency (see Chapter 2).

Table 3. Mean squared error (*MSE*), the three terms in its decomposition (squared bias, *SDSD* and *LCS*) and modeling efficiency (*EF*) for various model outputs.

	MSE	*Bias*^{2}	*SDSD*	*LCS*	*EF*
Yield	1.96	0.003 (0%)	0.25 (13%)	1.71 (87%)	0.56
Final biomass	5.53	0.23 (4%)	0.20 (4%)	5.09 (92%)	0.71
Harvest index	0.00706	0.00053 (7%)	0.00602 (85%)	0.00051 (7%)	−0.016
Biomass, all dates	3.36	0.18 (5%)	0.12 (4%)	3.06 (91%)	0.92
LAI, all dates	0.63	0.02 (3%)	0.06 (9%)	0.56 (89%)	0.78

The most arresting feature is the problem with harvest index. The modeling efficiency is very low, so that the model is in fact slightly worse than just using the average of all measurements as a predictor. It is the *SDSD* term that makes the major contribution to the mean squared error. This indicates that the overall variability is very different between observed and calculated values. This is discussed in more detail below.

Figures 1 and 2 show calculated *versus* observed yield and yield residuals *versus* observed yield, respectively. One noticeable feature of these graphs is that yield is generally over-predicted for small yield values and under-predicted for the highest yields. This is perhaps easier to see from the residuals (Fig. 2) than from the graph of calculated *versus* observed values (Fig. 1). It is also clear that there are 3 values with very low observed yield values, and there is severe over-prediction for all three of these situations. The corresponding graphs for final biomass (not shown) do not show any obvious trends, though the three situations with very low observed yields again stand out.

Figures 3 and 4 show the results for harvest index. Figure 3 shows clearly that the range of calculated values is much smaller than the range of observed values, as already noted. The harvest index residuals show a tendency to over-prediction for small yield values and over-prediction for the highest yield values, similar to the results for yield (Fig. 4). The values for the three situations with very small yields are very severely over-predicted. The conclusion is that a good starting point for improving the model would be to try to improve the calculation of harvest index. In particular, harvest index should be made more variable, decreasing more in unfavorable conditions. It would also be worthwhile

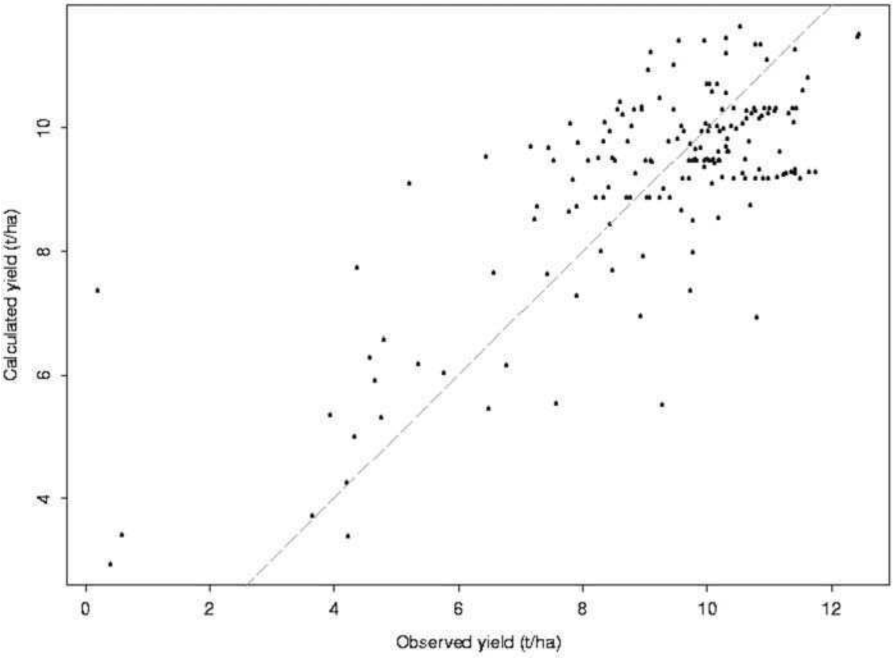


Figure 1. Calculated *versus* observed yields.

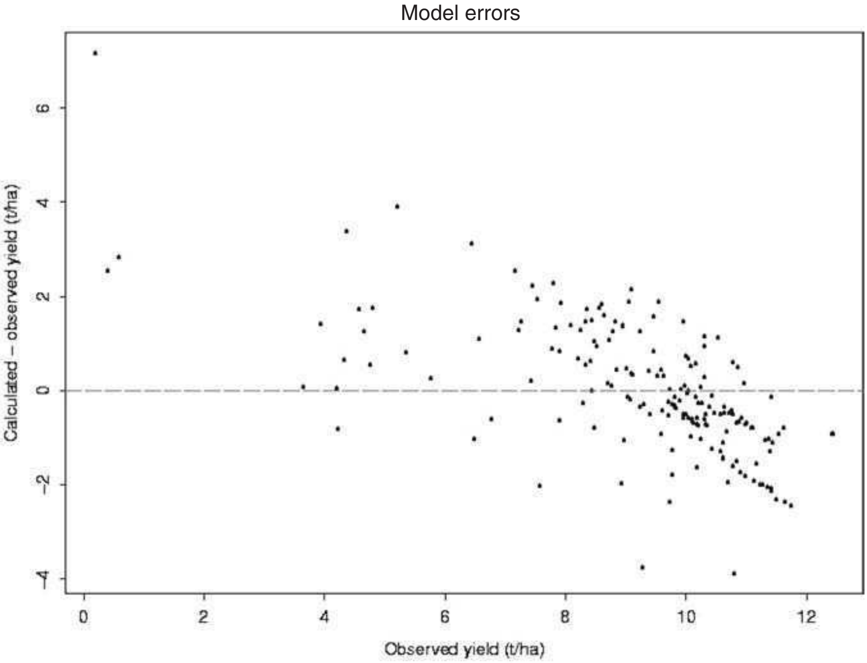


Figure 2. Yield residuals *versus* observed yield.

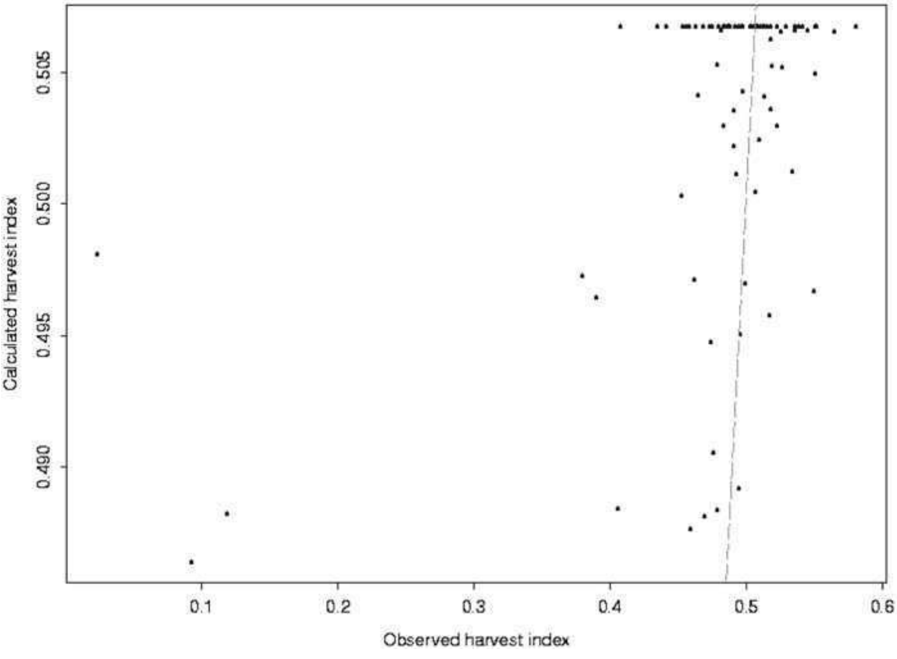


Figure 3. Calculated *versus* observed harvest index.

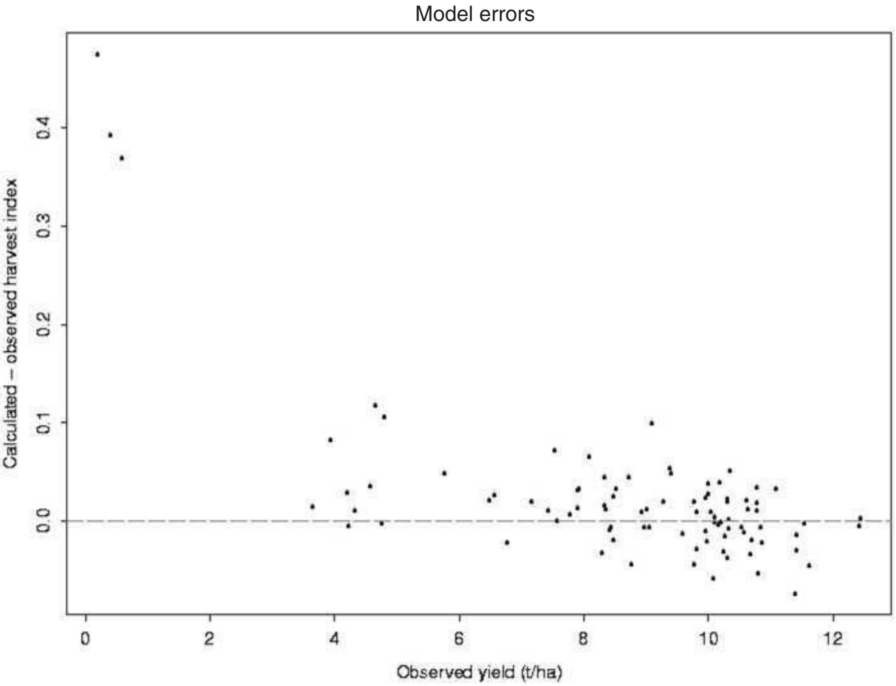


Figure 4. Harvest index residuals *versus* observed yield.

to examine in detail the three situations with very small yields. For these situations, both final biomass and harvest index are appreciably over-predicted.

Figures 5 and 6 show respectively calculated *versus* observed LAI values (all dates) and LAI residuals *versus* observed LAI. There seems to be over-prediction for small LAI values and under-prediction for large values. This is another systematic error that should be investigated further. The graphs of biomass for all dates are not presented because they do not exhibit any obvious trend.

4.3. Residuals as functions of explanatory variables

Residuals plotted against explanatory variables should not reveal any clear trends. If there are such trends, then the effect of that explanatory variable is incorrectly specified in the model. Crop models have literally thousands of explanatory variables (due to their dependence on daily climate), so it is neither of interest nor feasible to study each individual explanatory variable. For the climate explanatory variables, we use summary variables related to average climate over part or all of the growth period.

4.3.1. Variety

The only differences between varieties according to the model are in phenology. For each variety the model uses published values for the thermal time between emergence and flowering and between flowering and maturity. However, the training data cover a number

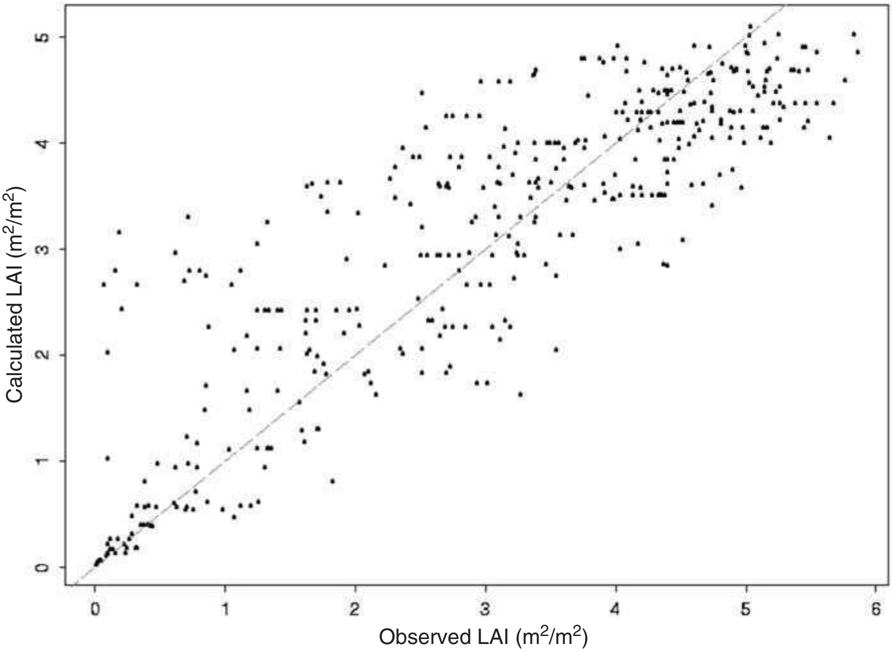


Figure 5. Calculated *versus* observed LAI.

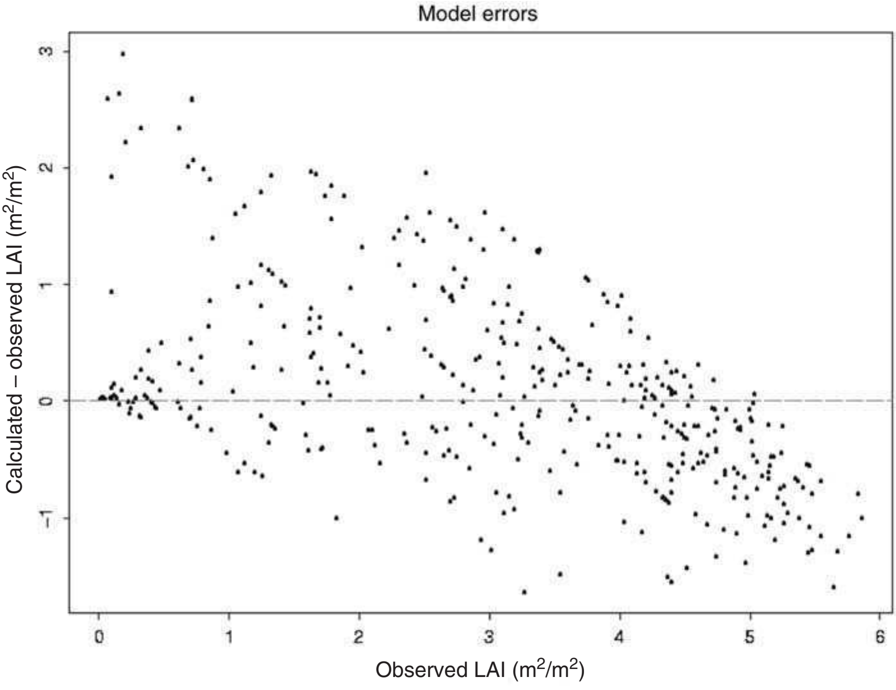


Figure 6. LAI residuals *versus* observed LAI.

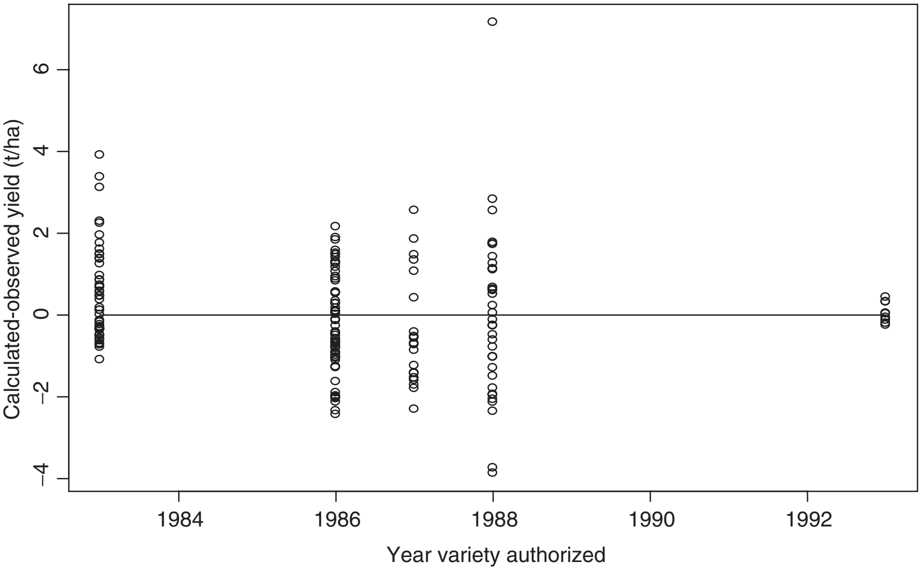


Figure 7. Yield residuals *versus* the year of introduction of the variety.

of years, and so it was thought that there might also be an increase in average yield not explained by phenology in going from earlier to later varieties. Figure 7 shows the yield residuals as a function of the year of introduction of the different varieties. No clear trend is apparent.

4.3.2. Density

The role of planting density according to the model is very simple. Leaf area index is simply multiplied by the number of plants per unit area. No account is taken of reduced growth per plant due to competition at high densities. This could lead to overestimating leaf area index at high densities. The graph of LAI residuals *versus* density (not shown) did not show such a trend.

4.3.3. Temperature

We defined two summary variables, namely calculated days from sowing to flowering and calculated days from flowering to maturity. These variables measure approximately average temperature over each period. Figure 8 shows yield residuals as a function of days from sowing to flowering. There is no clear trend in the residuals. There was no clear trend as a function of days from flowering to maturity either (not shown).

4.3.4. Soil moisture

The effect of soil moisture is of major importance, but analyzing the residuals as a function of soil moisture poses two problems. First, we have very few measurements of soil

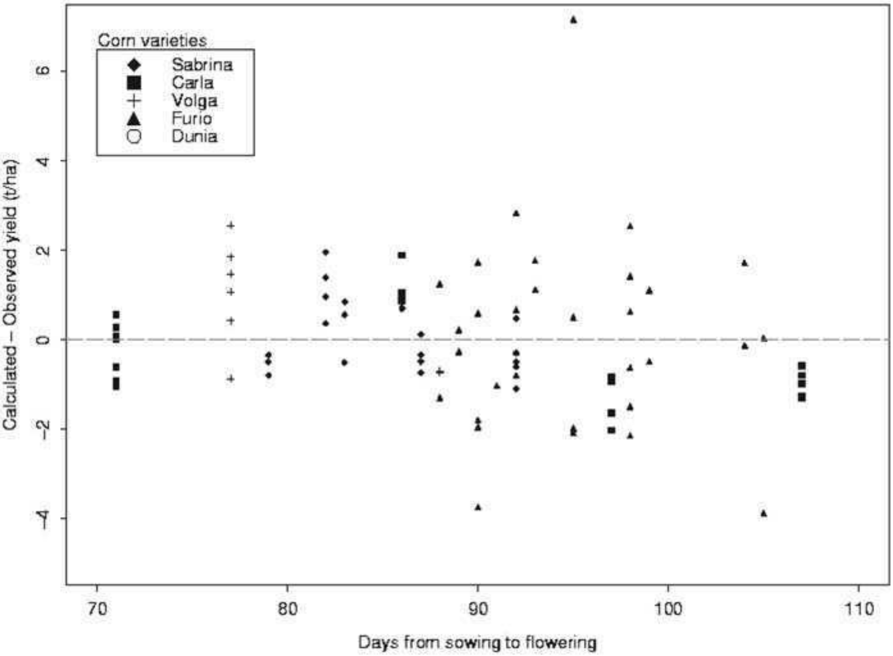


Figure 8. Yield residuals *versus* days from sowing to flowering.

moisture. Second, as for temperature, we are not interested in the effects of soil moisture for each day. We therefore defined a summary stress index, defined as the ratio of final biomass to final biomass in the absence of water stress, using calculated values for both terms. A stress index of 1 corresponds to no stress, and a value of 0 to no growth.

Figure 9 shows observed yield as a function of the stress index. As expected, yield clearly increases as the stress index increases (i.e. as the effect of stress decreases). However, the yield residuals showed no obvious trend with stress index (graph not shown).

5. Conclusions

Here, as in most crop models, there is not enough data to permit adjustment of all the model parameters. (This despite the major effort in collecting past data, and the relative simplicity of the model.) There is the problem of deciding which and how many parameters to adjust, the danger being that adjusting too many parameters will lead to a model with poor predictive ability. We based the choice of the number of parameters to estimate on the estimated prediction error for each number of parameters. This illustrates clearly the close connection between evaluation and parameter estimation. Parameter estimation should be based, as far as possible, on the criteria that are important for judging model quality.

We then evaluated the agreement between the adjusted model and the data. The objective was not only to judge the quality of the model, but also to reveal specific types of

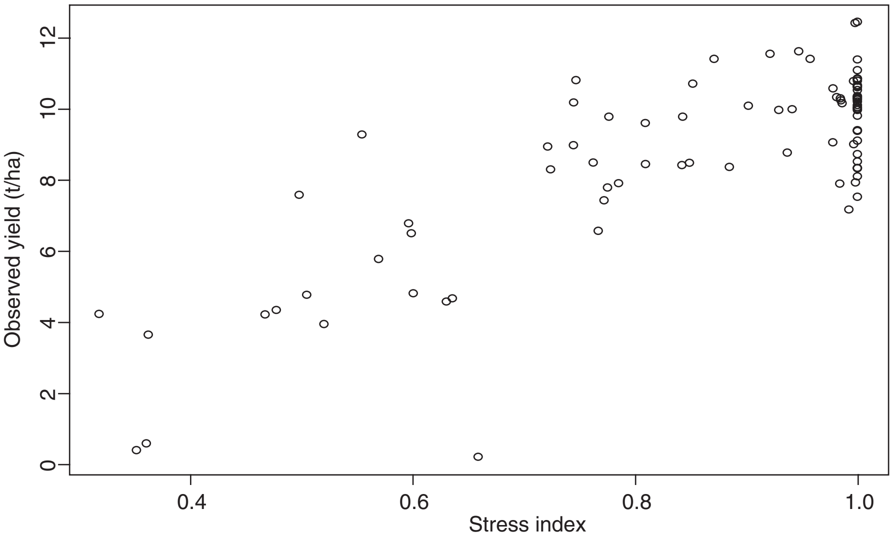


Figure 9. Observed yield *versus* stress index.

error that could be targeted in the next round of model improvement. The major conclusion here is that there is a problem in the way harvest index is modeled. The model predicts very little variability in harvest index compared to the observed variability. That is, it seems that harvest index is not sufficiently sensitive to conditions, and in particular to water stress. It should be noted that the necessary correction is more fundamental than simply adjusting a parameter, because the major parameter that determines the effect of water stress on harvest index was among the adjusted parameters.

The analysis here shows that it is worthwhile to investigate the discrepancies between model predictions and the data in many different ways since each representation brings out a different aspect of model quality.

Reference

Wallach, D., Goffinet, B., Bergez, J.-E., Debaeke, P., Leenhardt, D., Aubertot, J.-N., 2001. Parameter estimation for crop models: a new approach and application to a corn model. Agronomy Journal 93, 757–766.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 13

Evaluation of a model for kiwifruit

F. Lescourret and D. Wallach

1. **Introduction** **349**
2. **A model for kiwifruit** **349**
3. **Qualitative evaluation** **350**
4. **Evaluation of the predictions of fruit size distribution** **351**
5. **Conclusions** **355**
References **356**

1. Introduction

Kiwifruit is a fruiting perennial vine that only crops on new growth originating from 1-year-old stems called canes (Doyle et al., 1989). The plant is dioecious and so plots have both male and female plants. The number of canes is adjusted by winter pruning to get about 20–30 canes per female plant, the corresponding expected fruit number being 700–1000 (Agostini, 1995). Numerous studies have reported a positive relationship between seed number (up to 2000) and fruit size at harvest, which is thought to result from the promotion of fruit cell growth by some seed-elicited hormonal factors (Hopping, 1976). A major question is how to plant and then manage a plot of kiwifruit in order to maximize profit. Among the major management options are planting geometry, pruning of the vines and fruit thinning. Profit is not simply related to total yield, because the value of the crop also depends on the sizes of the individual fruits. It is thus important to predict how the distribution of fruit size depends on management. This point does not only concern kiwifruit but all the fruit crops where a fresh individual product is sold. It leads to models that are very different from models of annual crops, and the nature of the outputs leads to different requirements for model evaluation.

In the following section first we very briefly describe the model. Then we describe how the model was evaluated. This involved two different studies. The first involved a qualitative evaluation of the model, where various aspects of the behavior of the model were compared with results from the literature. In the second study the distribution of fruit sizes was compared quantitatively with experimental data.

2. A model for kiwifruit

The model for kiwifruit orchards used here is composed of three sub-models (Lescourret et al., 1999). The first describes flowering (Agostini et al., 1999). The initial number of buds is generated by drawing at random from a distribution function which is chosen to mimic the range of observed values. The number of buds that develop and the number of flowers that abort are also obtained by drawing from appropriate distribution functions at random. Finally, the distribution of flowering over time is also treated as random. Temperature, variety, size (the number of new canes and cane length) and thinning at the flowering stage, if applied, are taken into account in this sub-model. For the female vines, the flowering sub-model operates at the cane level and the results are summed up at the vine level. For the male vines, the flowering sub-model operates at the vine level. The outputs of the flowering sub-model, which are inputs of the second sub-model described below, are the number of flowers that open per day on each vine, i.e. that are ready to receive (on female vines) or to shed pollen (on male vines).

The second sub-model describes the pollination and fertilization of flowers and fruit setting (Lescourret et al., 1998a,b, 1999). It operates at the flower level. The number of pollen grains that are deposited on the stigmas of female flowers, the fraction that are viable, the number of ovules that are fertilized and the fraction of fruits that are set are all treated as random variables with appropriate distributions. This sub-model takes into account the effect of rainfall as well as planting options (proximity of males and females, distances between and within rows, male:female ratio, male varieties). The outputs are the number of seeds in each fruit of each vine of the orchard.

The final sub-model, which operates at the fruit level, is based on a deterministic description of fruit growth as a function of the number of seeds (predicted by the second sub-model), the fruit load of the vine (modulated by thinning) and the level of water stress which is related to climate and irrigation.

The effect of including randomness in the model is twofold. First of all, different flowers have different numbers of ovules that are fertilized leading to a distribution in fruit size. As already noted this is important since the value of the crop depends on the fruit size distribution. The second effect of randomness is that each simulation run with the model gives a different result. However, as we show, the differences between different model runs are relatively slight. This allows us to present simulated results for just a single simulation, unless noted otherwise.

3. Qualitative evaluation

There are experiments reported in the literature on the effect of various management decisions on kiwifruit orchards. In general, the information is not sufficiently complete to allow one to run the model and make quantitative comparisons with the data. Often the results are presented in the form of graphs or figures. However, one can compare model behavior and reported results qualitatively. This was done in the early stages of our modeling project. Note that in these studies some of the model parameters had to be estimated based simply on intelligent guessing.

The first example of qualitative evaluation is based on the study by Testolin (1991), who examined the consequences of an extreme planting design wherein there is a single male vine fertilizing the flowers. The results are presented in the form of a three-dimensional graph of the numbers of seeds per vine. The output of our pollination/fertilization sub-model, shows similar behavior (Fig. 1). This is in particular evidence in favor of the

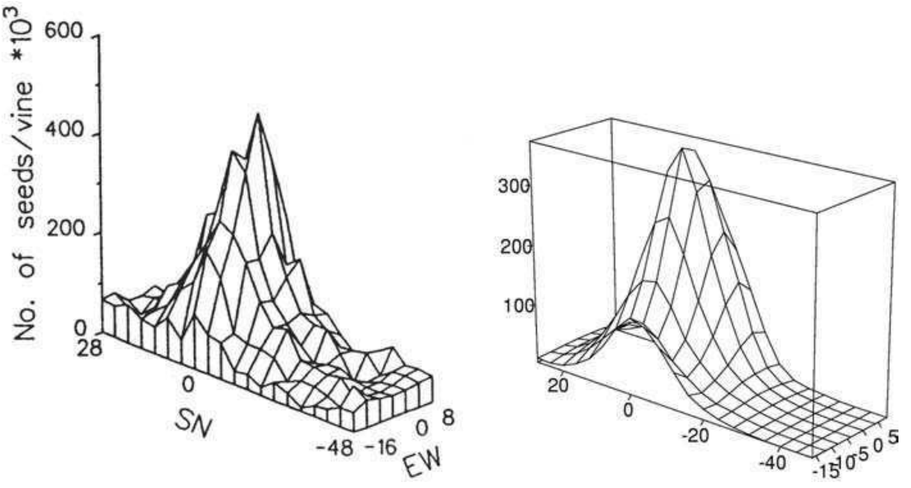


Figure 1. The total number of seeds per vine as a function of distance from a single male vine. Experimental results of Testolin (1991, left) and results from one simulation run of the pollination sub-model (right).

hypothesis in the model that the number of pollen grains per female flower declines exponentially with the distance between the male vine (source) and female vine (target).

The second example of qualitative evaluation is based on the results obtained by Antognozzi et al. (1991) and Lahav et al. (1989), who studied the effects of different modalities of thinning (no thinning, thinning at the flower bud stage or thinning at the set fruit stage) on mean fruit size per vine at harvest. Information on the number of seeds per fruit and on the intensity and date of thinning was either imprecise or lacking. Corresponding values plus parameters were estimated based on intelligent guessing (see above) to run the model. Comparison of the observations made by these authors and our simulations (Fig. 2) show good agreement between data on thinning at the flower bud stage or at the set fruit stage.

4. Evaluation of the predictions of fruit size distribution

The objective here is to compare observed and calculated distributions of fruit sizes. The experimental data come from the harvest of a plot in a kiwifruit orchard in Corsica, in 1995 (Lescourret et al., 1999). The original data give the number of harvested fruit in each of 9 weight categories. Figure 3 shows observed and simulated numbers of fruit in each of 4 weight categories based on European Union (EU) standards. The weight limits of the categories are presented in Table 1.

It is possible to compare separately, for each category, the simulated and observed numbers or fractions of harvested fruits. However, this is not convenient or easy to analyze, especially for larger numbers of categories. It is of interest then to summarize the

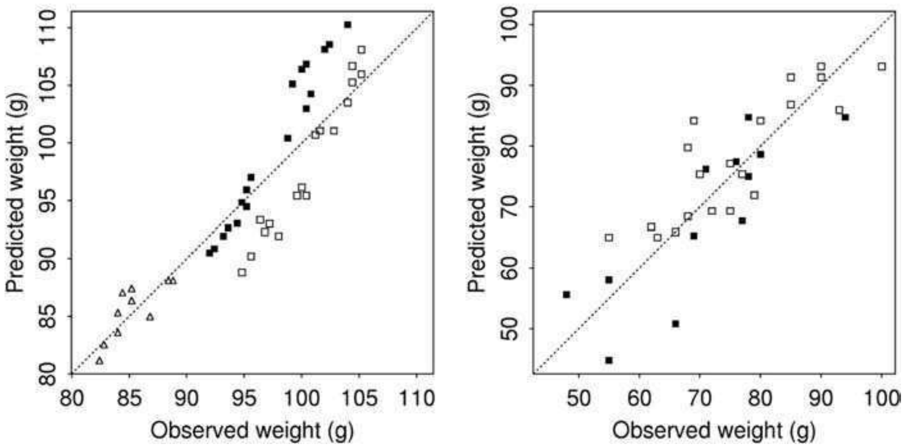


Figure 2. Simulated (by the fruit growth sub-model) *versus* observed average weight of fruit per vine at harvest. On the left, experimental results of Antognozzi et al. (1991). On the right experimental results of Lahav et al. (1989). △ – unthinned vines; □ – vines thinned at the flower bud stage and ■ – vines thinned at setting.

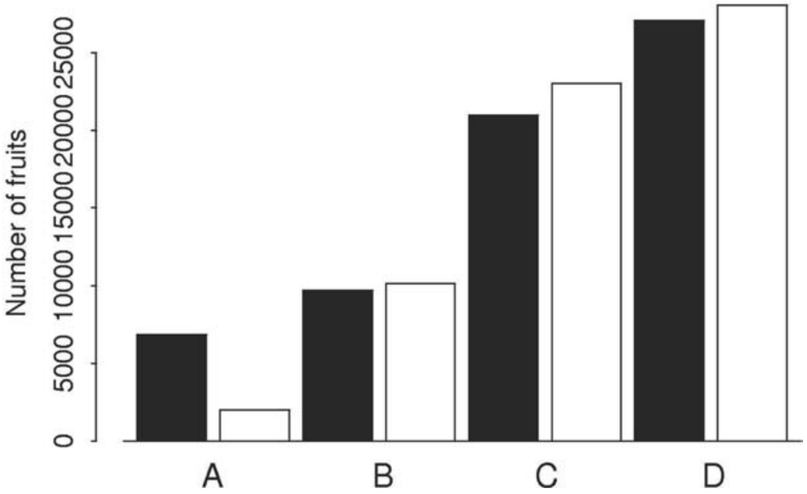


Figure 3. Number of fruit in each of 4 weight categories. Black bars represent observed values, white bars the simulated values from one run of the model. The weight limits for each category are given in Table 1.

comparisons with a single number, which represents a distance between the observed and simulated distributions.

A simple approach is to calculate a summary variable which is some weighted sum or average of the number of fruit per category, and then to compare observed and calculated values of this new variable. This comparison then simply involves comparing two numbers, and the distance can be the difference between them.

Table 1. Upper weight limits for fruit size categories.

Number of categories	Upper weight limits for each category		
	EU categories II and III	EU category I	EU category "Extra quality"
3	$z_1 = 70$ g	$z_2 = 90$ g	$z_3 =$ max weight
4	$z_1 = 70$ g	$z_2 = 90$ g	$z_3 = 105$ g $z_4 =$ max weight
5	$z_1 = 70$ g	$z_2 = 90$ g	$z_3 = 105$ g $z_4 = 125$ g $z_5 =$ max weight
9	$z_1 = 70$ $z_2 = 80$	$z_3 = 90$	$z_4 = 95$ $z_5 = 105$ $z_6 = 115$ $z_7 = 125$ $z_8 = 140$ $z_9 =$ max weight

In our case, a summary variable of particular interest is the monetary value of the kiwifruit crop, which is given by

$$VC = \sum N_i\ V_i$$

where VC is the total value of the crop, N_i is the number of fruit in weight category i, V_i is the value per fruit in category i and the sum is over categories. The observed and calculated values of VC based on the results shown in Figure 3 are $VC^{\text{obs}} = 3490€$ and $VC^{\text{sim}} = 3493€$. The difference is a negligible 3€. These values are based on average prices paid to fruit growers over the years 1991–2001 (data from the Kiwifruit Producers Organization communicated by J.M. Fournier).

A second distance that more directly reflects the differences between the observed and simulated distributions is the Kolmogorov–Smirnov statistic presented in the chapter "Evaluating crop models." Here we adapt this statistic to our specific case. We first define

$$S_{\text{obs},i} = \frac{\textit{total observed number of harvested fruits in categories 1},\ldots i)}{n_{\text{obs}}}$$

$$S_{\text{sim},i} = \frac{\textit{total simulated number of harvested fruits in categories 1},\ldots, i)}{n_{\text{sim}}}$$

We then define the step cumulative distribution function as $S_{\text{obs}}(z) = S_{\text{obs},\,i}$, where z is harvested fruit weight and i is the weight class to which the weight z corresponds. For example, if the weight z places a fruit in Category 2, then $S_{\text{obs}}(z)$ is the sum of the simulated numbers of harvested fruits in Categories 1 and 2 divided by the total simulated number of harvested fruits. $S_{\text{sim}}(z)$ is defined similarly but using simulated values. Finally the statistic ks is defined as

$$ks = \max_z |S_{\text{obs}}(z) - S_{\text{sim}}(z)| = \max_i |S_{\text{obs},i} - S_{\text{sim},i}|$$

(Sprent, 1992). The Splus® function cdf.compare can be used to compare the distribution functions. The calculation of the *ks* statistic here is very easy since few categories are involved.

Figure 4 shows the observed and calculated step cumulative distribution functions for various numbers of categories. For all numbers of categories the two distribution functions seem quite similar, but the details of the comparison and the value of the *ks* statistic depend on the number of categories. The largest difference between the observed and calculated distribution functions occurs at Category 1 if only 3 Categories are defined, but occurs at Category 6 in the case of 9 categories.

Using a probabilistic model, such as that employed here, each simulation produces a different result. To examine the variability of the results, we did multiple simulations with the pollination/fertilization sub-model, using only a single run of the flowering sub-model. The upper panel of Figure 5 shows the resulting number of harvested fruit in each of 4 weight categories, for 100 simulations. As can be seen, the variability is quite limited. The lower histogram in Figure 5 shows the distribution of the *ks* statistic for the 100 simulations. The mean value is 0.08 and the standard deviation is 0.004, giving a coefficient of variation of 5%. The distribution is slightly skewed. Simulations employing

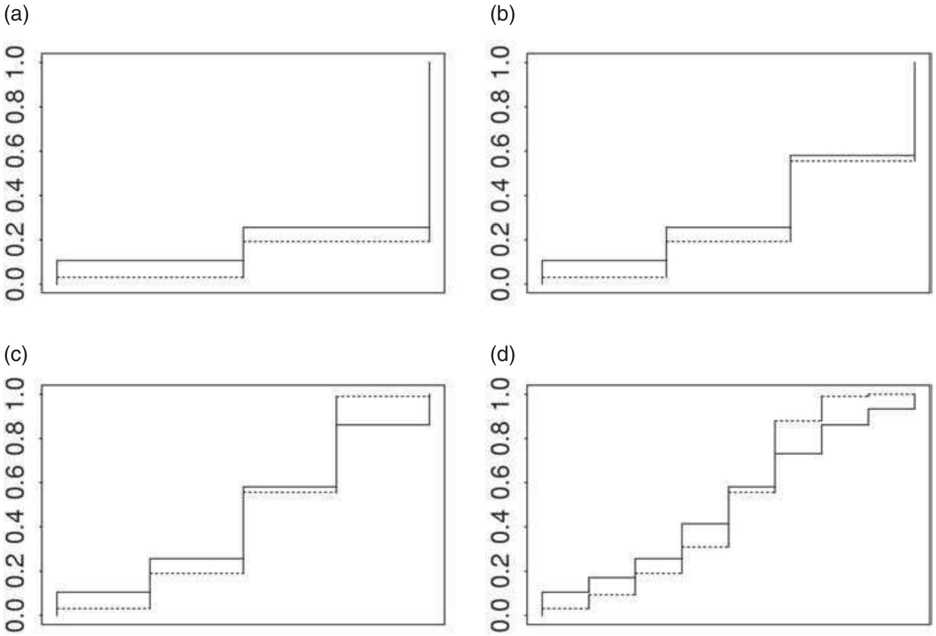


Figure 4. Values of the observed and calculated step distribution functions $S_{\mathrm{obs}}(z)$ and $S_{\mathrm{sim}}(z)$ for 3, 4, 5 or 9 weight categories (a–d, with $ks = 0.07$, 0.07, 0.13 and 0.15, respectively).

the random aspect of the flowering sub-model produced similar results. This suggests that we can meaningfully make comparisons using just one model run, since other runs will not be very different.

5. Conclusions

This chapter shows that model evaluation should be adapted to the objectives and nature of the model. The model used here was a good case since features like randomness and the type of outputs – distribution of individual products amongst categories – are unusual compared to classical crop models. The techniques used here are part of what Balci (1994) called VV&T techniques, which means Validation (building the right model), Verification (building the model right) and Testing (performing validation and verification). VV&T is a continuous activity throughout the entire cycle of a simulation study (Balci), and it is a key point because credibility of simulation results is especially important in models dealing with agricultural management.

Acknowledgments

The authors would like to thank Mr. J.M. Fournier at the BIK (Bureau Interprofessionnel du kiwi, the kiwifruit producers organization), who kindly supplied price charts for French-produced kiwifruit.

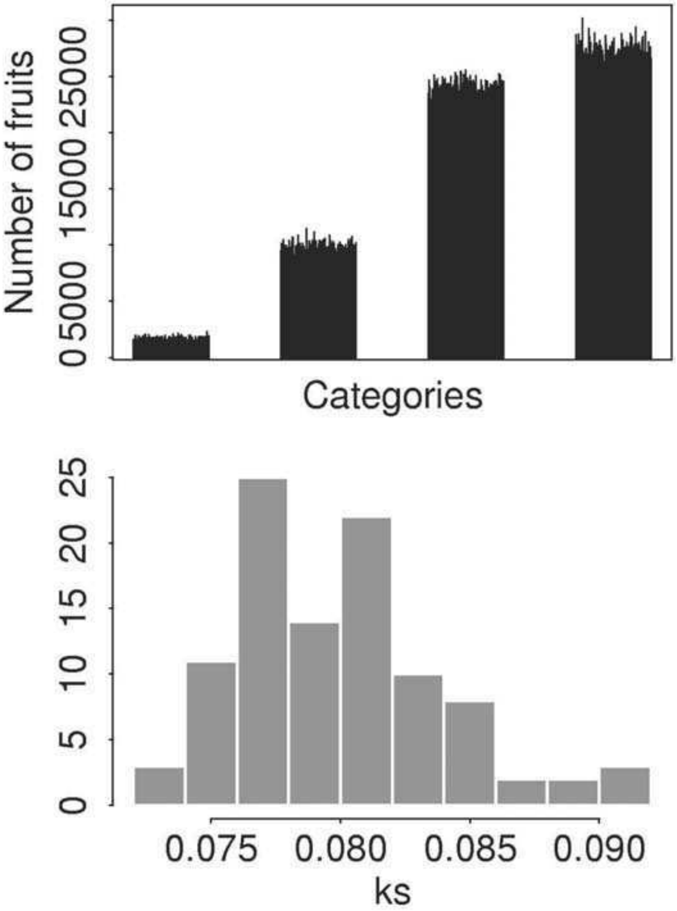


Figure 5. Classification of 100 simulated harvests into four size categories (cf. Figure 2). Each vertical line in a black bar represents the number of fruits in the corresponding category according to one simulation (above). A histogram of the *ks* values for the 100 simulations is in the lower figure.

References

Agostini, D., 1995. La floraison du kiwi (Actinidia deliciosa cv. Hayward): analyse de la variabilité et simulation par un modèle stochastique. Ph.D. Thesis, Lyon, France.

Agostini, D., Habib, R., Chadoeuf, J., 1999. A stochastic approach for a model of flowering in kiwifruit 'Hayward'. Journal of Horticultural Science and Biotechnology 74, 30–38.

Antognozzi, E., Tombesi, A., Ferranti, F., Frenguelli, G., 1991. Influence of sink competition on peduncle histogenesis in kiwifruit. New Zealand Journal of Crop and Horticultural Science 19, 433–439.

Balci, O., 1994. Validation, verification, and testing techniques throughout the life cycle of a simulation study. Annals of Operation Research 53, 121–173.

Doyle, C.J., Moore, W.B., Henzell, R.F., 1989. Modelling the economic consequences of potential management changes in a mature kiwifruit orchard in New Zealand. Agricultural Systems 31, 321–347.

Hopping, M.E., 1976. Effect of exogenous auxins, gibberellins, and cytokinins on fruit development in chinese gooseberry (*Actinidia chinensis Planch*). New Zealand Journal of Botany 14, 69–75.

Lahav, E., Korkin, A., Adar, G., 1989. Thinning stage influences fruit size and yield in kiwifruit. HortScience 24, 438–440.

Lescourret, F., Habib, R., Génard, M., Agostini, D., Chadoeuf, J., 1998a. Pollination and fruit growth models for studying the management of kiwifruit orchards. I. Models description. Agricultural Systems 56, 67–89.

Lescourret, F., Génard, M., Habib, R., Pailly, O., 1998b. Pollination and fruit growth models for studying the management of kiwifruit orchards. II. Models behaviour. Agricultural Systems 56, 91–123.

Lescourret, F., Blecher, N., Habib, R., Chadoeuf, J., Agostini, D., Pailly, O., Vaissière, B., Poggi, I., 1999. Development of a simulation model for studying kiwifruit orchard management. Agricultural Systems 59, 215–239.

Sprent, P., 1992. Pratique des statistiques non paramétriques. INRA Editions, Paris.

Testolin, R., 1991. Male density and arrangement in kiwifruit orchards. Scientia Horticulturae 48, 41–50.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 14

Sensitivity and uncertainty analysis of a static denitrification model

B. Gabrielle

1. **Introduction** **359**
2. **Materials and methods** **360**
3. **Results** **362**
4. **Conclusion** **365**

References **365**

1. Introduction

Denitrification is the process in the global nitrogen cycle which transforms soil and aquatic nitrate to the gases N_2 and N_2O, which are then returned to the atmosphere. It is a major pathway for atmospheric emissions of N_2O, a greenhouse gas, from terrestrial agroecosystems. The literature reports much data on gaseous flux measurements in various situations, and particularly in agricultural systems. However, there have been relatively few attempts to model the process, primarily because of the large variability in the experimental data at the field scale. In addition, these models require site-specific parameters due to the fact that soil structure and aggregate characteristics exert a major influence on denitrification. These parameters are usually estimated indirectly using laboratory incubation studies (Hénault and Germon, 2000). They may alternatively be estimated on the basis of measured denitrification data, by fitting simulated data to field observations, or set to default values.

As is the case for any model, there are thus uncertainties associated with the parameters of NEMIS, due to the variance of the parameter estimators for a given site. This variance may, for instance, be estimated when fitting model parameters to observed data. Another source of uncertainty, more specific to crop models, and truer still of the NEMIS model, is that parameters vary widely across agricultural fields (Hénault and Germon, 2000). A final source of uncertainty lies in the input data known to control denitrification (soil temperature, moisture, carbon and nitrogen content), which are spatially variable within an agricultural field. There are thus two sources of uncertainty that should be dealt with while simulating a new field situation: the uncertainty due to parameter variance, and that due the spatial variability of model input variables within the particular field selected. Secondly, the variation of parameter values across fields should also be addressed, for instance by analyzing a range of fields representative of the area under study.

In this example, I show how some of the methods described in Chapter 3 can be used to assess the influence of these sources of uncertainty on model outputs, at the field-scale, and across two particular field conditions. The objective is to identify the soil input variables or parameters which should be determined with most accuracy while using a model to predict denitrification rates. Such results may be used to issue guidelines while designing an experiment to predict denitrification at the field-scale. This example also examines how the sensitivity results may vary across fields – i.e. whether the variables or parameters identified as most sensitive in a given field are likely to be sensitive in another field characterized by different pedoclimatic conditions.

The sensitivity analysis (SA) methods used here fall in the variance-based category. The uncertainty in model input data (e.g. soil moisture content) was estimated from within-field spatial variability studies, whereas within-field parameter uncertainty was estimated from bootstrap samples of experimental denitrification data sets. The rationale for using bootstrap *versus* a more ordinary Monte-Carlo sampling technique was that it made it possible to take correlations between parameters into account. Finally, I used data sets collected at two field sites to capture the variability in SA results across agricultural fields.

2. Materials and methods

2.1. The NEMIS denitrification model

The NEMIS denitrification model (Hénault and Germon, 2000) uses a semi-empirical equation to calculate daily denitrification flux from the topsoil ($\hat{Y}_{\mathrm{m}}^{S}$):

$$\begin{cases} Y_{\mathrm{m}}^{S} = PDR\left(\dfrac{SW/V_0 - w_0}{1 - w_0}\right)^{d} & \text{if } (SW/V_0 - w_0) > 0 \text{ and } T > 0 \\ \qquad \times \dfrac{[NO_3^-]}{K_{\mathrm{N}} + [NO_3^-]} 10^{[A(T-20)]} & \\ Y_{\mathrm{m}}^{S} = 0 & \text{otherwise} \end{cases} \tag{1}$$

where SW is the soil volumetric moisture content (m^3 m^{-3}), $[NO_3^-]$ the nitrate concentration (mg N kg^{-1} soil), T the temperature (°C), and V_0 the soil porosity (m^3 m^{-3}). Equation (1) expresses the denitrification rate as the product of a potential rate, *PDR* (kg N ha^{-1} day^{-1}), with three unitless factors accounting for the response of denitrification to soil environmental conditions. The effect of temperature is modelled with a Q_{10} function ($Q_{10} = 10^{10A}$, with A in °C^{-1}). The response of denitrification to nitrate follows Michaelis–Menten kinetics, with half-saturation constant K_{N} (mg N kg^{-1}). The water content SW is taken as a predictor of the soil anoxic volume, with a power relationship parameterized by the unitless coefficient d. Parameter w_0 is the water-filled pore space threshold below which no denitrification occurs.

2.2. Experimental data

Two published denitrification data sets are used, against which NEMIS was initially tested (Hénault, 1993). They represent contrasting pedoclimatic and cropping conditions: a light clay soil managed as a fertilized grassland in Germany (Corré et al., 1990), and a barley-cropped clay loam with high organic C in Southern England (Webster et al., 1993). Denitrification rates were measured over a one year period on 22 (Webster et al. data set) to 46 (Corré et al. data set) dates. In both experiments, the authors incubated intact soil cores *in situ* or in the laboratory, and used the acetylene inhibition technique to measure denitrification. Model inputs (soil nitrate, water content and temperature) were also recorded. The measurements were made on 5–30 replicates, and presented coefficients of variation ranging from 20% to over 100%. Although very significant, this variability was not used in the uncertainty on input variables since only the natural within-field spatial variability was of concern in this study.

2.3. Sensitivity analysis

Figure 1 shows the sequence of steps in the sensitivity analysis (SA) undertaken with NEMIS. Prior to the global SA, a one-at-a-time analysis showed that, among the five

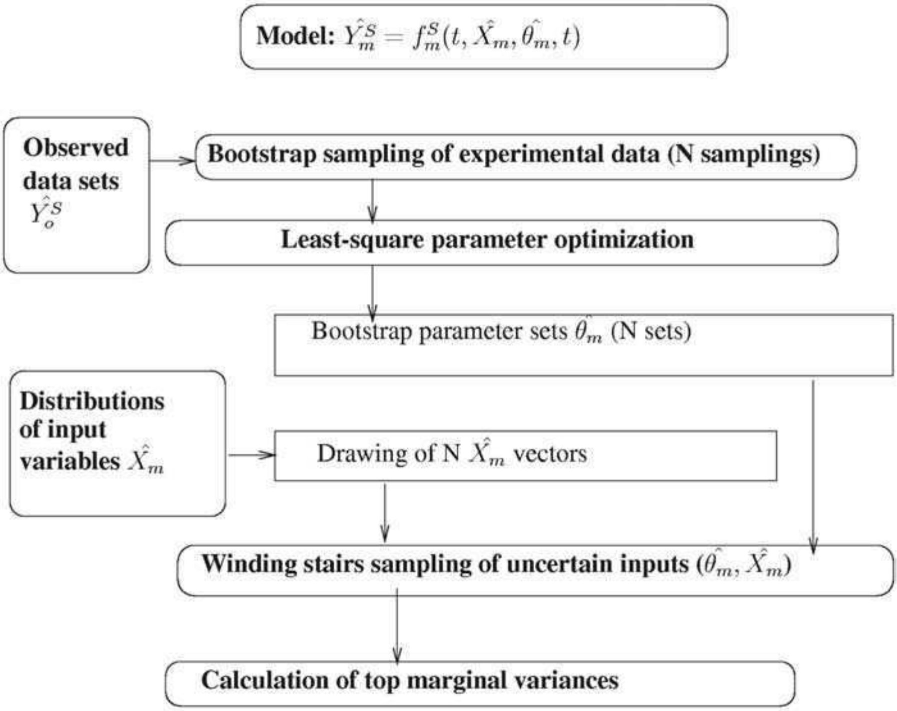

Model: $\hat{Y}_m^S = f_m^S(t, \hat{X}_m, \hat{\theta}_m, t)$
Observed data sets $\hat{Y}_o^S$
Bootstrap sampling of experimental data (N samplings)
Least-square parameter optimization
Bootstrap parameter sets $\hat{\theta}_m$ (N sets)
Distributions of input variables $\hat{X}_m$
Drawing of N $\hat{X}_m$ vectors
Winding stairs sampling of uncertain inputs $(\hat{\theta}_m, \hat{X}_m)$
Calculation of top marginal variances

Figure 1. Diagram of the uncertainty analysis of the NEMIS model.

parameters of Eq. (1), parameter d had very little influence. It was thus left at its nominal value of 1.74 throughout the analysis.

The first step of the SA then consisted of generating random samples of parameter values and input variables, whose source of uncertainty were parameter estimator variance and spatial variability within the field, respectively. First, we generated bootstrap estimates for the four model parameters considered, in the two experimental sites under study. For each site, 1000 Least-Square (L-S) estimates of model parameters were calculated with a constrained, Gauss–Newton-like algorithm (Numerical Algorithms Group, 1993), from bootstrap samples of the data drawn with a balanced method (Shao and Tu, 1995; p. 211).

Next, 1000 random values of each of NEMIS' three input variables (*SW*, *T* and NO_3^-) were drawn independently from Gaussian distributions. For soil temperature and water content, we used coefficients of variations of 5% and 10%, respectively (Boulier, 1985; Cellier et al., 1996). For nitrate concentrations, we used a CV of 20%, as reported by Cambardella et al. (1994). Finally, we assessed the respective contribution of the above sources of error to the variance of the simulated denitrification rates $\mathrm{Var}(Y_{\mathrm{m}}^{S})$ with the winding stairs method proposed by Jansen et al. (1994). The method is a modified Latin Hypercube sampling which makes it possible to estimate the firstorder sensitivity coefficient S_i (see Chapter 3) of model output Y to a particular input variable or parameter, denoted X_i. S_i is defined as:

$$S_i = \frac{\mathrm{Var}[E(Y_{\mathrm{m}}^{S}|X_i = x_i)]}{\mathrm{Var}(Y_{\mathrm{m}}^{S})}$$

S_i indicates the average influence of the input X_i on the model output.

3. Results

3.1. Boostrap parameter sets

Table 1 summarizes the univariate statistics of the model parameters and root mean squared errors (RMSE) derived from the bootstrap procedure for the two experimental data sets. The RMSE represented up to 90% of the mean observation, which indicates a weak ability of the model to explain the experimental data. Overall, the accuracy of the model ranged from 10 to 217 g N ha^{-1} day^{-1}, which is within the order of magnitude of the figures obtained by Johnsson et al. (1991) after calibrating a similar denitrification model on their data sets. Refer to Chapter 4 (Parameter estimation) for more details on these techniques.

Within each experimental field, parameter variance was moderate, with coefficients of variation (CVs) in the 10–30% range. The water threshold w_0 showed the least dispersion, as opposed to *PDR* which was the most variable parameter. An accurate identification of *PDR* is also likely to have been hampered because of its correlations with other parameters, as illustrated in Table 2. Such correlations between parameters indicate a poor resolution in the parameter space around the optimum set. Finally, the CVs associated with the *PDR* compare well with the 40–70% range reported by Hénault (1993) from laboratory measurements on a variety of soils.

Table 1. First moments of the bootstrap estimates of the NEMIS model parameters and root mean squared error (RMSE), for the 2 data sets. To remove tail effects, the statistics were calculated over the 2.5–97.5% percentile interval of the bootstrap distributions, whose boundaries are indicated in brackets. Below the RMSEs, the bracketed figures express the RMSEs as a percentage of the mean observed fluxes.

Data sets	Parameters (Means ±1 standard deviation)					Error
	N*	K_N (mg N/kg soil)	w_0 (%)	A ($\times 10^{-3}$ °C^{-1})	*PDR* (kg N ha^{-1} day^{-1})	RMSE
Webster et al. (1993)	22	5.00 ± 0.00	65.9 ± 2.3	49.2 ± 7.2	5.06 ± 0.90	0.010 ± 0.003
		(5.00–5.00)	(58.3–69.5)	(31.9–67.1)	(3.18–7.67)	(32.0)
Corré et al. (1990)	46	5.00 ± 1.16	50.0 ± 0.0	40.7 ± 9.3	7.82 ± 2.03	0.217 ± 0.060
		(5.00–10.04)	(50.0–50.0)	(31.9–62.7)	(4.19–10.00)	(87.5)

*Number of observations.

Table 2. Bootstrap estimates of the model parameters correlation matrix, for the two data sets.

Data sets		Correlation coefficients			
		K_N	w_0	*A*	*PDR*
Webster et al. (1993)	K_N	1.0	−0.062	−0.026	−0.217
	w_0		1.0	−0.789	0.892
	A			1.0	−0.526
	PDR				1.0
Corré et al. (1990)	K_N	1.0	−0.009	0.137	0.529
	w_0		1.0	0.049	0.073
	A			1.0	0.669
	PDR				1.0

Across the two experimental fields, the three model parameters considered as fixed by the model's author, whatever the environmental conditions (K_N, w_0 and A), differed only to a minor extent. In both cases the nitrate half-saturation constant K_N took the minimum value of 5 mg N kg^{-1} soil prescribed in the optimization, which implies that denitrification was little sensitive to nitrate for concentrations above this bottom value of 5 mg N kg^{-1} soil. The water-filled porosity threshold, w_0, ranged from its bottom value of 50% to a mid-range value of 66% for the experiment of Webster et al. (1993). Interestingly, this value is close to that of 62% reported by Rolston and Grundmann (1987) and Hénault (1993). The temperature response parameter A varied from 40 to 50 10^{-3} °C^{-1}, with the upper value being rather high for biological phenomena since it translates a Q_{10} value of 3.1. However, Stanford et al. (1975) reported Q_{10} values of 60–120 for temperatures below 11°C, because of an abrupt decrease of denitrification in this lower temperature range.

3.2. First-order sensitivities of various uncertainty sources

Figure 2 presents the relative contributions of the four identified sources of uncertainty to the variance of the simulated denitrification rates, for two of the data sets. Overall, the input variables contributed the greatest share in the total simulated variance, totalling 60–70%, with moisture content and temperature being most sensitive. The variability in parameters contributed less to the variance in output. Their share of total variance was around 20%, which is relatively small given that the individual parameters had rather greater coefficients of variation. The net contribution of the uncertainty in the parameter sets is then likely to have been lower than may have been expected for individual parameters because these were correlated to some extent.

The analysis of Figure 2 may point at strategies for reducing the uncertainty in the prediction of denitrification. Increasing the frequency of denitrification measurements at a given site could be expected to improve the L-S estimates of the model parameters, thereby reducing their uncertainty. On the other hand, taking more replicates may increase the accuracy on the measurements, and decrease the model's residual error. Finally, increasing the number of replicates while measuring the input variables would decrease their uncertainty. Figure 2 shows that this last option would be the most efficient, since about 65% of the variance in the simulated denitrification fluxes may be ascribed to the uncertainty

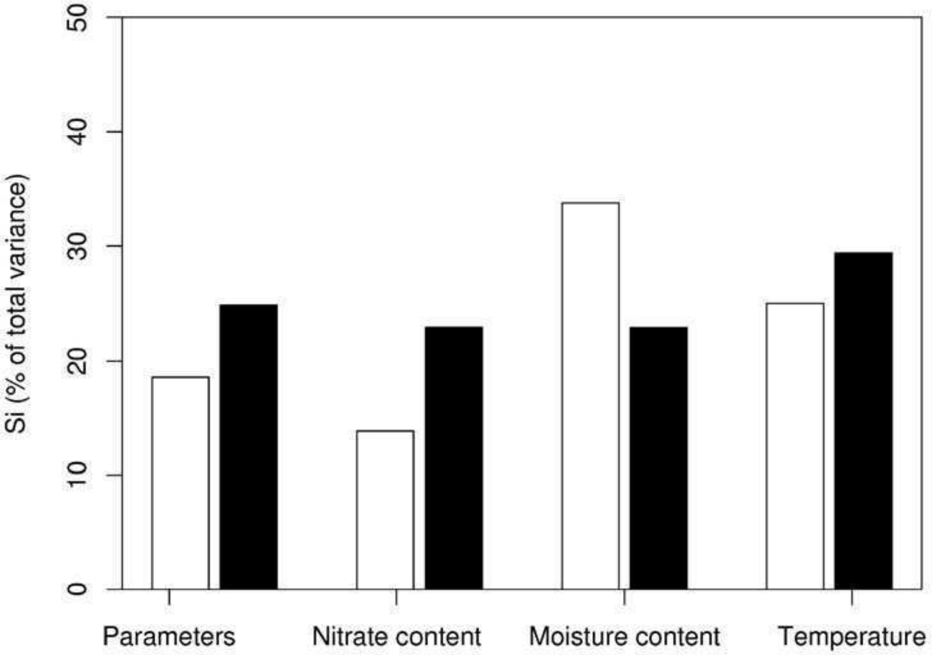


Figure 2. First-order sensitivity coefficients (S_i) of the four sources of uncertainty in the NEMIS denitrification model, for the data sets of Corré et al. (white bars) and Webster et al. (black bars). The sources comprise the three input variables (soil moisture, temperature, and nitrate content), and model parameters.

in the input variables. However, this strategy is also somewhat limited by the spatial variability inherent to these variables at the field scale, combined with the measurement errors. A better representation of the denitrification process at this scale may thus require a stochastic approach. Otherwise it seems essential to check the values of the parameters against actual measurements of field-denitrification, with a number of 20–40 observations allowing to reach a satisfactory accuracy of the model.

4. Conclusion

This example illustrated the use of variance-based sensitivity analysis, and showed how it could be applied to sort the various sources of uncertainty according to their effect on model outputs. This ranking may serve to guide future use and development of the model, for instance, by emphasizing the need to measure a particular input with sufficient accuracy.

Although somewhat complex, the method proposed here has the advantage of taking into account possible correlations between model parameters. This is definitely important since most crop models include a vast number of parameters compared to the number of outputs that can actually be tested against experimental data. This means that model parameters are frequently cross-correlated (as appeared here), and that ignoring this fact in the sensitivity analysis may result in misleading conclusions. The major drawback of this method is that it requires some parameters to be simultaneously optimized, a task which might be daunting with full-fledged, dynamic crop models. Traditional steepest-descent algorithms such as used here may indeed not always give good results (see Chapter 4 on Parameter estimation). Moreover, the information retrieved on model sensitivity to the set of parameters investigated might prove difficult to interpret since it no longer pertains to one single parameter but to the parameter set as a whole. Finally, as with all sensitivity analysis methods, the outcome depends on the experimental conditions tested, and on the values selected for the parameters not included in the analysis. However, in the study presented here it appeared that the patterns of the uncertainty analysis were similar across experimental sites.

Acknowledgments

I would like to acknowledge Catherine Hénault (INRA Dijon, France) for her guidance in the use of the NEMIS model, and C.P. Webster (IACR Rothamsted, UK) for editing suggestions on an earlier draft.

References

Boulier, J., 1985. Modélisation stochastique de l'infiltration en milieux poreux non uniformes; application à une micro parcelle irriguée. Ph.D. Thesis, Université Joseph Fourier, Grenoble.

Cambardella, C.A., Moorman, T.B., Novak, J.M., Parkin, T.B., Karen, D.L., Turco, R.F., Konopa, A.E., 1994. Field-scale variability of soil properties in Central Iowa soils. Soil Science Society American Journal 58, 1501–1511.

Cellier, P., Jacquet, A., Bautrais, P., Morlat, R., Delanchy, P., 1996. Modélisation du régime thermique des sols de vignoble du val de loire: relations avec des variables utilisables pour la caractérisation des terroirs. In: Colloque Les terroirs viticoles: concept, produit, valorisation, INRA-ISVV, Angers, pp. 107–112.

Corré, W.J., Dijkman, W., Sauerbeck, D., 1990. Denitrification of the top soil of production grassland. Mitteilungen der deutschen bodenkundlichen Gesellschaft 60, 183–188.

Hénault, C., 1993. Quantification de la dénitrification dans les sols à l'échelle de la parcelle cultivée, a l'aide d'un modèle prévisionnel. Ph.D. Thesis, Ecole Nationale Supérieure d'Agronomie, Montpellier.

Henault, C., Germon, J.C., 2000. NEMIS, a predictive model of denitrification on the field scale. European Journal of Soil Science 51, 257–270.

Jansen, M.J.W., Rossing, W.A.H., Daamen, R.A., 1994. Monte Carlo estimation of uncertainty contributions from several independent multivariate sources. In: Grasman, J., van Straten, G. (Eds), Predictability and Nonlinear Modelling in Natural Sciences and Economics, Kluwer Academic Publishing, Dordrecht, pp. 335–343.

Johnsson, H., Klemendtsson, L., Nilsson, A., Svensson, B.H., 1991. Simulation of field scale denitrification losses from soils with grass ley and barley. Plant and Soil 138, 287–302.

Numerical Algorithms Group, 1993. Minimizing or maximizing a function (E04JAF), volume Mark 16 of The NAG Fortran Library Manual, page ref. 1435/0. Oxford.

Rolston, D., Grundmann, G., 1987. A water function approximation to degree of anaerobiosis associated with denitrification. Soil Science 144, 437–441.

Shao, J., Tu, D., 1995. The Jacknife and Bootstrap. Springer Series in Statistics, New York.

Stanford, G., Dziena, S., Pol, R.A.V., 1975. Effect of temperature on denitrification rate in soil. Soil Sci. Soc. Am. Proc. 39, 867–870.

Webster, C.A., Shepherd, M.A., Goulding, K.W.T., Lord, E., 1993. Comparisons of methods for measuring the leaching of mineral N from arable land. Journal of Soil Science 44, 49–62.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 15

Sensitivity analysis of PASTIS, a model of nitrogen transport and transformation in the soil

P. Garnier

1. Introduction **367**
2. Model description **367**
3. Experimental data **371**
4. Sensitivity analysis **371**
5. Results **372**
6. Conclusion **374**
References **375**

1. Introduction

PASTIS (Prediction of Agricultural Solute Transformations In Soils) is a mechanistic model that simulates the movement of water, heat and solute in the soil as well as the transformations of soil carbon and nitrogen (Garnier et al., 2001, 2003). The model considers just one spatial dimension (depth) and simulates processes over a short period (a crop cycle). The model is used as a research tool.

This chapter presents the results of a sensitivity analysis of various model outputs to several of the model inputs. The objectives of the sensitivity analysis were threefold. The first objective was to identify the model parameters to which the outputs are particularly sensitive. It is important to determine the values of these parameters accurately. The second objective was to determine the processes to which the outputs are particularly sensitive in order to better understand the functioning of the system. The third objective concerned the fit of the model to experimental data. Sensitivity analysis in general is purely model based. It explores how model outputs vary when model inputs are changed. However, it is also of interest to see how model agreement with data varies when model inputs are varied.

The following section describes the model. Section 3 describes the experimental situation to which the sensitivity analysis is applied. Section 4 defines the sensitivity indices that are calculated, Section 5 presents results and Section 6 conclusions.

2. Model description

PASTIS consists of two submodels, a transformation submodel named CANTIS (Carbon And Nitrogen Transformations In Soil) and a transport submodel. The transformation submodel calculates the amount of mineralized nitrate in the soil as a function of temperature and water potential which are calculated by the transport submodel. The transport submodel calculates the nitrate transported by convection and dispersion. The model state variables, input variables, output variables and parameters are listed in Table 1.

Table 1. Model input variables, state variables, output variables and parameters.

Input variables	Initial conditions	*$NO^-_{3,\mathrm{INI}}(z)$: initial nitrate content (kg N/ha) $\theta_{\mathrm{INI}}(z)$: initial water content (cm^3/cm^3) $C_{\mathrm{INI},i}$: initial carbon content in pool i (kg C/ha) $C{:}N_{\mathrm{INI},i}$: initial C:N ratio in pool i
	Boundary conditions	*$PET(t)$: potential evapotranspiration (cm/h) $R(t)$: rain (cm) $h_{\mathrm{B}}(t)$: water pressure head at bottom of profile (cm) $T_{\mathrm{B}}\ (t)$: temperature at bottom of profile $T_{\mathrm{T}}(t)$: temperature at soil surface
State variables		$C_i\ (z,t)$: carbon content of the organic matter pool i, $h(z,t)$: matric potential (cm) $\theta(z,t)$: water content (cm^3/cm^3) $K(z,t)$: hydraulic conductivity (cm/h) $T(z,t)$: soil temperature (K) $S(z,t)$: solute concentration (kg m^{-3})
Input parameters	Biological parameters	*k_i: decomposition rate of the organic matter pool i under reference conditions *B_{T}: temperature factor coefficient T_{ref}: 15 °C h': water potential at which microbial activity ceases (−75 800 cm of water) h_{ref}: −100 cm of water K_{MZ}: parameter of the biomass-dependent function
	Physical parameters	C_{h}: capillary capacity (m^{-1}) *a_0: first coefficient for hydraulic conductivity function a_i: other coefficients for hydraulic conductivity function $\theta_{\mathrm{s}}, \theta_{\mathrm{r}}$: saturated and residual water content (cm^{3}cm^{-3}) α, n, m: coefficients for water retention function D: dispersion coefficient (m^{2}s^{-1}). D': dispersivity coefficient (m^2 s^{-1}) q: Darcy flux (m s^{-1}) *λ'_{T}: thermal conductivity coefficient λ_{T}: thermal conductivity (W m^{-1} K^{-1}) C_{T}: volumetric thermal capacity of soil (J m^{-3} K^{-1}) $\mathrm{C_w}$: volumetric thermal capacity of water (J m^{-3} K^{-1})

continued

Table 1.—Cont'd

Output variables	**$\hat{Y}_{\mathrm{LN}}$: nitrogen leached below 150 cm (1 year)
	$\hat{Y}_{NO_3}(z,t)$: nitrate content
	$\hat{Y}_{\mathrm{CO_2}}(t)$: mineralized CO2
	$\hat{Y}_{\mathrm{GNM}}(t)$: gross N mineralization
	$\hat{Y}_h(z,t)$: water pressure head (cm)
	$\hat{Y}_\theta(z,t)$: volumetric water content (cm^3/cm^3)
	$\hat{Y}_{\mathrm{T}}(z,t)$: temperature (ºC)

*indicates a factor in the sensitivity analysis.
**indicates an output in the sensitivity analysis.

2.1. The transformation submodel

The CANTIS model simulates the carbon and nitrogen cycles (Fig. 1). The processes that are modeled are decomposition of organic matter, mineralization, immobilization, nitrification and humification. Soil organic matter is divided into five main organic pools. The microbial population is split into an autochtonous biomass (AUB) that decomposes the humified organic matter (HOM), and a zymogenous biomass (ZYB) that decomposes fresh (FOM) and soluble (SOL) organic matter. The FOM pool is composed of four biochemical fractions, namely, rapidly decomposable material (RDM), hemicelluloses (HCE), cellulose (CEL) and lignin (LIG).

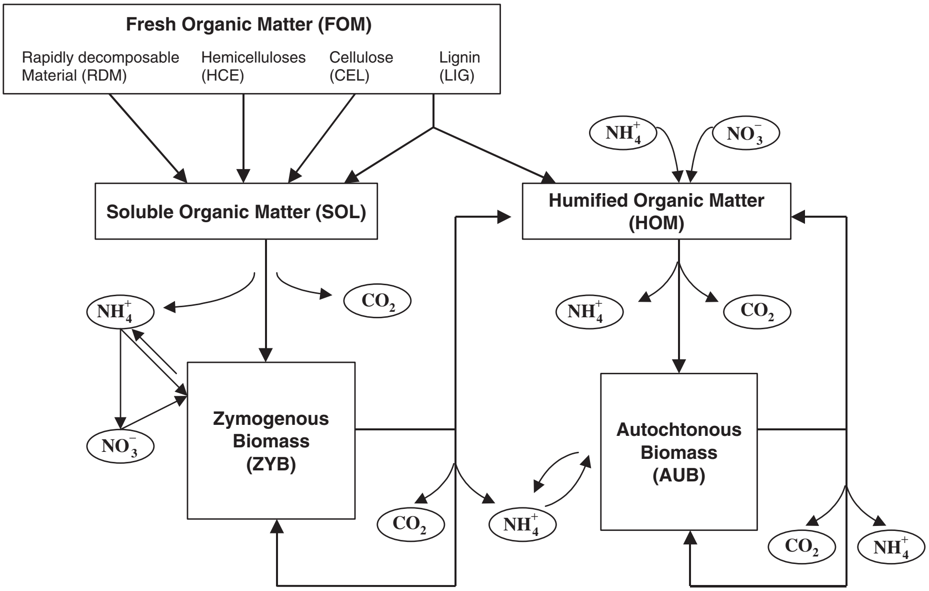


Figure 1. Flow diagram of the CANTIS model.

Table 2. Equations for decomposition limiting factors.

Factors	Equation	References
Temperature limitation factor	$f_{\mathrm{T}} = \exp^{B_{\mathrm{T}}(T - T_{\mathrm{ref}})}$	Rodrigo et al. (1997)
Moisture limitation factor	$f_{\mathrm{W}} = \dfrac{\ln(h/h')}{\ln(h_{\mathrm{ref}}/h')} \quad h' < h < h_{\mathrm{ref}}$ $f_{\mathrm{W}} = 1 \quad h \geq h_{\mathrm{ref}}$ $f_{\mathrm{W}} = 0 \quad h \leq h'$	Andrén et al. (1992)
Contact limitation factor related to zymogenous biomass	$f_{\mathrm{B}} = \dfrac{B_Z}{K_{\mathrm{MZ}} + B_Z}$	Hadas et al. (1998)

Decomposition of fresh or soluble organic matter is assumed to follow first-order kinetics relative to microbial biomass size:

$$\frac{dC_i}{dt} = -k_i C_i f_{\mathrm{T}} f_{\mathrm{W}} f_{\mathrm{B}} f_{\mathrm{N}}$$

where C_i is carbon content of organic matter pool i, k_i is the decomposition rate under reference conditions and f_{T}, f_{W}, f_{B} and f_{N} are factors related respectively to temperature, water, zymogenous biomass, and nitrate. Expressions for the first three of those factors are given in Table 2. The factor f_{N} is equal to 1 when the amount of mineral nitrogen available is sufficient for microbial needs, and less than 1 when the available nitrogen limits decomposition rate (Recous et al., 1995).

2.2. The transport submodel

The transport equations are given in Table 3. Water flow is described using Richards' equation. The classical convection–dispersion equation is used to simulate solute movement. Heat transport is described using the convection–diffusion equation.

Table 3. Transport equations. t is time (s), z is depth (m).

	Equations	Simulated variable	Functions necessary to solve the equations
Water flow	$C_h \dfrac{\partial h}{\partial t} = \dfrac{\partial}{\partial z}\left[K\left(\dfrac{\partial h}{\partial z} - 1\right)\right]$	h	$h(\theta) = \dfrac{1}{\alpha}\left[\left(\dfrac{\theta - \theta_r}{\theta_s - \theta_r}\right)^{-1/m} - 1\right]^{1/n}$ $K(\theta) = a_0 + a_1\theta + a_2\theta^2 + a_3\theta^3$
Solute movement	$\dfrac{\partial(\theta S)}{\partial t} = \dfrac{\partial}{\partial z}(\theta D \dfrac{\partial S}{\partial z} - qS)$	S	$D(\theta) = D\theta$
Heat transport	$C_{\mathrm{T}} \dfrac{\partial T}{\partial t} = \dfrac{\partial}{\partial z}(\lambda_{\mathrm{T}} \dfrac{\partial T}{\partial z} - qC_{\mathrm{w}}T)$	T	$\lambda_{\mathrm{T}}(\theta) = \lambda'_{\mathrm{T}}\theta$

3. Experimental data

An experiment was carried out from October 1993 to September 1994 in a bare field with loamy soil located at Mons-en-Chaussée in Northern France. The absence of a crop allowed more precise estimates of the C and N pools than would be possible otherwise.

The input variables listed in Table 1 were measured. These include rainfall (R) and potential evapotranspiration (PET), temperature at the soil surface (T_{T}) and at a depth of 1.5 m (T_{B}), and matric potential at 1.5 m (h_{B}). Measurements were made every hour. Soil samples were collected to measure initial water content (θ_{INI}) and initial mineral nitrogen ($NO^{-}_{3,\mathrm{INI}}$) every 20 cm from 0 to 1.5 m. The amount of initial carbon (C_i^{INI}) and the initial C:N ratio ($C : N_i^{\mathrm{INI}}$) were measured in the pools of fresh organic matter, humified organic matter and microbial biomass.

Measurements of the output variables of Table 1 were also made (every hour each day). Our experimental field was equipped to measure matric potential Y_h, volumetric water content Y_θ and temperature Y_{T} from the surface to a depth of 1.5 m in increments of 20 cm. Also nitrate content ($Y_{\mathrm{NO_3}}$) every 20 cm and CO_2 flux at the soil surface ($Y_{\mathrm{CO_2}}$) were measured.

The physical parameters and some of the biological parameters were measured in soil samples in the laboratory (Garnier et al., 2001). The remaining parameters were estimated by fitting the CANTIS submodel to incubation data (C and N mineralization curves).

4. Sensitivity analysis

Sensitivity analysis was carried out for two output variables, namely the amount of nitrate mineralized from 0 to 60 cm ($\hat{Y}_{\mathrm{MN}}$) and the amount of nitrate leached below 150 cm soil depth ($\hat{Y}_{\mathrm{LN}}$). Furthermore, we examined the sensitivity of model efficiency for nitrate content, which measures how well the model reproduces observed nitrate content. The definition of modeling efficiency here is

$$Y_{\mathrm{EF}} = \frac{\sum_{i=1}^{N_t}[\sum_{j=1}^{N_z}(Y_{\mathrm{NO_3}}(t_i,z_j)-\bar{Y}_{\mathrm{NO_3}})^2-\sum_{j=1}^{N_z}(\hat{Y}_{\mathrm{NO_3}}(t_i,z_j)-Y_{\mathrm{NO_3}}(t_i,z_j))^2]}{\sum_{i=1}^{N_t}[\sum_{j=1}^{N_z}(Y_{\mathrm{NO_3}}(t_i,z_j)-\bar{Y}_{\mathrm{NO_3}})^2]}$$

where N_t and N_z are respectively the number of dates and the number of depths represented in the data.

The input factors for the sensitivity analysis were chosen to include factors likely to be important for different processes. The coefficient a_0 of the hydraulic conductivity function and the potential evapotranspiration PET were chosen because they should be important for drainage. To study the effect of temperature on mineralization, the thermal conductivity coefficient λ'_{T} and the temperature factor coefficient of the humified organic matter decomposition B_{T} were included. The autochthonous biomass decomposition rate (k_A) was included because it strongly influences mineralization. Finally, initial nitrate content $NO^{-}_{3,\mathrm{INI}}$ was included because it is difficult to measure (coefficient of variation of 20%). It is therefore important to determine the consequences of errors in this input variable. Sensitivity analysis of the model was carried out for one factor at a time, all the other factors being fixed at their nominal values. Each factor was varied in the range (nominal value − 60%) to (nominal value − 60%). The sensitivity index (*SI*) for model output Y with respect to factor X was calculated as:

$$SI = \frac{\underset{X \in U}{\text{Max}}\,[Y\,(X)] - \underset{X \in U}{\text{Min}}\,[Y\,(X)]}{\underset{X \in U}{\text{Max}}\,[Y\,(X)]} \tag{1}$$

where U is the range of values explored for the factor X.

5. Results

The response profiles are shown in Figure 2 for each output variable – input factor pair. Figure 3 shows a comparison of observed and simulated nitrate content in two soil layers over time for the nominal values of all factors and when K_{A} or *PET* is changed.

Consider first the effect of the autochtonous biomass decomposition rate k_{A}. This parameter had a strong effect on mineralization, (Fig. 2a) but only a small effect on the amount of leached nitrate (Fig. 2b). N mineralization in the first layer (0–30 cm) was reduced in the spring and summer of 1994 when k_{A} was decreased (Fig. 3a), but k_{A} had only a slight effect on the earlier period from October to March because the degree of mineralization was low during this period. Changing k_{A} had only a small influence on the simulated amount of leached nitrate because the nitrate produced during the second period moved downwards during the autumn and winter of 1994 and this period is not simulated. The sensitivity analysis results would probably have been different if the experiment had continued over a period of two years. The model efficiency remained close to its initial value of 0.75 as k_{A} varied (Fig. 2c). This is probably because k_{A} mainly affects nitrate in the upper layers whereas modeling efficiency concerns nitrate in all layers.

The temperature factor coefficient B_{T} had a strong effect on mineralization of N (Fig. 2a). The effect of B_{T} depends on the temperature. If the temperature is below T_{ref} (15°C) then an increase in B_{T} increases mineralization, whereas, if the temperature exceeds 15°C B_{T} decreases mineralization. The amount of leached nitrate and model efficiency were not strongly affected by B_{T}.

The initial nitrate content of the first layer $NO^{-}_{3,\mathrm{INI}}$ had a strong influence on the amount of leached nitrate and on model efficiency (Fig. 2b,c) but did not affect the mineralization of N (Fig. 2a). The mean coefficient of variation of nitrate measured in the first layer was 21%.

Figure 2 shows that a_0 has a strong influence on the amount of leached nitrate and on model efficiency but not on the mineralization of N. The transfer velocity of nitrate is increased when a_0 is increased. This effect reduced the remaining nitrate content at the bottom of the soil profile at the end of the experiment. The model efficiency is increased when a_0 has a higher value.

The thermal conductivity λ'_{T} had a very small effect on the model outputs. The simulation of the nitrate dynamics is slightly improved when the thermal conductivity is increased.

Finally, Figure 2 shows that the potential evapotranspiration *PET* had a strong influence on the amount of leached nitrate, but only a slight influence on the amount of mineralized

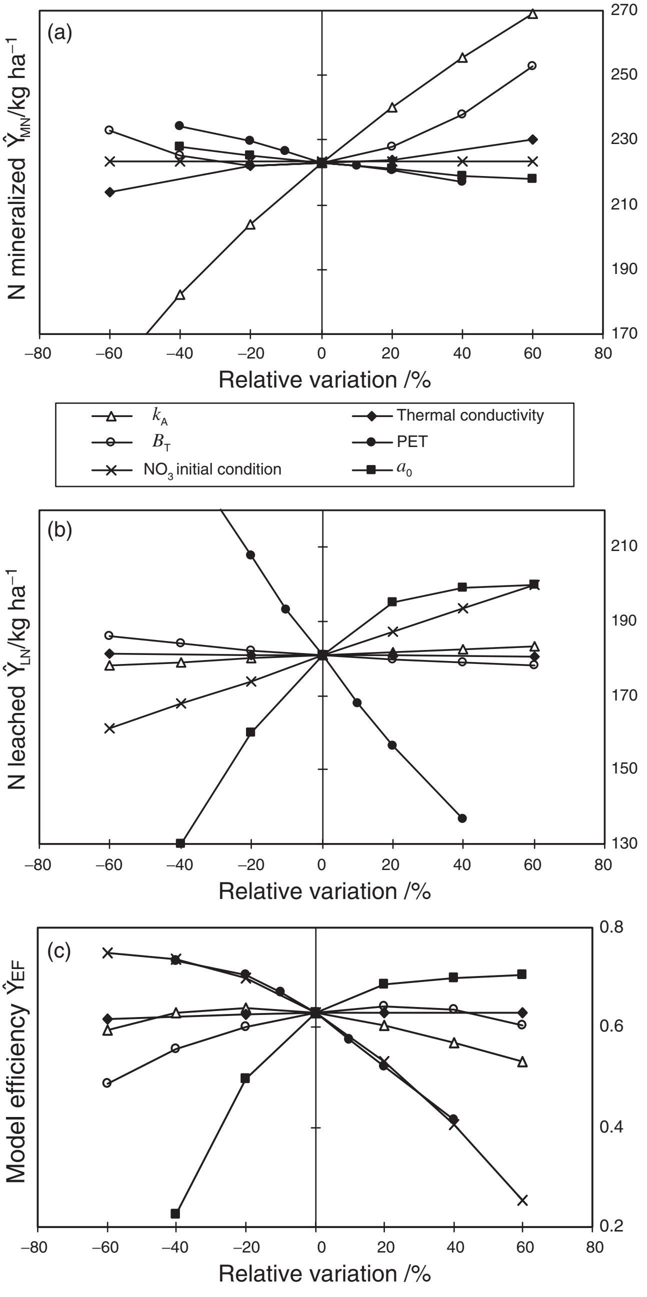


Figure 2. Sensitivity analysis response profiles.

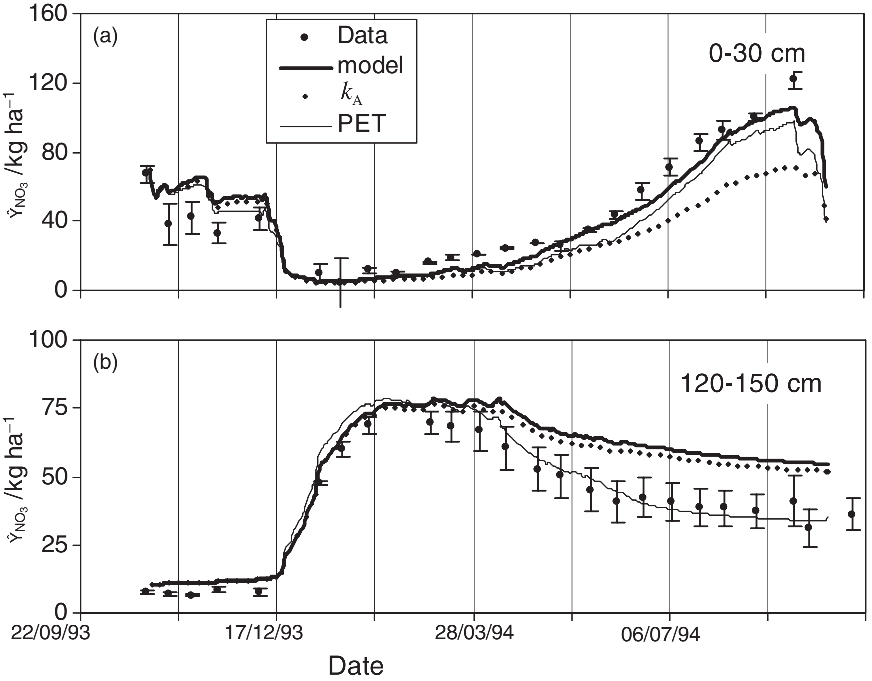


Figure 3. Graphical comparison of measured and simulated NO_3^- contents in the 0–30 cm (a) and 120–150 cm layers (b). The dotted line represents simulations when K_A is decreased by 60%. The thin continuous line represents simulations when *PET* is decreased by 40%.

nitrogen. When *PET* is decreased by 40%, the nitrate contents of the deeper layers are strongly reduced and are very close to the measured values (Fig. 3b), and the efficiency is increased (Fig. 2c).

The sensitivity indices are presented in Table 4. These indices are a summary of the more detailed information in the response profiles. It can be seen that mineralized nitrogen $\hat{Y}_{MN}$ is most sensitive to the biological input parameters k_A and B_T. Leached nitrogen $\hat{Y}_{LN}$ is most sensitive to the input variables *PET* and $NO_{3,INI}^-$ and to the physical parameter a_0. Model efficiency Y_{EF}, like $\hat{Y}_{LN}$, is most sensitive to *PET*, $NO_{3,INI}^-$ and a_0. The parameter λ_T' seems to have relatively little effect on model outputs.

6. Conclusion

The sensitivity analysis shows that model efficiency depends mainly on the hydraulic conductivity a_0, on *PET* and on $NO_{3,INI}^-$, and not on biological inputs or parameters. We conclude that the overestimation of simulated nitrate content at 150 cm was probably

Table 4. Sensitivity indices.

	k_A	B_T	PET	λ'_T	a_0	$NO^-_{3,INI}$
$\hat{Y}_{MN}$	0.414	0.123	0.105	0.070	0.039	0
	PET	a_0	$NO^-_{3,INI}$	B_T	K_A	λ'_T
$\hat{Y}_{LN}$	0.547	0.35	0.191	0.043	0.027	0.006
	a_0	$NO^-_{3,INI}$	PET	B_T	K_A	λ'_T
$\hat{Y}_{EF}$	0.686	0.655	0.56	0.351	0.159	0.032

due to uncertainty in the simulation of transport processes rather than to the simulation of biological processes.

The sensitivity analysis further shows that the factors that had a strong effect on nitrate leaching but not on mineralization (a_0, PET and $NO^-_{3,INI}$) also had a strong effect on model efficiency. On the other hand, the factors that had a strong effect on mineralization but not on leaching (the autochtonous biomass decomposition rate k_A and the temperature factor coefficient of HOM decomposition B_T) had only a small effect on model efficiency.

The fact that it is the factors that affect leaching, and not the factors that affect mineralization, that also affect model efficiency is probably due to the fact that only one year was simulated. This tends to increase the importance of transport processes compared to biological processes. The nitrate produced after May had no time to leach out before the end of the experiment in September. An experimental period of 2 years would no doubt have increased the sensitivity of the model to biological parameters. A related problem is that model efficiency based on nitrate content data is probably not a good measure of how well the model simulates biological processes.

References

Andrén, O., Steen, E., Raghai, K., 1992. Modelling the effects of moisture on barley straw and root decomposing in the field. Soil Biology and Biochemistry 24, 727–736.

Garnier, P., Néel, C., Mary, B., Lafolie, F., 2001. Evaluation of a nitrogen transport and transformation model in a bare soil. European Journal of Soil Science 52, 253–268.

Garnier, P., Néel, C., Aita, C., Recous, S., Lafolie, F., Mary, B., 2003. Modelling carbon and nitrogen dynamics in a bare soil with and without straw incorporation. European Journal of Soil Science 54, 555–568.

Hadas, A., Parkin, T.B., Stahl, P.D., 1998. Reduced CO_2 release from decomposing wheat straw under N-limiting conditions: simulation of carbon turnover. European Journal of Soil Science 49, 487–494.

Recous, S., Robin, D., Darwis, S., Mary, B., 1995. Soil inorganic N availability: effect on maize residue decomposition. Soil Biology and Biochemistry 27, 1529–1538.

Rodrigo, A., Recous, S., Néel, C., Mary, B., 1997. Effects of moisture and temperature on microbial processes in soils: comparison of nine simulation models. Ecological Modelling 102, 325–339.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 16

Sensitivity analysis of GENESYS, a model for studying the effect of cropping systems on gene flow

N. Colbach and N. Molinari

1. Introduction **377**
2. Presentation of the model GENESYS **379**
3. Materials and methods **379**
4. Results **383**
5. Discussion **385**
6. Conclusion **389**
References **389**

1. Introduction

GENESYS-COLZA (Colbach et al., 2001a,b) quantifies the effects of cropping systems (regional field pattern, crop succession, cultivation techniques, oilseed rape varieties) on gene flow from new rapeseed varieties to volunteers and feral populations over time in agricultural regions. The present version models a transgene coding for herbicide-tolerance to a non-selective herbicide. This gene flow can lead to (a) the spread of herbicide-tolerant rape volunteers and to (b) the contamination of non-GM (genetically modified) harvests by transgenes, making them unsuitable for a non-GM commercialisation label. The aim of the model is to rank cropping systems according to their risk of gene flow in order to (a) evaluate gene flow in existing cropping and farming systems, to (b) design new field and regional cropping systems limiting gene flow, and to (c) identify the characteristics of rape varieties that increase or decrease gene flow.

The model was evaluated by confronting its simulated output with independent field data (Colbach et al., 2005a). It has also been used to evaluate and design cropping systems in the case of coexisting GM, non-GM and organic crops (Angevin et al., 2002; Colbach et al., 2004a). During these studies, questions concerning the survey of cropping history in farmers' fields and the choice of input variables for the simulations arose frequently:

- When gene flows are simulated in a cluster of fields, how far is it necessary to survey and simulate fields around the cluster, in order to obtain accurate simulations in the cluster?
- For how many past years is it necessary to know and simulate the cropping history of all these fields? After how many years of simulation does the effect of the initial seed bank, which is usually unknown in a field, become insignificant?

- Which cultivation techniques must be described with the greatest precision for accurate simulations? Which are the cultivation techniques that influence gene flow most and are to be modified first to limit gene flow?

Conceptually, the simplest method for sensitivity analysis is to vary repeatedly one input variable at a time while keeping the others fixed. This type of one-at-a-time analysis only assesses sensitivity relative to point estimates and not to the entire input distributions, which better reflect the uncertainty on the model inputs (Hamby, 1994). Because of the number and the complexity of input variables and their interactions present in GeneSys, this conventional technique where each input variable varies and is analysed separately is neither appropriate nor feasible. The model structure requires to assess sensitivity with regard to the combined variability resulting from considering all input variables simultaneously, which has never been attempted yet.

The number of input variables was, however, much too large to study them all simultaneously. Consequently, the analysis was split into several parts (Fig. 1) and only the results concerning the effect of the initial seed bank are presented here. Furthermore, the sensitivity of the model to input variables depends on the output variable studied. Only two major output variables were studied here, (a) the pollution rate of conventional rape harvests by a transgene or any other extraneous gene that would render the harvest less valuable or improper for commercialising; (b) the density of rape volunteers in winter cereals where the volunteers are considered as weeds and responsible for yield losses. The present study was carried out, assuming that the newly introduced gene/allele was a

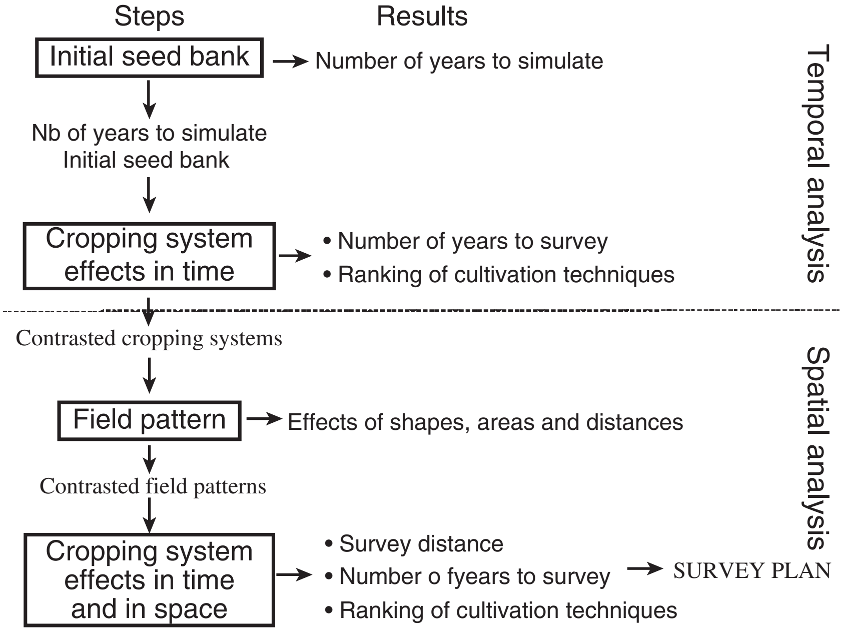

Steps
Results
Initial seed bank
Number of years to simulate
Nb of years to simulate
Initial seed bank
Cropping system effects in time
• Number of years to survey
• Ranking of cultivation techniques
Temporal analysis
Contrasted cropping systems
Field pattern
Effects of shapes, areas and distances
Contrasted field patterns
Spatial analysis
Cropping system effects in time and in space
• Survey distance
• Number o fyears to survey
• Ranking of cultivation techniques
SURVEY PLAN

Figure 1. Temporal and spatial steps in the sensitivity analysis of the GeneSys model.

dominant allele *A* and that the new variety was a homozygous *AA*, as is the case for the new herbicide-tolerant transgenic varieties.

2. Presentation of the model GeneSys

Only the temporal part of the GeneSys model is considered here, and so only one field is considered for each simulation. Details are given by Colbach et al. (2001a,b). The model uses the following input variables:

(1) the **crop** grown each year on the simulated field, with eight modalities, i.e. GM rape, non-GM rape, winter or spring cereals, unsown, autumn-sown, spring-sown or permanent set-aside.
(2) the **cultivation techniques** used to manage each crop, comprising stubble breaking, tillage, sowing date and density, herbicide spraying, cutting and harvest loss (i.e. seeds lost before or while harvesting rape crops).
(3) the **genetic variables.** The genotypes of the transgenic and conventional rape varieties are *AA* and *aa*, respectively, where *A* is the transgene or any dominant allele coding for herbicide tolerance and *a* the associated recessive allele. The varieties can differ in pollen and seed production, and the difference depends on whether the plants grow in rape crops or in other crops. Self-pollination rates also vary according to genotype.

The initial seed bank must be determined on the onset of simulation, with the number of seeds and their genotype proportions in each soil layer.

These input variables influence the annual life cycle of both cultivated and volunteer rape plants, which comprises seedlings, adults, flowers, seed production and seed bank left after harvest. The life cycle is repeated for each simulated year. For each life stage, both the number of individuals per square meter and the proportion of the three genotypes *AA, Aa* and *aa* are calculated.

The main output variables are, for each simulated year, the adult plants (whether cultivated or volunteer), the seed production and the seed bank left after harvest. For each of these variables, both the number of individuals per square meter and the genotype proportions are calculated.

3. Materials and methods

The sensitivity analysis to the initial seed bank was based on extensive simulations followed by statistical analyses on the simulated output. In the simulations, the input variables describing the seed bank were treated as controlled experimental factors (e.g. total seed density, proportion of *AA* seeds . . .). All combinations of factor levels were used in the simulations except unrealistic ones (e.g. deeply buried recently shed seeds). The model input variables describing the cropping systems, on the other hand, were determined by Monte Carlo sampling. Each Monte Carlo sample of cropping systems was crossed with each combination of the controlled factors related to the initial seed bank. After the simulations were completed, the simulated output was analysed statistically, as the result of a virtual experiment.

3.1. Choice of input factors

3.1.1. Initial seed bank

The series of initial seed banks was obtained by combining the following factors:

- the distribution among seed age classes, with 3 modalities: either 100% of seeds less than one-year-old (hence "young seeds"), 100% of seeds older than one year ("old seeds") or a 50–50% mixture of each class.
- the distribution of old seeds among the four 5-cm-thick soil layers, with 3 modalities: either 100% of seeds in the top 5 cm, 100% of seeds on the deepest 5 cm, or a homogeneous distribution among the four layers, with 25% of seeds in each layer. Young seeds can only be found in the top layer as the initial seed bank illustrates a seed bank left immediately after a crop harvest.
- genotype proportions with 7 modalities: 100% of *AA*, *Aa* or *aa* seeds, respectively; 50–50 mixtures of *AA* and *Aa*, *AA* and *aa* or *Aa* and *aa*, respectively; or a homogeneous mixture with a third of *AA*, *Aa* and *aa*.
- total number of seeds of the seed bank, with 4 levels: 0, 100, 1000 or 10 000 seeds per m^2. The latter value is approximately the amount of seeds left after a rapeseed harvest with 5% of seed loss. When the total seed bank was zero, the first three factors did not vary.

In total, there were $1 \times 1 \times 7 \times 3$ (young seeds) $+ 2 \times 3 \times 7 \times 3$ (old seeds and mixture of old and new) $+ 1$ (empty seed bank) $= 148$ initial seed banks tested.

3.1.2. Cropping system

3.1.2.1. Genetic input factors

In all simulations, the transgenic allele was a dominant *A*. The first genetic input factor was the variety type when rape was grown, with two modalities: transgenic varieties *AA* and conventional ones *aa*. The remaining seven genetic input factors were the self-pollination rates of *AA*, *Aa* and *aa* genotypes; the relative pollen emission and yield rates for GM *vs.* non-GM plants in rape crops; and the relative pollen emission and yield rates in environments other than rape. The levels of these seven quantitative factors were sampled randomly within [0; 1] at the beginning of each simulation, according to the uniform distribution.

3.1.2.2. Crop succession and management

Each year other than the final one, the simulated crop was sampled with equal probabilities among the eight possible crops given in Section 2, according to the uniform distribution. The only exception was permanent set-aside which could only follow a sown set-aside.

When the analysed output variable was harvest pollution, the last crop was always a non-GM rape. When the rape volunteers in winter cereals were studied, the last crop was a winter cereal.

Each year, cultivation techniques were also sampled with uniform probability, but the sampled techniques depended on the simulated crop (Table 1). For instance, a rape crop could be preceded by chisel ploughing, mouldboard ploughing or zero tillage and be sown between 1 August and 30 October, whereas unsown set-aside could never be preceded by tillage and was of course never sown.

3.2. Duration of simulation runs and number of replications

The duration of each simulation run was 25 years. For each of the 148 initial seed banks tested, 10 000 simulations with random cropping system variables were run. For better precision when comparing initial seed banks, the same 10 000 cropping systems, determined by Monte Carlo sampling, were crossed with the 148 initial seed banks.

These numbers of replications were ridiculously low compared to the enormous number of possible combinations of input variables and/or parameters. However, it will never be possible to explore more than a tiny proportion of these combinations. Therefore, to test the stability of the observed simulation results, all the analysis steps were repeated with a different set of simulated replications.

3.3. Statistical methods

In order to synthesise the results and to identify the major input factors, the simulated output was analysed, using analysis of variance (see the Chapter 3). The analysed output variable was either (a) the proportion of GM seeds (genotypes *AA* or *Aa*) in non-GM rape harvests (hence “harvest pollution”); or (b) the total density of rapeseed volunteers in winter cereals.

For each year *t*, the analysis was restricted to the simulation runs with either a non-GM rape variety in year *t* (output variable = harvest pollution) or winter cereals in year *t* (output = volunteer density). The main model was:

$$\begin{aligned} \textit{output variable (t)} &= \textit{constant} + \textit{initial seed bank effect} + \textit{cropping system effect} \\ &\quad + \textit{error} \end{aligned} \tag{1}$$

The sensitivities to the initial seed bank and to the cropping system were evaluated by comparing the r^2 of Eq. (1) with the r^2 of Eq. (2) and Eq. (3) below,

$$\textit{output variable (t)} = \textit{constant} + \textit{cropping system effect} + \textit{error} \tag{2}$$

$$\textit{output variable (t)} = \textit{constant} + \textit{initial seed bank effect} + \textit{error} \tag{3}$$

Table 1. Range of possible values for cultivation techniques, depending on the simulated crop.

Technique	Rapeseed		Cereals		Set-aside				Border
	Transgenic	Conventional	Winter	Spring	Autumn-sown	Spring-sown	Unsown	Permanent	
Stubble breaking	A/P	A/P	A/P	A/P	A/P	A/P	A	A	A
Tillage	Chisel/ plough/ no tillage	Chisel/ plough/ no tillage	Chisel/ plough/ no tillage	Chisel/ plough/ no tillage	Chisel/ plough/ no tillage	Chisel/ plough/ no tillage	No tillage	No tillage	No tillage
Sowing date	[1 Aug., 30 Oct.]	[1 Aug., 30 Oct.]	[1 Sept., 30 Nov.]	[1 Feb., 31 May]	[1 Aug., 30 Nov.]	[1 Feb., 31 May]	No sowing	No sowing	No sowing
Sowing density (seeds/m^2)	[1, 150]	[1, 150]	[1, 450]	[1, 450]	[1, 600]	[1, 600]	0	0	0
1st and 2nd cutting	A	A	A	A	A	A	A/P	A/P	A/P
Cutting date (days after flowering onset)									
1st cutting	A	A	A	A	A	A	[0, 39]	[0, 39]	[0, 39]
2nd cutting	A	A	A	A	A	A	[0, 22]	[0, 22]	[0, 22]
Herbicides 1 and 2% (mortality of genotypes)									
aa	[0, 100]	0	[0, 100]	[0, 100]	[0, 100]	[0, 100]	[0, 100]	[0, 100]	[0, 100]
AA and *Aa*	0	0	[0, 100]	[0, 100]	[0, 100]	[0, 100]	[0, 100]	[0, 100]	[0, 100]
Harvest loss (%)	[0, 100]	[0, 100]	100	100	100	100	100	100	100
Grazing	A	A	A	A	A	A	A	A/P	A

A – absence; P – presence. For a given crop or border, cultivation techniques are chosen randomly according to a uniform distribution among the possible levels (qualitative variables) or in the range of possible values (quantitative values).

r^2 is defined as $r^2 = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 / \sum_{i=1}^{N}(Y_i - \bar{Y})^2$ where Y_i is the ith output value simulated by GeneSys; $\hat{Y}_i$ the corresponding value predicted by one of the linear models, and $\bar{Y}$ the average value of the output simulated by GeneSys.

In these models, initial seed bank was a categorical factor with 148 levels; cropping system was also treated as a categorical factor, with each Monte Carlo sample being considered as a distinct modality. Because in each year, only those runs with non-GM rape or winter cereals could be used, depending on whether harvest pollution or volunteer density was analysed, the cropping system factor had approximately 1300 modalities instead of 10 000. In order to identify the seed bank characteristics that influence gene flow, the following linear model was also tested:

$$\begin{aligned} \textit{output variable}\,(t) = {} & \textit{constant} + \textit{cropping system effect} + \textit{seed bank density effect} \\ & + \textit{seed age effect} + \textit{vertical distribution effect} \\ & + \textit{genotypic composition effect} + \textit{error} \end{aligned} \tag{4}$$

All seed bank variables were treated as categorical factors. Seed bank density had four levels (0, 100, 1000 or 10 000 seeds/m^2), seed age three levels (young seeds, old seeds or a mixture of both), vertical distribution three levels, nested in the seed age (young seeds: top; old seeds: top, bottom, uniform) and genotype composition had 7 levels (100% *AA*, 100% *Aa*, 100% *aa*, 50% *AA*–50% *Aa*, 50% *AA*–50% *aa*, 50% *Aa*–50% *aa* and 33% *AA*–33% *Aa*–33% *aa*).

These analyses of variance were carried out with the GLM procedure of SAS (Statistical Analysis System, SES institute Inc, Cary, NC, USA).

4. Results

4.1. Harvest pollution

The year following the initial seed bank, the variability in harvest pollution of the non-GM rape was explained in approximately the same proportion by the main effect of the initial seed bank and by the main effect of the cropping system, with a strong interaction between the two factors (Fig. 2). Indeed, the r^2 of the two models reduced to the single factors *initial seed bank* and *cropping system*, respectively, were roughly half of the r^2 of the additive model comprising the both factors. But this model did not comprise any interaction and its r^2 was only 0.60 and therefore considerably less than 1. The effect of the initial seed bank decreased rapidly with time and from the 7th year onwards, the part of variability (0.0004) explained by the effect of the initial seed bank decreased below 0.001 and became negligible, especially compared to that explained by cropping system effects (0.98).

The most important characteristic of the initial seed bank was its genotypic composition which influenced harvest pollution for approximately 9 years (Table 2A) until the

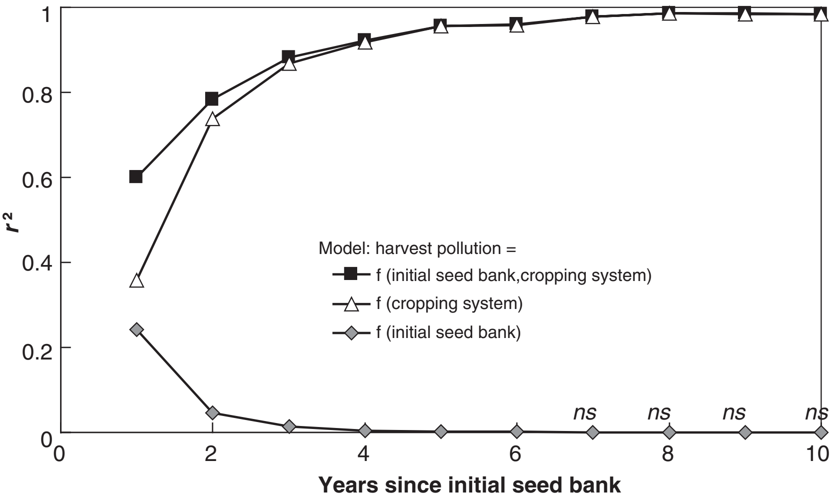


Figure 2. Proportion of variability in non-GM rape harvest pollution explained by initial seed bank (♦), cropping system (△) or both factors (■), as a function of time since initial seed bank. ns = not significant at alpha = 0.01 in model [3] (Colbach et al., 2004c).

part of explained variability (partial r^2) decreased below 0.0001 and became negligible. Harvest pollution increased with the frequency of the transgene in the seed bank (Table 2B). Total seed density was significant during 4 years and harvest pollution increased with density. Seed age influenced harvest pollution during 3 years only, with a mixture of both young and old seeds resulting in the highest risk and a seed bank comprising old seeds only resulting in the lowest risk. The initial vertical seed distribution influenced harvest pollution only during the first two year, with superficial or uniformly distributed seed banks resulting in a higher risk.

4.2. Rape volunteer density

The effect of the initial seed bank on rape density in winter cereals was much smaller than on harvest pollution and its effect had already disappeared after two years (Fig. 3), when the variability explained by the seed bank effect decreased below 0.001. However, there was even more interaction between seed bank and cropping system than in the case of harvest pollution as the r^2 of the additive model comprising both seed bank and cropping system effects (but not their interaction) was only 0.26 for the first year after simulation onset.

The ranking of the various components of the seed bank was different from the one for harvest pollution. In contrast to harvest pollution, the genotype composition of the initial seed bank did not influence volunteer density at all (Table 3A). The other factors were ranked as for harvest pollution but their effects did not last as long, with only 3 years after simulation onset for seed density and one year for seed age and seed distribution. The levels of seed density and vertical distribution were ranked identically for both volunteer density and harvest pollution (Table 3B). For seed age and genotypic composition, the ranking was

Table 2. Effects of characteristics of initial seed bank on harvest pollution of non-GM rape. Results of linear model: harvest pollution = constant + cropping system effect + seed bank density effect + seed age effect + vertical distribution effect + genotypic composition effect + error (Colbach et al., 2004b).

A. Variability explained by each variable (partial r^2)

Years since initial seed bank	Partial r^2 (sum of squares/total sum of squares)					r^2
	Cropping system	Seeds/m^2	Seed age	Vertical distribution	Genotypic composition	
1	**0.3596**	**0.1055**	**0.0021**	**0.0049**	**0.0757**	0.5480
2	**0.7385**	**0.0118**	**0.0016**	**0.0006**	**0.0219**	0.7746
3	**0.8683**	**0.0023**	**0.0005**	0.0001	**0.0077**	0.8791
4	**0.9178**	**0.0003**	<0.0001	<0.0001	**0.0037**	0.9221
5	**0.9551**	<0.0001	<0.0001	<0.0001	**0.0014**	0.9567
6	**0.9579**	<0.0001	<0.0001	<0.0001	**0.0011**	0.9590
7	**0.9777**	<0.0001	<0.0001	<0.0001	**0.0004**	0.9782
8	**0.9865**	<0.0001	<0.0001	<0.0001	**0.0002**	0.9865
9	**0.9850**	<0.0001	<0.0001	<0.0001	**0.0001**	0.9850
10	**0.9839**	<0.0001	<0.0001	<0.0001	**0.0001**	0.9839

B. Comparison of means for the year following the initial seed bank

Seeds/m^2	Mean pollution	Seed age	Mean pollution	Vertical distribution	Mean pollution	Genotypic composition	Mean pollution
10 000	0.1690	young + old	0.0947	Top	0.1006	*AA*	0.1651
1000	0.0749			Uniform distribution	0.0942	*AA–Aa*	0.1396
100	0.0208	young	0.0941			*Aa*	0.1094
		old	0.0758	Bottom	0.0699	*AA–Aa–aa*	0.0834
						AA–aa	0.0746
						Aa–aa	0.0454
						aa	0.0000

different: volunteers densities were highest if the initial seed bank consisted exclusively of old seeds and increased with the proportion of *aa* seeds in the initial seed bank.

5. Discussion

5.1. Explaining effects

The experimental plan used for exploring seed bank effects in combination with cropping system avoided concluding on the effects of a given input variable while only using one set of values for the remaining input variables, which has frequently been done in previous sensitivity analyses (Hamby, 1994). The analyses of variance allowed us to identify the most important input variables and their effects on the simulated output.

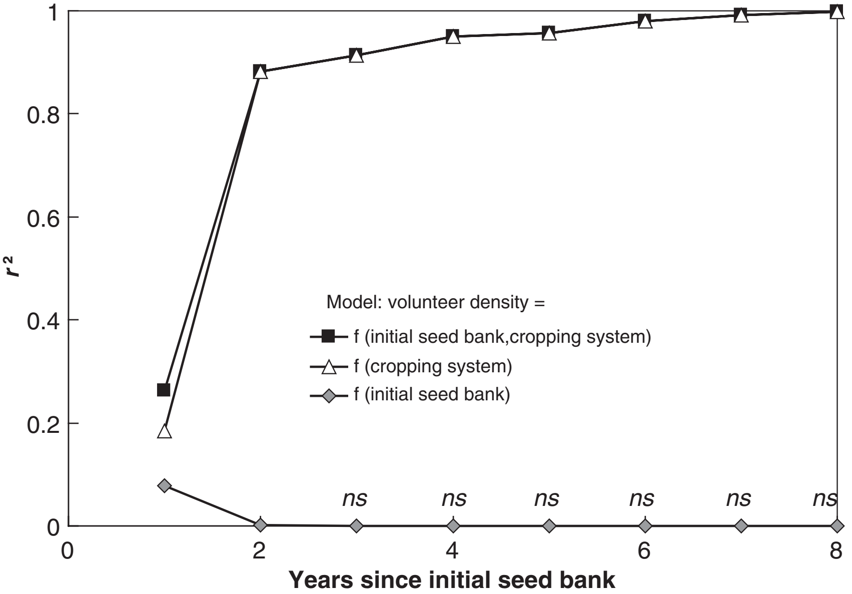


Figure 3. Proportion of variability in rape volunteer density in winter cereals explained by initial seed bank (♦), cropping system (△) or both factors (■), as a function of time since initial seed bank. ns = not significant at alpha = 0.01 in model [3].

The results of the two output variables were consistent with the known biological and physical processes occurring in the field. First, the relative importance of the various components of the initial seed bank depended on the analysed output. For instance, for output variables illustrating genotype proportions such as harvest pollution of non-GM rape (i.e. the rate of *AA* and *Aa* seeds in *aa* harvests), the most important characteristic of the initial seed bank was its genotypic composition, whereas frequency output variables such as volunteer densities in winter cereals depended mostly on the seed density of the initial seed bank.

The effects of each seed bank characteristic could also be explained consistently. The effect of initial seed density was easy to understand: whatever the analysed output variable, the more seeds there were initially, the higher the simulated output. The effect of the genotypic composition of the initial seed bank depended on the output variable. The relationship was again obvious for harvest pollution which increased with the content of *AA* and *Aa* seeds in the initial seed bank. In contrast, volunteer densities slightly increased with the *aa* content in the seed bank. Indeed, depending on the variety characteristics chosen for GM and non-GM varieties in the cropping system, the *aa* genotype could produce more pollen and seeds than *AA* or *Aa* genotypes. This difference is usually referred to as "cost of herbicide tolerance".

The effect of seed age was more difficult to analyse. Harvest pollution was highest when the seed bank consisted mostly of young seeds whereas the opposite was true for volunteer densities. Consequently, as the main advantage of old seeds over young seeds

Table 3. Effects of characteristics of initial seed bank on the density of rape volunteers in winter cereals. Results of linear model: volunteer density = constant + seed bank density effect + seed age effect + vertical distribution effect + genotypic composition effect + error.

A. Variability explained by each variable (partial r^2)

Years since initial seed bank		Partial r^2				r^2
	Cropping system	Seeds/m^2	Seed age	Vertical distribution	Genotypic composition	
1	**0.1847**	**0.0480**	**0.0078**	**0.0030**	0.0001	0.2437
2	**0.8824**	**0.0006**	<0.0001	<0.0001	<0.0001	0.8832
3	**0.9138**	**0.0004**	<0.0001	<0.0001	<0.0001	0.9143
4	**0.9499**	<0.0001	<0.0001	<0.0001	<0.0001	0.9501
5	**0.9570**	<0.0001	<0.0001	<0.0001	<0.0001	0.9571
6	**0.9801**	<0.0001	<0.0001	<0.0001	<0.0001	0.9802
7	**0.9921**	<0.0001	<0.0001	<0.0001	<0.0001	0.9921
8	**0.9981**	<0.0001	<0.0001	<0.0001	<0.0001	0.9981
9	**0.9986**	<0.0001	<0.0001	<0.0001	<0.0001	0.9986
10	**0.9138**	<0.0001	<0.0001	<0.0001	<0.0001	0.9138

B. Comparison of means for the year following the initial seed bank

Seeds/m^2	Mean density	Seed age	Mean density	Vertical distribution	Mean density	Genotypic composition	Mean density
10 000	1.1631	Old	0.6848	Top	0.6003	*aa*	0.4658
1000	0.1316	Young	0.4490	Uniform distribution	0.4228	*Aa–aa*	0.4493
100	0.0133	Young + old	0.1742			*AA–aa*	0.4493
				Bottom	0.2849	*AA–Aa–aa*	0.4401
						Aa	0.4159
						AA–Aa	0.4159
						AA	0.4159

is their better survival ability, it appeared that seed survival was a major process for volunteer densities in winter cereals, even for the year following the initial seed bank. For the earlier sown rape crops of which harvest pollution was analysed, this aspect was less important as there would be less time for seed bank to decrease between the onset of the simulation and the sowing of the crop. In the case of rape crops, the main aspect seemed to be the relative emergence of volunteers (which are at least partially GM) *vs.* cultivated non-GM plants and the emergence of volunteers is best if seeds are close to soil surface which is usually the case of younger seeds.

This aspect is consistent with the ranking observed for the vertical seed distribution of the initial seed bank. Whatever the analysed output variable, the risk was highest when seeds were concentrated in the top soil layer and lowest when they were buried. Indeed, emergence decreases with seed depth and even soil-inverting tillage modes such as mouldboard ploughing will only carry back part of the buried seeds (Cousens and Moss, 1990; Colbach et al., 2000; Roger-Estrade et al., 2002).

The results varied little between the two simulated data sets, and this despite the low number of simulations compared to the huge number of possible combinations of input variables.

5.2. Determination of input data for evaluations and simulations

The first step of the sensitivity analysis was to evaluate the long-term effect of the initial seed bank. Its result is crucial for the future evaluation and use of GENESYS as the initial seed bank is difficult to estimate for a single field (Dessaint et al., 1992) and impossible to determine for all fields and borders constituting a field plan. The analysis showed that the effect of the initial seed bank became negligible after approximately 7 years of simulation when analysing harvest pollution of conventional rape. This duration is the necessary span of simulated years to initialise the system and to have a realistic seed bank to start the "real" simulation. Consequently, if the user wants to analyse harvest pollution in a region with a seven-year-long rotation comprising transgenic rapeseed, he/she should start with an empty seed bank followed by a transgenic rapeseed and then simulate for 14 years at least and only analyse the last 7 years of the simulation. If, however, the user is interested in what happens when a transgenic crop is introduced into a region formerly cultivated with non-transgenic rapeseed, then he/she first needs to simulate a 7-year crop succession with non-transgenic rape before switching to a rotation with transgenic rape, and then analyse the last rotation only.

The main effect of initial seed bank on volunteers in winter wheat lasted only for two years but as the analysis indicated a strong interaction between the seed bank and the cropping system effects, the necessary simulation span should be extended for a few years, as a precaution.

However, whatever the analysed output variable, when the model is used to compare simulations and observations as in the case of model evaluation, it is not always possible to gather the field history for such a long period. It is therefore necessary to analyse the sensitivity of the model to the cropping system variables to identify the most pertinent ones that must be obtained in field surveys. It is probably not necessary to estimate all variables with the same degree of precision.

5.3. Determination of pertinent changes in cropping systems

The sole analysis of the effect of initial seed bank is of course insufficient to advise farmers on how to modify their cropping systems to limit gene flow, especially as this first step of the sensitivity analysis showed cropping system to be the dominant factor in this process. However, a few ideas already take shape. For instance, the importance for harvest pollution of the genotype composition of the seed bank would indicate the ratio of GM to non-GM varieties in the rotation as well as the use of herbicides eliminating exclusively one genotype to be major factors. In the case of volunteer density, it would rather be the overall frequency of rape crops, regardless of their variety, illustrated by the importance of the seed bank density. The impact of seed age could be translated as the time since the last rape crop. The analysis of the effect of vertical distributions indicates that tillage modes are not that easy to reason. When all seeds are initially close to soil surface, the optimal solution is easy to find: to till with a mouldboard plough to create a "bottom" seed bank. However, in the next year, the seed bank would resemble a uniform distribution, with the older surviving seeds still at the bottom and the newer seeds shed on top. This situation would present a risk as high as the sole top configuration. The effects of tillage as well as of other cropping system effects must therefore be studied more in detail. Furthermore, the present analysis was limited a single field and the effects of initial seed bank disappear probably more rapidly in case of pollen and seed import from neighbouring fields.

6. Conclusion

This first part of the sensitivity analysis already contributed to the set of rules for the future use of the GENESYS model. It determined the minimum simulation duration to initialise the seed bank for further realistic simulations. This kind of information is crucial for the future evaluation of the model as field seed bank is usually unknown.

However, while the present study revealed the overall importance of cropping system, it did not analyse the elements of cropping system individually. This is absolutely necessary both for determining survey plans for farmers' fields and designing new cropping systems. Furthermore, the present study was restricted to a single field whereas both the actual processes related to gene flow (i.e. pollen and seed dispersal) and the associated model also comprise a spatial dimension. Not only is it necessary to study the sensitivity of the model to field areas, forms and distances, but also the ranking and effects of cropping system elements which may vary for fields interacting with their neighbouring fields. These aspects, i.e. cropping system in time and in space, are analysed in further studies (Colbach et al., 2004b, 2005b).

References

Angevin, F., Colbach, N., Meynard, J.M., Roturier, C., 2002. Analysis of necessary adjustments of farming practices. In: Bock, A.-K., Lheureux, K., Libeau-Dulos, M., Nilsagard, H., Rodriguez-Cerezo, E. (Eds), Scenarios for Co-existence of Genetically Modified, Conventional and Organic Crops in European Agriculture, EUR 20394 EN. Technical Report Series of the Joint Research Center of the European Commission.

Colbach, N., Roger-Estrade, J., Chauvel, B., Caneill, J., 2000. Modelling vertical and lateral seed bank movements during mouldboard ploughing. European Journal of Agronomy 13, 111–124.

Colbach, N., Clermont-Dauphin, C., Meynard, J.M., 2001a. GENESYS: a model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. I. Temporal evolution of a population of rapeseed volunteers in a field. Agriculture, Ecosystems and Environment 83, 235–253.

Colbach, N., Clermont-Dauphin, C., Meynard, J.M., 2001b. GENESYS: a model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. II. Genetic exchanges among volunteer and cropped populations in a small region. Agriculture, Ecosystems and Environment 83, 255–270.

Colbach, N., Angevin, F., Meynard, J.M., Messéan, A., 2004a. Using the GeneSys model quantifying the effect of cropping systems on gene escape from GM rape varieties to evaluate and design cropping systems. OCL 11, 11–20.

Colbach, N., Molinari, N., Clermont, C., 2004b. Sensitivity analyses for a model simulating demography and genotype evolutions with time. Application to GeneSys modelling gene flow between rapeseed varieties and volunteers. Ecological Modelling 179, 91–113.

Colbach, N., Fargue, A., Sausse, C., Angevin, F., 2005a. Evaluation and use of a spatio-temporal model of cropping system effects on gene flow. Example of the GeneSys model applied to three co-existing herbicide tolerance transgenes. European Journal of Agronomy 22, 417–440.

Colbach, N., Molinari, N., Meynard, J.M., Messéan, A., 2005b. Integrating spatial aspects into sensitivity analyses for models simulating demography and genotype evolutions with time. Application to GeneSys modelling gene flow between rapeseed varieties and volunteers. Agronomy for Sustainable Development in press.

Cousens, R., Moss, S.R., 1990. A model of the effects of cultivation on the vertical distribution of weed seeds within the soil. Weed Research 30, 61–70.

Dessaint, F., Barralis, G., Beuret, E., Caixinhas, M.L., Post, B.J., Zanin, G., 1992. Étude coopérative EWRS: la détermination du potentiel semencier: II. estimation de la précision relative sur la moyenne à partir de composites. Weed Research 32, 95–101.

Hamby, D.M., 1994. A review of techniques for parameter sensitivity analysis of environmental models. Environmental Monitoring and Assessment 32, 135–154.

Roger-Estrade, J., Colbach, N., Leterme, P., Richard, G., Caneill, J., 2001. Modelling vertical and lateral weed seed movements during mouldboard ploughing with a skim-coulter. Soil and Tillage Research 63, 35–49.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 17

Data assimilation and parameter estimation for precision agriculture using the crop model STICS

M. Guérif, V. Houlès, D. Makowski
and C. Lauvernet

1. Introduction 391
2. Materials and method 392
3. Results and discussion 394
4. Conclusion 397
References 398

1. Introduction

Crop models simulating soil–plant dynamics as a function of weather and cultural practices are convenient tools for improving nitrogen fertilization decision support systems. One possible approach consists of simulating, before fertilizer application, the effect of different nitrogen doses for a series of possible weather patterns, and selecting the dose that optimizes a given agro-environmental criterion (Meynard et al., 2002; Houlès et al., 2004). There is a big challenge from both the economic and environmental points of view to develop such approaches to support precision agriculture and to recommend spatially variable fertilizer doses. An important problem is therefore to determine the input variables and parameter values at a high spatial resolution. Remote sensing images acquired from satellites or airplanes provide high resolution information on crop characteristics (growth, nitrogen status) through inversion of radiative transfer models (Moulin et al., 2003). This information allows agronomists to estimate, through data assimilation techniques, some of the unknown input variables and parameters and to perform a spatial calibration of the crop model (Guérif and Duke, 1998; Guérif et al., 2001).

This approach is illustrated in this chapter with the STICS wheat crop model for one agricultural plot located in France. In this case study, leaf area index and the amount of nitrogen absorbed by the crop are estimated from airborne remote sensing for 280 cells (20 m × 20 m resolution). Ten model parameters are then estimated cell by cell from these data. The benefits of this approach are discussed.

2. Materials and method

2.1. The STICS model

STICS is a crop model developed by INRA to describe on a daily time step the C, N, and water cycles in the soil–plant system (Brisson et al., 1998). Its main features are its modular construction, uniform level of complexity, and the generic nature of the formalisms chosen to represent the processes. As a consequence, this model can be used to describe a lot of different crops. Among the many (more than 200) parameters of the model equations, some are not supposed to be adjusted to data; some others may be determined for each plant or for each genotype; finally, a number of these parameters represent soil characteristics and may be estimated from soil measurements (soil depth, water holding capacity, apparent density, depth of rooting impediments due to soil constraints, clay content, organic nitrogen content of the cultivated horizon, etc.). Several of the many state variables of the model describe the crop growth. Some of these variables are the leaf area index (*LAI*), the crop biomass, and the crop nitrogen uptake (*QN*). Other variables represent soil physicochemical characteristics such as the water and nitrogen content of different soil layers. STICS can be used to predict many output variables. In this study, we consider only two of them, namely grain yield and grain protein content. These two output variables can be used to compute interesting objective functions for optimizing the amount of applied fertilizer (Houlès et al., 2004).

The model can be expressed as:

$$Z_{t+1} = F(Z_t, X_t; \theta) + \varepsilon_t$$

where Z_t is the $(p \times 1)$ vector including the true p state variables at time t, X_t is the vector including the input variables (weather, applied nitrogen fertilizer, etc.) for day t, θ is defined here as a set of 10 parameters to be adjusted cell by cell, and ε_t is the errors vector $(p \times 1)$. The ten parameters are defined in Table 1 and will be referred hereafter as "the parameters". Eight of them represent soil characteristics and the other two represent some crop characteristics.

Table 1. Prior information for the 10 parameter values. H_n refers to the nth soil horizon.

Parameter	Acronym	Lower bound	Higher bound
Organic nitrogen content (H_1) (%)	Norg	0.04	0.17
Lime content (H_1) (%)	Calc	0	40
Rooting impediment depth (cm)	Obstarac	50	150
Water content at field capacity (H_1) (%)	Hcc(H_1)	17	22
Water content at field capacity (H_2) (%)	Hcc(H_2)	14	22
Water content at field capacity (H_3) (%)	Hcc(H_3)	14	26
Bulk density (H_2) (g cm^{-3})	DA(H_2)	1.45	1.6
Mineral nitrogen content at sowing (H_1) (kg ha^{-1})	Nmin_ini(H_1)	50	85
Life duration of leaves (°C day)	durvieF	140	220
LAI growth coefficient	vlaimax	1.5	2.5

2.2. Data

Data were obtained from a precision agriculture experiment carried out near Laon (northern France) during the year 1999–2000 (Guérif et al., 2001) in a farmer's wheat plot of 10 ha. Airborne remote sensing measurements were made at four dates in April, May, and June during the growth cycle with a CASI sensor; yield (using a yield monitor) and protein content (using sampling and laboratory analysis) were measured at harvest. Soil measurements made on a grid defined in the plot were used to derive the prior distribution of the soil parameters.

The hyper spectral reflectance measurements were inverted through radiative transfer models to estimate *LAI* and chlorophyll content of the leaf (denoted *Cab*, g m^{-2}) with a spatial resolution of 20 m (Moulin et al., 2003). *QN* was estimated from *LAI* and *Cab* using mathematical functions established from destructive measurements. *LAI* and *QN* data were then called "measurements" and these measurements were denoted m_{jtc} where j is the index of the measurement type ($j = 1, 2$ for *LAI* and *QN* respectively), t is the time index ($t = 1, \ldots, 4$), and c is the cell index ($c = 1, \ldots, 280$).

The simulation units were the 280 cells (20 m × 20 m resolution) of the plot and a set of measurements denoted M_c, $c = 1, \ldots, 280$, was defined for each cell. Each set M_c includes eight measurements (4 measurements of *LAI* and 4 measurements of *QN*). For illustration, data obtained at two dates in the 280 cells are displayed in Figure 1.

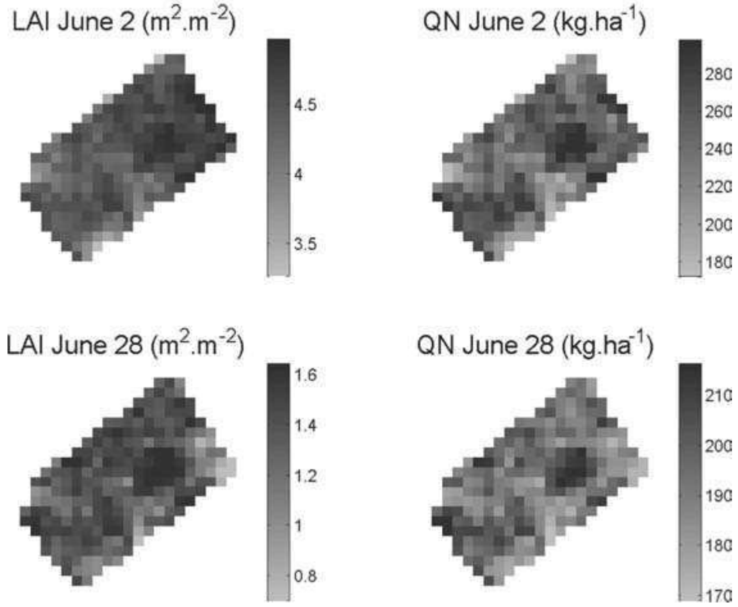


Figure 1. Spatial distribution of "observed" *LAI* and *QN* values at two dates in the 280 cells of the plot.

2.3. Statistical method

The parameters were estimated using the GLUE method described in Chapter 4. This method was used to estimate the ten parameters for each of the 280 cells from the measurements M_c, $c = 1, \ldots, 280$. Note that only ten parameters were estimated from the data and that the other parameters were set equal to fixed values. The 10 parameters were selected from a set of 32 soil parameters and two plant parameters from a sensitivity analysis (Houlès, 2004).

As explained in Chapter 4, the principle of the method is to discretize the parameter space by generating randomly a large number of parameter values θ_i $(i = 1, \ldots, N)$ from a prior parameter distribution. For each cell, the posterior parameter distribution $P(\theta|M_c)$ was approximated by calculating weights p_i at each parameter value θ_i from likelihood $P(M_c|\theta_i)$ and prior density $P(\theta_i)$ (see Chapter 4 for more details).

The prior distribution of the 10 parameters was defined as a uniform distribution whose upper and lower bounds (Table 1) were determined from soil measurements as explained above. The parameters were assumed independent and the likelihood function $P(M_c|\theta_i)$ was assumed normal.

The implementation of the estimation method requires the definition of the sample size *N*. After testing six different values, the value $N = 200\,000$ was selected. Model outputs were simulated with STICS for each generated parameter vector θ_i $(i = 1, \ldots, N)$ and each cell. The likelihood and weight values are then calculated from the model simulations.

2.4. Evaluation of the model predictions

The value of using cell-specific *LAI* and *QN* data for predicting yield and grain protein content was evaluated by calculating root mean squared error (*RMSE*) (see Chapter 2). Two types of predictions were derived with STICS. First, the model parameters were fixed to the mean values of the prior distributions, i.e. to the central values of the intervals displayed in Table 1. In this case, the model returned a single yield value and a single grain protein content value for all the 280 cells. Second, the model parameters were fixed to the mean values of the posterior distributions computed for each cell. In this case, the model predictions were different between cells because the posterior distributions were determined from the remote sensing data. A *RMSE* value was computed for each type of output variable and for each series of predictions.

3. Results and discussion

3.1. Posterior parameter distribution

For each cell, a posterior distribution is obtained for each of the 10 parameters. As an example, Figure 2 presents the posterior distribution for four parameters and three cells.

The results show that the 10 parameters can be divided into two groups:

- "active" parameters have their posterior distributions greatly modified (mean and variance) compared to the uniform prior distribution (e.g. Obstarac and durvieF).

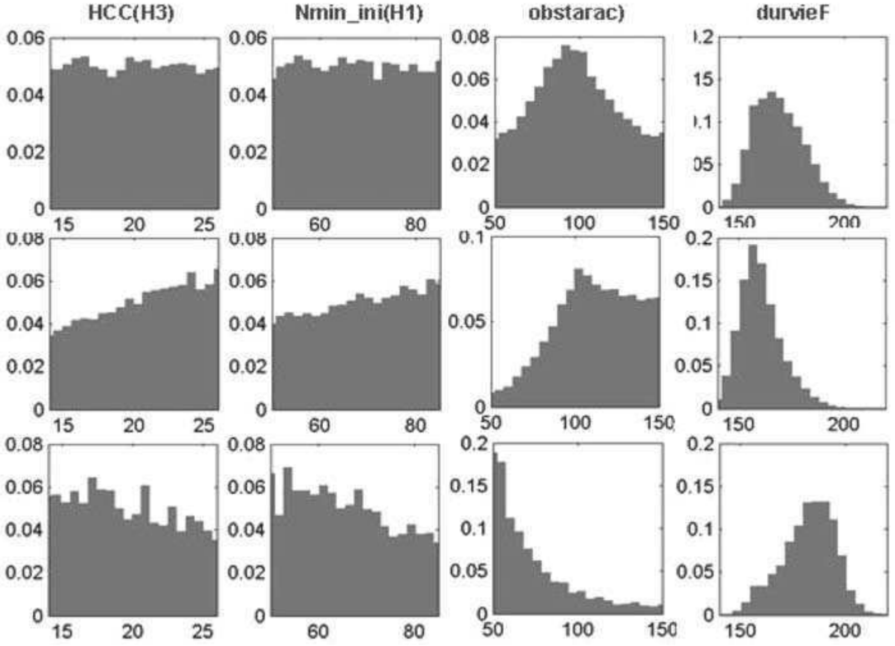


Figure 2. Posterior distribution obtained for four parameters and three cells. The x axis gives the parameter values. The y axis gives the density. Each line corresponds to one cell and each column corresponds to one parameter.

These parameters have a strong effect on the simulated values of the state variables (*LAI, QN*) and on the likelihood;

- "inactive" parameters have a mean value of the posterior distribution very close to the mean value of the prior distribution (e.g. HCC(H_3) and Nmin_ini(H_1)). These parameters do not have a strong influence on the model output and, consequently, their posterior parameter distributions do not differ much from their prior distributions.

For the "active" parameters, the posterior standard deviation is smaller than the prior standard deviation, indicating that the use of the measurements has reduced the uncertainty associated with parameter values.

The means of the posterior parameter distributions can be used as parameter estimates. Figure 3 shows the means of the posterior distributions for four parameters and for the 280 cells of the plot. The mean values of the posterior distributions of the parameters HCC(H_3) and Nmin_ini(H_1) do not vary much among cells. Those parameters can be considered as "inactive" because their posterior means do not differ much from the prior means. On the contrary, considerable variability is observed for the parameters Obstarac and durvieF. The spatial distribution of the estimated values exhibits a spatial structure that reflects the structure observed in the remote sensing images (Fig. 1). Figure 3 also reveals that the estimated values of the parameters Obstarac and durvieF are negatively correlated (high values of durvieF correspond to low values of Obstarac and *vice versa*).

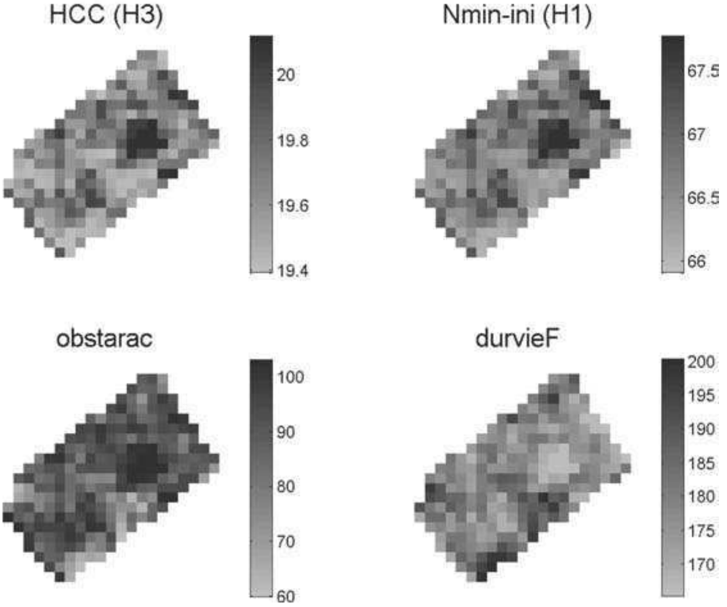


Figure 3. Spatial distribution of four estimated parameters.

3.2. Evaluation of model predictions obtained with and without data assimilation

The accuracies of the two series of model predictions are compared *via RMSE* and *RRMSE* (see Chapter 2). These two criteria are calculated from the yield and grain protein content values measured at harvest for each of the 280 cells of the plot. The results are shown in Table 2.

Grain yield predictions are improved when the parameters are adjusted to cell-specific data; *RMSE* and *RRMSE* are divided by more than 2 when the STICS parameters are fixed to the posterior means. On the contrary, the grain protein content predictions are less accurate with data assimilation than without. This rather surprising result may be due to several reasons:

- grain protein content may be only weakly correlated to the assimilated variables *LAI* and *QN*;

Table 2. Root mean squared error (*RMSE*) and relative root mean squared error (*RRMSE*) for yield and grain protein with and without data assimilation.

Output variable	*RMSE*		*RRMSE*	
	Without data assimilation	With data assimilation	Without data assimilation	With data assimilation
Yield ($t \cdot ha^{-1}$)	1.80	0.72	22.2	8.91
Grain protein content (%)	0.54	1.21	4.95	11

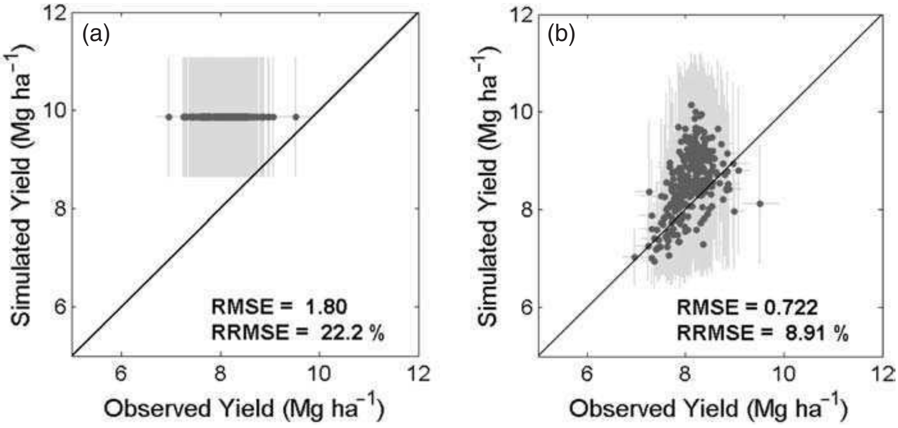


Figure 4. Values of simulated grain yield *versus* observed ones for the 280 cells of the plot (a) before and (b) after assimilation of 4 dates of *LAI* and *QN* data. Each point represents the average simulation for a pixel; the grey bars above and below each point represent ±1 standard deviation of the simulations made with (a) the prior and (b) the posterior distribution of the parameters.

- grain protein content may depend on other parameters than those considered in this study;
- the model equations used to predict grain protein content may be unsatisfactory;
- the method used to estimate the model parameters may not be very efficient: the estimated values that optimize the retrieval of *LAI* and *QN* are not the "true" ones and degrade the simulation of grain protein content.

Beyond the possible reduction of the errors in simulating some output variables (true for grain yield, false for protein content with our model), the method provides spatial distribution of the output variables, whereas the prior information on the parameters gives a single estimate for the whole cells, as illustrated for grain yield in Figure 4.

4. Conclusion

This case study illustrates the potential interest of data assimilation for precision agriculture. We showed how the data provided by remote sensing systems can be used to adjust the model parameters to site characteristics. Data assimilation methods may also be used for assimilating other spatialized observations such as yield maps provided by combine harvesters. This kind of data may be more strongly correlated to the variable predicted at harvest than *LAI* and *QN*.

A number of issues still have to be investigated, notably about the definition of prior parameter distributions, the description of the model and measurement errors, the selection of the parameters to be estimated, the optimization of the number and dates of measurements, and the choice of the statistical method for estimating parameters and/or state variables. In this paper, only some of the model parameters were estimated using a variant of the method GLUE. Other estimation methods could be used. For example, parameters could be estimated by minimizing a cost function taking into account both the data and the prior information about parameter values. The minimization of such a function is difficult for complex models but algorithms were developed for this purpose (e.g. Le Dimet and Talagrand, 1986). These algorithms using the adjoint model for calculating the gradient are being developed for complex models as STICS (Lauvernet, 2005). It would be interesting to compare this approach to the results obtained by updating the state variables with the Ensemble Kalman filter or the particle filter introduced in Chapter 5.

References

Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M.-H., Ruget, F., Nicoullaud, B., Gate, P., Devienne-Barret, F., Antonioletto, R., Durr, C., Richard, G., Beaudoin, N., Recous, S., Tayot, X., Plenet, D., Cellier, P., Machet, J.-M., Meynard, J.-M., Delécolle, R., 1998. STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn. Agronomie 18, 311–346.

Guérif, M., Duke, C., 1998. Calibration of the SUCROS emergence and early growth module for sugar beet using optical remote sensing data assimilation. European Journal of Agronomy 9, 127–136.

Guérif, M., Beaudoin, N., Durr, C., Machet, J.M., Mary, B., Michot, D., Moulin, S., Nicoullaud, B., Richard, G., 2001. Designing a field experiment for assessing soil and crop spatial variability and defining site specific management strategies. Proceedings 3rd European Conference on Precision Agriculture, Montpellier, June, pp. 677–682.

Houlès, V., 2004. Mise au point d'un outil de modulation intra-parcellaire de la fertilisation azotée du blé d'hiver basé sur la télédétection et un modèle de culture. Ph.D. Thesis, INA-PG, Paris, 319 pp. (http://www.inra.fr/ea/theses/Theses2004.html#Houles).

Houlès, V., Mary, B., Guérif, M., Makowski, D., Justes, E., 2004. Evaluation of the crop model STICS to recommend nitrogen fertilization rates according to agro-environmental criteria. Agronomie 24, 1–9.

Lauvernet, C., 2005. Assimilation variationnelle des observations de télédétection dans les modèles de fonctionnement de la végétation: utilisation du modèle adjoint et prise en compte des contraintes spatiales. Ph.D. Thesis, Université Joseph Fourier – Grenoble 1.

Le Dimet, F.X., Talagrand, O., 1986. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. Tellus 38A, 97–110.

Meynard, J.-M., Cerf, M., Guichard, L., Jeuffroy, M.-H., Makowski, D., 2002. Which decision support tools for the environmental management of nitrogen? Agronomie 22, 817–829.

Moulin, S., Guérif, M., Baret, F., 2003. Model inversion procedure for retrieving wheat biophysical variables from hyperspectral measurements. IEEE International Geoscience and Remote Sensing Symposium, Toulouse, France.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 18

Application of Extended and Ensemble Kalman Filters to soil carbon estimation

J.W. Jones and W.D. Graham

1. Introduction **399**
2. The soil carbon model **399**
3. Measurements **400**
4. The Ensemble Kalman Filter (EnKF) **401**
5. The Extended Kalman Filter (EKF) **402**
6. Comparison of Extended and Ensemble Kalman Filters **405**
References **407**

1. Introduction

Variations of the Kalman Filter have been developed for nonlinear models (e.g. Gelb, 1974; Albiol et al., 1993; Graham, 2002). One variation, the Extended Kalman Filter (Gelb, 1974), uses linear approximations of the expressions for propagating the conditional mean of the state and its associated covariance matrix. This approach requires derivation of analytical expressions for propagation based on the first term of a Taylor series expansion. The second approach, the Ensemble Kalman Filter (Burgers et al., 1998; Eknes and Evensen, 2002; Margulis et al., 2002), uses Monte Carlo sampling techniques to generate an ensemble of state variable realizations that are each propagated and updated using the Kalman update equations. The Ensemble Kalman Filter was used by Jones et al. (2004) for optimally estimating soil carbon and a decomposition rate parameter over time for a single field using a nonlinear model. This case study uses that same simple, nonlinear soil carbon model to illustrate the use of the Ensemble and Extended Kalman Filter methods.

2. The soil carbon model

The model has one state variable, the mass of carbon (Z_t, $\mathrm{kg\,ha^{-1}}$) in the top 20 cm of soil in a single field. Changes in soil C are simulated dynamically on an yearly basis. The model also has one unknown parameter, R, the fraction of soil C that is decomposed per year (yr^{-1}). The equations that describe the dynamics of this system are adapted from Jones et al. (2004):

$$\begin{aligned}
&Z_{t+1} = Z_t - R \cdot Z_t + b \cdot U_t + \varepsilon_t \\
&R \sim N(\mu_R, \sigma_R^2) \\
&Z_0 \sim N(\mu_{Z_0}, \sigma_{Z_0}^2) \\
&\varepsilon_t \sim N(0, \sigma_\varepsilon^2)
\end{aligned} \tag{1}$$

where t is time (in years) from an arbitrary starting year when initial values of soil C are known, Z_t is the true soil C in year t, R is the true soil C decomposition rate (yr^{-1}), U_t is the amount of C in crop biomass added to the soil in year t (kg[C]ha^{-1} yr^{-1}), b is the fraction of crop biomass C that is added to the soil in year t that remains after one year (yr^{-1}), and ε_t is a random term representing model error (kg[C] ha^{-1}). This simple model is nonlinear; variables R and Z_t are multiplied in Eq. (1). In this example, we assume that U_t and b are known constants.

There are three sources of uncertainty in this model: uncertainty about the true decomposition rate R, uncertainty about the initial value of soil C, Z_0, and uncertainty in the model structure (ε_t). It is assumed that the model error ε_t, the parameter R, and the initial value Z_0 are normally distributed and uncorrelated. The model error (ε_t) is also assumed uncorrelated with time (i.e. white noise).

3. Measurements

Soil C measurements (m_t) may be made yearly or less frequently, but measurements of R are not possible. Thus, the model has two variables that are to be estimated, but only one is observable. Furthermore, it is assumed that soil C measurement error is normally distributed, independent in time and independent from Z and R. Thus,

$$\begin{aligned}
&m_t = \mu . \varphi_t + \tau_t \\
&\varphi_t = \begin{pmatrix} Z_t \\ R \end{pmatrix} \\
&\mu = (1, 0) \\
&\tau_t \sim N(0, \sigma_m^2)
\end{aligned} \tag{2}$$

where m_t is the measurement of soil C in year t, kg[C] ha^{-1} and τ_t is measurement error, and σ_m^2 is the variance of soil C measurement error. Measurements made from time 0 until the current time t are represented by the vector $M_{1:t}$, where $M_{1:t} = (m_1, m_2, \ldots, m_t)^{\mathrm{T}}$.

Our objective is to estimate $\varphi_t = \binom{Z_t}{R}$ from the measurements $M_{1:t}$. As the model is nonlinear, it is not possible to determine the analytical expression of the posterior distribution of φ_t i.e. the distribution of $\varphi_t|M_{1:t}$. In the next sections, we describe two methods to estimate the expected values and variance–covariance matrix of $\varphi_t|M_{1:t}$.

4. The Ensemble Kalman Filter (EnKF)

The EnKF was already introduced in Chapter 5. This method uses a Monte Carlo approach to generate an ensemble of N realizations of system states, propagating each realization using the stochastic model Eq. (1), and updating each realization at each measurement time (Burgers et al., 1998; Margulis et al., 2002). For the simple soil C model, each ensemble realization consists of value for $\varphi_t = \begin{pmatrix} Z_t \\ R \end{pmatrix}$, denoted by $\varphi_t^j = \begin{pmatrix} Z_t^j \\ R^j \end{pmatrix}$, for the *j*th ensemble member at time t. After updating this ensemble member using the measurement at time t, it is written as $\varphi_{t,K}^j = \begin{pmatrix} Z_{t,K}^j \\ R_K^j \end{pmatrix}$. $\varphi_{t,K}^j$ is related to φ_t^j by

$$\varphi_{t,K}^j = \varphi_t^j + K_t^{\mathrm{e}}\left(m_t^j - \mu\varphi_t^j\right)$$

where $m_t^j = m_t + \tau_t^j$, $\tau_t^j \sim N(0, \sigma_m^2)$, K_t^{e} is a (2×1) vector of Kalman gains defined by $K_t^{\mathrm{e}} = \Sigma_t^{\mathrm{e}}\mu^{\mathrm{T}}(\mu\Sigma_t^{\mathrm{e}}\mu^{\mathrm{T}} + \sigma_m^2)^{-1}$, and Σ_t^{e} is the (2×2) variance–covariance matrix of N state vectors φ_t^j, $j = 1, \ldots, N$. According to the latter equation, we have

$$\begin{aligned} Z_{t,K}^j &= Z_t^j + K_{X,t}(m_t + \tau_t^j - Z_t^j) \\ R_K^j &= R^j + K_{R,t}(m_t + \tau_t^j - Z_t^j) \end{aligned}$$

where

$$K_{Z,t} = \frac{\operatorname{var}(Z_t^j|M_{1:t-1})}{\operatorname{var}(Z_t^j|M_{1:t-1}) + \sigma_m^2},$$

$$K_{R,t} = \frac{\operatorname{cov}(Z_t^j, R^j|M_{1:t-1})}{\operatorname{var}(Z_t^j|M_{1:t-1}) + \sigma_m^2},$$

and τ_t^j is a synthetic random measurement error, drawn from $N(0, \sigma_m^2)$, that represents uncertainty in the *j*th ensemble member measurement (Margulis et al., 2002). $K_{Z,t}$ and $K_{R,t}$ are Kalman gains at time t for Z and R, respectively. Finally, soil C and decomposition rate are estimated after updating all ensemble members by computing the means of $Z_{t,K}^j$ and R_K^j over all N ensemble members. For additional details, see Jones et al. (2004). Note that the EnKF does not have explicit equations for updating the covariance matrix. Instead, this matrix is computed from the ensemble members numerically, prior to updating each ensemble member and then again after updating them to obtain prior and posterior estimates, respectively (see Chapter 5).

5. The Extended Kalman Filter (EKF)

To present the EKF method, it is convenient to express the model Eq. (1) as

$$
\begin{aligned}
\frac{\mathrm{d}Z_t}{\mathrm{d}t} &= -RZ_t + bU_t + \varepsilon_t = f_1(Z_t, R, t) \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= 0 = f_2(Z_t, R, t)
\end{aligned}
\tag{3}
$$

In the Extended Kalman Filter, the differential equations (3) describing the system dynamics are expanded about the current estimate (the conditional means) of the state variables using a Taylor series, assuming that all of the partial derivatives exist. All but the first term in the Taylor series are dropped to create a first-order approximation of the system state equations. The conditional mean of the state and its associated covariance matrix are propagated through time using the first-order (linearized) relationships, and updates are made at measurement times (Gelb, 1974; Welch and Bishop, 2002; Pastres et al., 2003).

Equation (3) describes the two functions (f_1 and f_2) of the soil C model. Equations are needed to propagate the conditional means of system states (Z and R) as well as the variances and covariances of these states. The covariance matrix (P_t) defining these terms is given by:

$$
P_t = \begin{bmatrix} \mathrm{var}(\hat{Z}_t) & \mathrm{cov}(\hat{Z}_t, \hat{R}) \\ \mathrm{cov}(\hat{Z}_t, \hat{R}) & \mathrm{var}(\hat{R}) \end{bmatrix}
\tag{4}
$$

where the "hats" over the variables indicate the best estimates of the conditional mean of the state variables, not the true values. The major challenge of the EKF approach is the development of the set of equations for propagating the covariance matrix between measurement times. This is where the Taylor series is necessary. To implement the Taylor series, we define F as the matrix of partial derivatives of f_1 and f_2 with respect to Z and R:

$$
F = \begin{bmatrix} \dfrac{\partial f_1}{\partial Z} & \dfrac{\partial f_1}{\partial R} \\ \dfrac{\partial f_2}{\partial Z} & \dfrac{\partial f_2}{\partial R} \end{bmatrix} = \begin{bmatrix} -R & -Z \\ 0 & 0 \end{bmatrix}
\tag{5}
$$

The equation to update the state covariance matrix (P) is given by (Gelb, 1974):

$$
\dot{P}_t^- = F \cdot P_t + P_t \cdot F^{\mathrm{T}} Q
\tag{6}
$$

where the (−) superscript on P designates the covariance matrix estimated at time t before a measurement is used to update the estimate. Q is the initial model covariance matrix:

$$
Q = \begin{bmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & 0 \end{bmatrix}
\tag{7}
$$

Equations (5)–(7) are expanded to develop equations for propagating each of the terms in the covariance matrix Eq. (4). The propagation of the expected values (conditional means, $\hat{R}_t$ and $\hat{Z}_t$) is used to obtain best estimates of the state variables. The system of equations for propagating conditional mean and covariance matrix terms between measurement times becomes:

$$\begin{aligned}
\frac{\mathrm{d}\hat{Z}_t}{\mathrm{d}t} &= -\hat{R}\hat{Z}_t + bU_t \\
\frac{\mathrm{d}\hat{R}}{\mathrm{d}t} &= 0 \\
\frac{\mathrm{dvar}(\hat{Z}_t)}{\mathrm{d}t} &= -2[\hat{R}\mathrm{var}(\hat{Z}_t) + \hat{Z}_t\mathrm{cov}(\hat{Z}_t, \hat{R})] + \sigma_\varepsilon^2 \\
\frac{\mathrm{dvar}(\hat{R})}{\mathrm{d}t} &= 0 \\
\frac{\mathrm{dcov}(\hat{Z}_t, \hat{R})}{\mathrm{d}t} &= -\hat{R}\mathrm{cov}(\hat{Z}_t, \hat{R}) - \hat{Z}_t\mathrm{var}(\hat{R})
\end{aligned} \tag{8}$$

The initial conditions for solving this system of equations are assumed to be:

$$\begin{aligned}
\hat{Z}_0 &= \mu_{Z_0} \\
\hat{R} &= \mu_R \\
\mathrm{var}(\hat{Z}_0) &= \sigma_{Z_0}^2 \\
\mathrm{cov}(\hat{Z}_0, \hat{R}) &= 0
\end{aligned} \tag{9}$$

The five equations in Eq. (8) were transformed into discrete difference equations and a time step of one year was used to propagate the variables over time until a measurement was available. Then, all values were updated using the Kalman Gain matrix (K). The following equation was derived to compute K for our example problem, based on Gelb (1974):

$$\begin{pmatrix} K_{Z,t} \\ K_{R,t} \end{pmatrix} = \begin{bmatrix} \mathrm{var}\ (\hat{Z}_t|M_{1:t-1}) & \mathrm{cov}(\hat{Z}_t, \hat{R}|M_{1:t-1}) \\ \mathrm{cov}(\hat{Z}_t, \hat{R}|M_{1:t-1}) & \mathrm{var}(\hat{R}|M_{1:t-1}) \end{bmatrix} \cdot \mu^{\mathrm{T}} \cdot$$

$$\left\{ \mu \cdot \begin{bmatrix} \mathrm{var}\ (\hat{Z}_t|M_{1:t-1}) & \mathrm{cov}(\hat{Z}_t, \hat{R}|M_{1:t-1}) \\ \mathrm{cov}(\hat{Z}_t, \hat{R}|M_{1:t-1}) & \mathrm{var}(\hat{R}|M_{1:t-1}) \end{bmatrix} \cdot \mu^{\mathrm{T}} + \sigma_m^2 \right\}^{-1}$$

where $\mu = (1\ \ 0)$. The conditional notation designates that the variances and covariance at time t are based on all measurements through time $t-1$ (before updates are made).

Substituting variables into the above equations results in the following Kalman gains:

$$
\begin{aligned}
K_{Z,t} &= \frac{\operatorname{var}(\hat{Z}_t|M_{1:t-1})}{\operatorname{var}(\hat{Z}_t|M_{1:t-1}) + \sigma_m^2} \\
K_{R,t} &= \frac{\operatorname{cov}(\hat{Z}_t, \hat{R}|M_{1:t-1})}{\operatorname{var}(\hat{Z}_t|M_{1:t-1}) + \sigma_m^2}
\end{aligned}
\tag{10}
$$

Note that although R is not measured, the measurement of soil C provides information for refining the estimate of R *via* the covariance term. Also note that these gains vary with time; they are recalculated each time a measurement is made. If measurements are not made in a particular year, model predictions provide estimates of soil C and the update step is omitted. When a measurement is made, state variables are updated by:

$$
\begin{aligned}
\hat{Z}_t|M_{1:t} &= \hat{Z}_t|M_{1:t-1} + K_{Z,t}(m_t - \hat{Z}_t|M_{1:t-1}) \\
\hat{R}|M_{1:t} &= \hat{R}|M_{1:t-1} + K_{R,t}(m_t - \hat{Z}_t|M_{1:t-1})
\end{aligned}
\tag{11}
$$

The Kalman Gain variables are used to weight the updated estimate on the basis of error variances. Note that if measurement error variance is very small relative to model prediction variance, then $K_{Z,t}$ approaches 1.0 and the updated model prediction Eq. (11) will be approximately the measured value. In contrast, if measurement error is large relative to prediction error, $K_{Z,t}$ will be closer to 0.0, and the updated soil C estimate will be near the predicted value. Furthermore, if the covariance term used to compute $K_{R,t}$ is small, the updated R will remain close to its estimate from the previous step. However, if the covariance term is large, differences between measured and predicted soil C will result in adjustments to $\hat{R}$ in the update step.

The covariance matrix is updated using the equation (Gelb et al., 1974):

$$
P^{+} = (I - K\mu)P^{-} \tag{12}
$$

where $\mu = (1 \;\; 0)$. The (+) superscript on P designates the covariance matrix after it is updated using measurement at time t. For our problem, the terms in the updated covariance matrix derived from Eq. (12) are:

$$
\begin{aligned}
\operatorname{var}(\hat{Z}_t|M_{1:t}) &= \operatorname{var}(\hat{Z}_t|M_{1:t-1}) \cdot \left[1 - \frac{\operatorname{var}(\hat{Z}_t|M_{1:t-1})}{\operatorname{var}(\hat{Z}_t|M_{1:t-1}) + \sigma_m^2}\right] \\
\operatorname{var}(\hat{R}|M_{1:t}) &= \operatorname{var}(\hat{R}|M_{1:t-1}) - \left[\frac{\operatorname{cov}(\hat{Z}_t, \hat{R}|M_{1:t-1})^2}{\operatorname{var}(\hat{Z}_t|M_{1:t-1}) + \sigma_m^2}\right] \\
\operatorname{cov}(\hat{Z}_t, \hat{R}|M_{1:t}) &= \operatorname{cov}(\hat{Z}_t, \hat{R}|M_{1:t-1}) - \left[\frac{\operatorname{cov}(\hat{Z}_t, \hat{R}|M_{1:t-1}).\operatorname{var}(\hat{Z}_t|M_{1:t-1})}{\operatorname{var}(\hat{Z}_t|M_{1:t-1}) + \sigma_m^2}\right]
\end{aligned}
\tag{13}
$$

These equations were programmed using a discrete time step of one year to implement the EKF for optimal estimation of soil C (Z) and decomposition rate (R) for this problem.

6. Comparison of Extended and Ensemble Kalman Filters

Thirty years of measurements were generated by using Eq. (1) with true values of R and initial soil C (Z_0) to generate true values of Z_t, then randomly sampling from the distribution of τ_t at each annual time step to compute m_t using Eq. (2). Realistic parameters and error terms were selected for examples presented in this chapter and are summarized in Table 1 (from Jones et al., 2004). Two numerical cases are compared. The first case compared EKF and EnKF using the simple model and the values in Table 1 with measurements each year. The second case demonstrates the effects of measurements made at 5-year intervals. Comparisons of Z, R, and their variances are shown.

Figure 1 shows EKF and EnKF estimates of soil organic C for the inputs used (Table 1) as well as annual measurements (generated as discussed above). It also shows the "true" soil C values created as discussed above. Estimates made by both Kalman Filters are smooth and in most years are closer to the "true" values than measured values. Estimates of R decreased from an unconditional mean of 0.020 yr^{-1} to values less than the "true" value of 0.010 after about six years and then converged to its true value after about 20 years. Standard errors of soil C estimates first increased from the initial value of about 141 to over 400, then decreased to about 320 (Fig. 1). This value is less than half of the standard error in estimates if measurements alone were used each year (707 kg[C] ha^{-1}). The standard error of R decreased considerably over time.

When measurements were only made every 5 years, model predictions of soil C each year still followed "true" values very well (Fig. 2). However, standard errors of those estimates were higher, increasing between measurements but decreasing to less than measurement error every time a measurement was available. After twenty years, standard errors of Kalman Filter estimates remained lower than those of measurements alone for all years and was less than 500 kg[C] ha^{-1} in years with measurements. Estimates of R and its standard errors for the 5-year measurement interval were similar to those for 1-year measurement.

Table 1. Values of parameters, initial conditions, and inputs (adapted from Jones et al., 2004).

Variable	Definition	Units	Value
Z_0	Initial estimate of soil C at time 0	kg[C] ha^{-1}	16 000
$\sigma_{Z_0}^2$	Variance of initial soil C estimate	(kg[C] $\text{ha}^{-1})^2$	20 000
R	True value of mineralization parameter	yr^{-1}	0.010
σ_m^2	Variance of measurement, constant over time	(kg[C] $\text{ha}^{-1})^2$	500 000
σ_ε^2	Variance in model estimates of soil C, each year time step	(kg[C] $\text{ha}^{-1})^2$	20 000
μ_R	Unconditional mean of soil C decomposition parameter	yr^{-1}	0.020
σ_R^2	Variance of decomposition rate parameter	yr^{-2}	0.0001
U_t	Input of C to the soil each year (assumed constant)	kg[C] ha^{-1}	2000
b	Proportion of annual soil C input that remains after one year	–	0.20

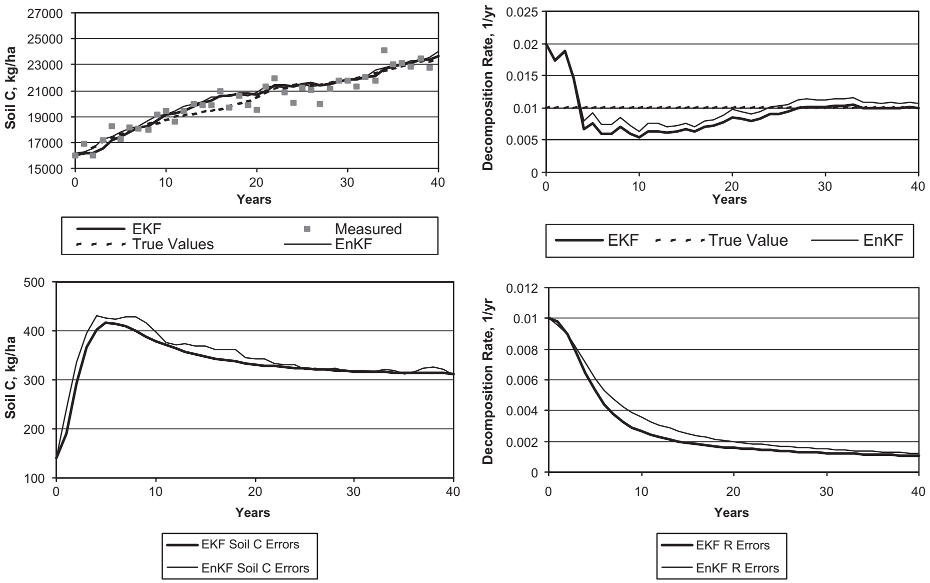


Figure 1. Comparison of EKF and EnKF methods of estimating soil C (upper left figure) and decomposition rate parameter (upper right figure) when measurements are made every year. True soil C and *R* values are also shown. Standard deviations of estimates are compared in the bottom two graphs for soil C and *R*, respectively. All comparisons are based on parameters and initial values in Table 1.

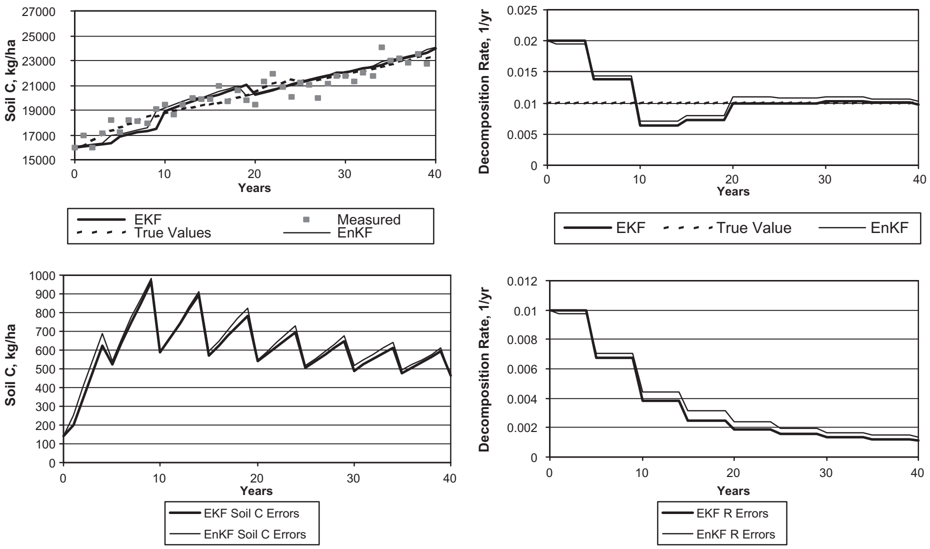


Figure 2. Comparison of EKF and EnKF methods of estimating soil C (upper left figure) and decomposition rate parameter (upper right figure) for the case when measurements are made every five years. True soil C and *R* values are also shown. Standard deviations of estimates are compared in the bottom two graphs for soil C and *R*, respectively.

Results for both EKF and EnKF were similar for the cases shown in Figures 1 and 2 as well as others that are not presented. The advantages of the EKF is in the speed with which computations are made. The computer code was easier to write and debug for EKF *vs.* EnKF. A disadvantage is the requirement of deriving the first order approximation equations for propagating system states and covariance matrix. It is not possible to use this approach on nonlinear models which cannot be written in compact analytical form. Thus, a clear advantage for the EnKF is that one does not have to develop those mathematical relationships. Although the EnKF is slower, one could not detect differences in runtime for this problem.

References

Albiol, J., Robuste, J., Casas, C., Poch, M., 1993. Biomass estimation in plant cell cultures using an extended Kalman filter. Biotechnology Progress 9, 174–178.

Burgers, G., van Leeuwen, P.J., Evensen, G., 1998. Analysis scheme in the ensemble Kalman filter. Monthly Weather Review 126, 1719–1724.

Eknes, M., Evensen, G., 2002. An ensemble Kalman filter with a 1-D marine ecosystem model. Journal of Marine Systems 36, 75–100.

Gelb, A., 1974. Applied Optimal Estimation. MIT Press, Cambridge.

Graham, W.D., 2002. Estimation and prediction of hydrogeochemical parameters using extended Kalman filtering. In: Govindaraju Rao, S. (Ed), Stochastic Methods in Subsurface Contaminant Hydrology. ASCE Publications, Reston, pp. 327–363.

Jones, J.W., Graham, W.D., Wallach, D., Bostick, W.M., Koo, J., 2004. Estimating soil carbon levels using an ensemble Kalman filter. Transactions of the ASAE 47, 331–339.

Margulis, S.A., McLaughlin, D., Entekhabi, D., Dunne, S., 2002. Land data assimilation and estimation of soil moisture using measurements from the Southern Great Plains 1997 field experiment. Water Resources Research 38, 1299–1318.

Pastres, R., Ciavatta, S., Solidoro, C., 2003. The extended Kalman filter as a toll for the assimilation of high frequency water quality data. Ecological Modelling 170, 227–235.

Welch, G., Bishop, G., 2002. An introduction to the Kalman filter, Chapel Hill: Report TR-95–041.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 19

Analyzing and improving corn irrigation strategies with MODERATO, a combination of a corn crop model and a decision model

J.-E. Bergez, J.-M. Deumier and B. Lacroix

1. Introduction **409**
2. The decision model **410**
3. The constraints **412**
4. Uses of MODERATO **413**
5. Conclusions **417**
References **417**

1. Introduction

Given the pressure on water resources, new irrigation scheduling approaches, not necessarily based on satisfying the full crop water requirement but rather aimed at increasing efficient use of the allocated irrigation water and use of stored ground water are needed (Kirda and Kanber, 1999). However, irrigation scheduling is among the most complex of crop management problems. There are in general a number of irrigation dates and that number is not known in advance but depends on circumstances. Whether or not to irrigate on a particular day depends on the state of the crop and of the soil, and thus not only on past irrigation decisions, but also on future irrigation decisions i.e. it is necessary to consider the overall irrigation strategy. Furthermore, irrigation requires resources, namely water, equipment and labor, that are often limiting factors and so it is necessary to take that into account.

The major irrigated crop in France, in terms of area as well as in terms of water use, is corn. The most common irrigation method for corn in France is by traveling gun, which is a high pressure gun mounted on a trolley which is self propelled. The gun irrigates a sector of the assigned area and then must be moved to the next sector. When the entire irrigation area assigned to the gun has been irrigated, a new round of irrigation can begin.

The purpose of MODERATO is to evaluate current irrigation strategies for corn and to propose improved strategies applicable to irrigation by traveling gun as well as to other methods. To achieve this, MODERATO combines a dynamic and biophysical corn crop model with a dynamic decision model (Bergez et al., 2001a). The crop model is described in Wallach et al. (2001) and also in Chapter 12 of this book. It will not be discussed further. The decision model consists of a set of decision rules for different management decisions, in particular, irrigation management decisions. The use of decision rules is important because they allow the decisions to depend on the specific conditions in each field in each year as well as on the level of available resources (Aubry et al., 1998).

The crop model and the decision model interact every day in MODERATO. The crop model updates the state variables each day and passes their values to the decision model together with the explanatory variables that day. Within that collection of variables are the indicator variables of the decision rules. The decision model then evaluates the decision rules to decide if a management action is to be taken. If so the information concerning the decision is passed back to the crop model (e.g. amount of water or sowing density).

The decision model of MODERATO is described in Section 2. Section 3 explains how resource limitations are taken into account. Section 4 describes the different types of study that can be performed using MODERATO are presented in. Section 5 contains conclusions.

2. The decision model

The full set of decisions that are taken into account in MODERATO is presented in Table 1. Table 1 also gives an example of each decision rule as used in a specific study by Bergez et al. (2002). The following discussion only concerns the decision rules for irrigation since MODERATO is particularly oriented toward irrigation decisions. These decision rules were developed in collaboration with irrigation engineers. Similar rules have been used in the past (Leroy et al., 1996, 1997).

The timing of irrigation is determined by five separate decision rules:

(1) Sowing rule. This rule determines whether or not to apply irrigation shortly after sowing and the amount of water to be applied.
(2) Starting rule. This rule determines the day on which to begin irrigation during the growing season and the amount for the first irrigation round.
(3) Next round rule. This rule is invoked after a round of irrigation has been terminated. It determines when to start the next round and the irrigation amount for rounds after the first.
(4) Climatic events rules. Irrigation can be temporarily suspended after a rainfall event or due to strong wind. This rule determines the length of the interruption.
(5) Stopping rule. This rule is invoked at the end of an irrigation round. It has one of three conclusions. Either the previous round of irrigation was the last, or another round of irrigation is to be performed and that will be the last, or another round of irrigation is to be performed and when finished this rule will be invoked again. If the next round is the last, the amount of irrigation is given.

Many of the rules in MODERATO are based on the general form:

IF (*condition 1a* OR *condition 1b*) AND (*condition 2a* OR *condition 2b*) THEN *decision; define amount.*

where conditions 1a and 1b concern crop development while conditions 2a and 2b refer to water status in the soil. The first condition in each pair (conditions 1a and 2a) uses meteorological variables as indicator variables while the second condition in each pair is based on state variables. The user can choose to ignore one of the two conditions in each part of the premise.

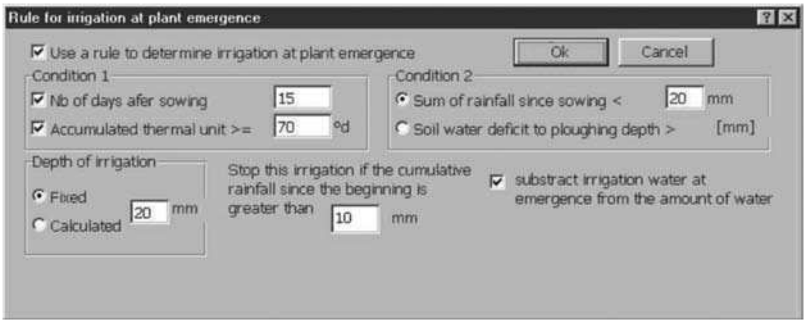
Table 1. The management decisions in MODERATO, with an example of decision rule for each decision.

Decision	Rules
Sowing date	IF (*April* 20 $<$ *date* $<$ *May* 30) AND (*cumulative rainfall during the previous* 3 *days* $<$ 15 *mm*) AND (*crop not yet sown*) THEN *sow*. IF (*date* $=$ *May* 30) AND (*crop not yet sown*) THEN *sow*
Sowing density	*Density* $=$ 80 000 *plants ha*$^{-1}$
Variety	*Variety* $=$ *Cécilia*
Fertilisation date	*Fertilisation date* $=$ *sowing date*
Fertilisation amount	*Fertilisation amount* $=$ 200 *kg N ha*$^{-1}$
Harvest date	IF (*grain moisture content* $<$ 20% OR *cumulative thermal units from sowing* $\geq$ 2100°C *days*) AND (*cumulative rainfall during the previous* 3 *days* $<$15 *mm*) AND (*crop not yet harvested*) THEN harvest. IF (*date* $=$ 14 *October*) AND (*crop not yet harvested*) THEN *harvest.*
Irrigation	Sowing rule: IF (*time after sowing* $=$ 15 *days* OR *degree days after sowing has just passed* 70°C *days*) AND (*cumulative rainfall since sowing* $<$ 20 *mm*) AND (*no irrigation after sowing yet applied*) THEN *irrigate with* 20 *mm*. Starting rule: IF (*date* $\geq$ *June* 15) AND (*cumulative rainfall in previous* 5 *days* $<$ 15 *mm*) AND (*cumulative potential evapotranspiration in previous* 5 *days* $>$ 15 *mm*) THEN *begin first irrigation round. Irrigation amount in first round* $=$ 30 *mm*. Next round rule: IF (*development stage* $<$ *flowering*) AND (*days since start of last round* $=$ 9) THEN *start next round.* IF (*flowering* $\leq$ *stage* $\leq$ 50% *grain moisture content*) AND (*days since start of last round* $=$ 6) THEN *start next round.* IF (*development stage* $>$ 50% *grain moisture content flowering*) AND (*days since start of last round* $=$ 9) THEN *start next round. Irrigation amount in rounds after first* $=$ 30 *mm.* Stopping rule: IF (*date* $=$ 1 *September*) AND (*cumulative rainfall over previous* 5 *days* $<$15 *mm*) AND (*cumulative potential evapotranspiration over previous* 5 *days* $<$20) THEN *do one more irrigation round after present round.* OTHERWISE *irrigation ceases*. Rainfall rule: IF (*cumulative rainfall over previous* 5 *days* $>$ 15 *mm*) THEN *no irrigation for min* (5, *cumulative rainfall over previous* 5 *days* (*in mm*)/4) *days*.

An example of a rule for irrigation after sowing is (see Fig. 1)

IF (*time after sowing* $=$ 15 days OR *cumulative degree days after sowing has just passed* 70°C *days*) AND (*cumulative rainfall since sowing* $<$ 20 *mm*) THEN *irrigate*.

The rule also specifies the amount of water to apply. The amount can either be a fixed quantity (for example, 20 mm for the rule defined in Fig. 1) or it can be calculated based on soil water depletion. This rule has an additional section which says that the irrigation round after sowing is abandoned if cumulative rainfall since the beginning of the round exceeds 10 mm before all sectors have been irrigated. Finally, in the case of irrigation after sowing, one can decide whether or not the total available water includes the water used for this round or not.



☑ Use a rule to determine irrigation at plant emergence

Condition 1
☑ Nb of days afer sowing 15
☑ Accumulated thermal unit >= 70 °d

Condition 2
● Sum of rainfall since sowing < 20 mm
○ Soil water deficit to ploughing depth > [mm]

Depth of irrigation
● Fixed 20 mm
○ Calculated

Stop this irrigation if the cumulative rainfall since the beginning is greater than 10 mm

☑ substract irrigation water at emergence from the amount of water

Figure 1. Graphical user interface for defining the sowing rule in MODERATO that determines whether or not to irrigate at sowing.

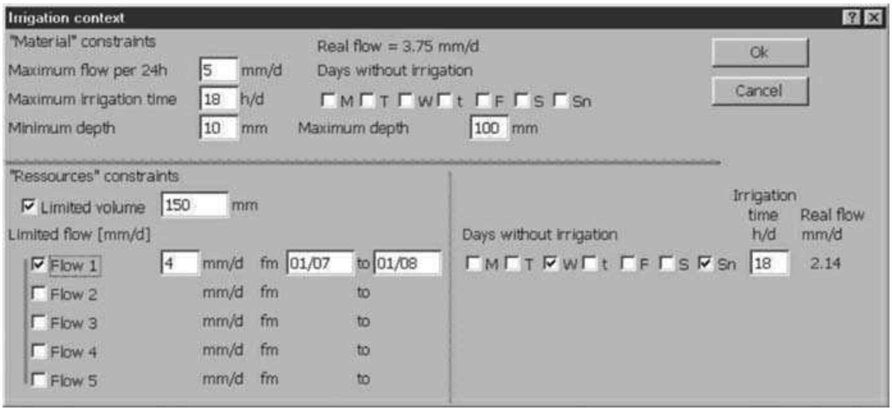
Note that the decision model does not have total flexibility but rather has a built-in structure adapted to the specific irrigation context in question. Flexibility arises from the possibility of using either one or both the conditions 1a and 1b combined with either one or both the conditions 2a and 2b. Also the choice of parameter values for the decision rules allows one to test a range of decision rules. However, we did not opt for total flexibility, which could be achieved by allowing the user to freely define his own decision rules (Shaffer and Brodahl, 1998). Instead, we chose a compromise between flexibility on the one hand and ease-of-use in inputting the decision rules on the other. As Figure 1 shows, the user can input decision rules quite easily by way of the graphical user interface.

The decision rules in MODERATO are meant to be compatible with actual practice but are not meant to mimic farmer behavior completely. Indeed, although surveys in south western France have shown that some farmers implicitly use decision rules similar to those given above, the indicator variables that they use may be quite different from those of MODERATO. For example, responses to a question about the basis for the start of irrigation included “when leaves start to roll”, “when the soil changes color”, “when tensiometer values reach a given threshold”, “when the neighbor starts to irrigate”, “when the soil dug up by moles is dry” etc. These indicator variables are not used by MODERATO and furthermore are not available from the crop model.

3. The constraints

There are various constraints that affect irrigation which can be taken into account with MODERATO (see Fig. 2). These are:

- Maximum flow rate. Flow rate may be limited by the available equipment, by the contract with the water provider or by regulation. Note that flow rate and amount to apply determine the time necessary to irrigate a sector, and thus also the minimum time for a full round of irrigation. In MODERATO, one can specify both an equipment limitation on flow rate and also lower flow rates for certain periods, such as might be imposed by a water provider.
- Total available water. Total available water may be limited by contract, or because the farmer is pumping from a reservoir of limited capacity.



"Material" constraints

Maximum flow per 24h 5 mm/d

Maximum irrigation time 18 h/d

Minimum depth 10 mm

Real flow = 3.75 mm/d

Days without irrigation

☐ M ☐ T ☐ W ☐ t ☐ F ☐ S ☐ Sn

Maximum depth 100 mm

"Ressources" constraints

☑ Limited volume 150 mm

Limited flow [mm/d]

					Days without irrigation	Irrigation time h/d	Real flow mm/d
☑ Flow 1	4 mm/d	fm	01/07	to 01/08	☐ M ☐ T ☑ W ☐ t ☐ F ☐ S ☑ Sn	18	2.14
☐ Flow 2	mm/d	fm		to			
☐ Flow 3	mm/d	fm		to			
☐ Flow 4	mm/d	fm		to			
☐ Flow 5	mm/d	fm		to			

Figure 2. Graphical user interface for defining the constraints to be taken into account in MODERATO.

- Available time for irrigation. In case of drought, irrigation can be prohibited on certain days. Labor may also not be available on certain days, or may only be available for a certain number of hours each day. The effect is to reduce the time during which the irrigation equipment can operate and thus to increase the minimum time for a full round of irrigation.
- One can specify a maximum and a minimum amount of irrigation water to be applied in a round.

4. Uses of MODERATO

4.1. Evaluating a strategy

To evaluate an irrigation strategy with MODERATO the user specifies the strategy by using the graphical user interface to define the decision rules (or if an expert user by using specific text files). MODERATO can then be used to simulate the results of that strategy for a series of climate scenarios or soil types. Companion tools have been developed to help analyze the results of the strategy for those climates. One tool creates a table giving the mean, standard deviation, minimum value, maximum value and the second and forth quartiles for grain yield, amount of water applied in irrigation, amount of water lost by drainage, water use efficiency, yield loss compared to yield in the absence of water stress, first and last days with irrigation and a measure of profit. The tool also indicates extreme results, which one might want to analyze in more detail. Other tools provide graphical representations of the results. One graph gives cumulative frequencies for grain yield, profit and starting and ending irrigation day. A second graph shows, for each climate, actual and potential grain yield, the amount of water applied in irrigation and profit (Fig. 3). A third graph displays soil water deficit on the first and last days of irrigation.

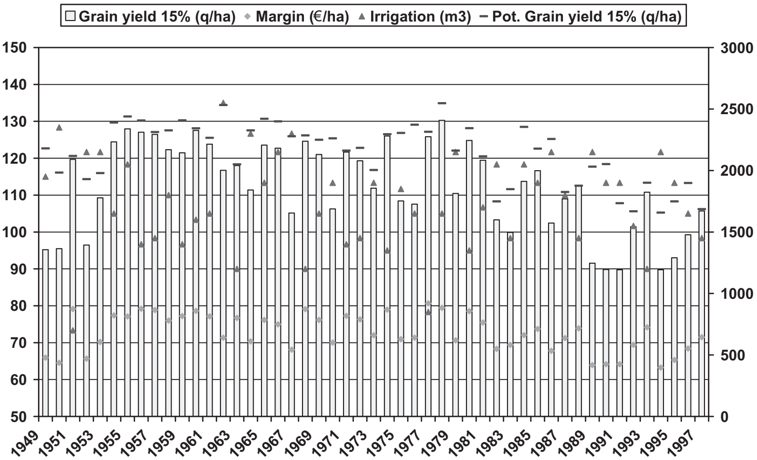


Figure 3. Graphical representation provided by MODERATO of results over a series of climates.

4.2. Analyzing the results for a specific climate

A set of 5 graphs is available to analyze in detail the results of an irrigation strategy for a particular climatic year and soil. The first graph provides a graphical representation of the climate variables. The second graph gives details of soil water dynamics. The third graph displays plant variables *versus* time. The fourth graph displays information about the irrigation applications. The last is a calendar, which shows the starting date of each round of irrigation, days when no irrigation is applied because of recent rainfall, etc (Fig. 4).

4.3. Providing parameters for tensiometer-based irrigation

A set of recommendations for corn irrigation based on measurements of soil moisture using tensiometers has recently been developed in France. The overall method is known as IRRINOV®MAIS (Deumier et al., 2002). The recommendations are based on thresholds of soil moisture. If soil moisture falls below the threshold, irrigation should be applied. The method proposes different thresholds depending on soil type, stage of crop development and level of irrigation equipment.

The thresholds are to a large extent based on experimental results. However, experimentation is limited as to the number of contexts that can be tested. MODERATO is therefore being used as a second approach to determining the thresholds. For this, an equation is required which converts between tensiometer values and soil moisture. The decision rules in MODERATO allow one to test different soil water thresholds for irrigation at various stages, and the conversion equation converts those thresholds into tensiometer values.

4.4. Optimizing decision rules

One method of seeking improved decision rules with MODERATO is by trial and error. One creates and evaluates different strategies in order to find strategies better than the

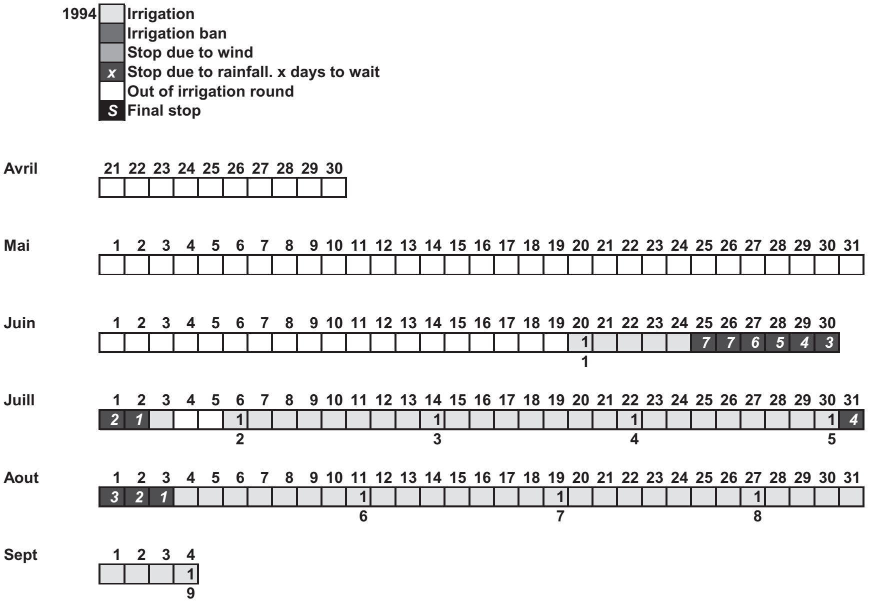


Figure 4. Graphical representation of the calendar of irrigation decisions for a particular context provided by MODERATO.

initial strategy. The tools of MODERATO for analyzing a strategy can be very valuable here in suggesting what changes might lead to improved strategies.

When one seeks only simple changes in the decision rules, it might be possible to do a systematic search. For example, different parameters related to the soil moisture thresholds for starting the first and subsequent rounds of irrigation were tested (Bergez et al., 2002). The other decision rules were kept constant (Fig. 5).

For more extensive changes compared to the initial strategy, automated optimization algorithms as described in Chapter 6 are necessary. One such algorithm was used to optimize the 8 parameters in the following decision rules (Bergez and Garcia, 2002; Bergez et al., 2001b):

(1) Starting rule. IF (*cumulative degree day after sowing* > ***T*1**) AND (*soil water deficit* > ***D*1**) THEN *begin first round of irrigation. Irrigation amount in first round* = ***I*1**.
(2) Next round rule. IF (*soil water deficit* > ***D*2**) THEN *start a new round of irrigation. Irrigation amount in rounds after first* = ***I*2**.
(3) Stopping rule. IF (*cumulative degree day after sowing* > ***T*3**) AND (*soil water deficit* > ***D*3**) THEN *do a last round of irrigation*. OTHERWISE *stop irrigation. Irrigation amount if last round is applied* = ***I*3**.

4.5. Evaluating heterogeneity between sectors

The irrigated area is not irrigated simultaneously but rather sector by sector. The results will then differ between sectors and the variability will be greater, the longer it takes

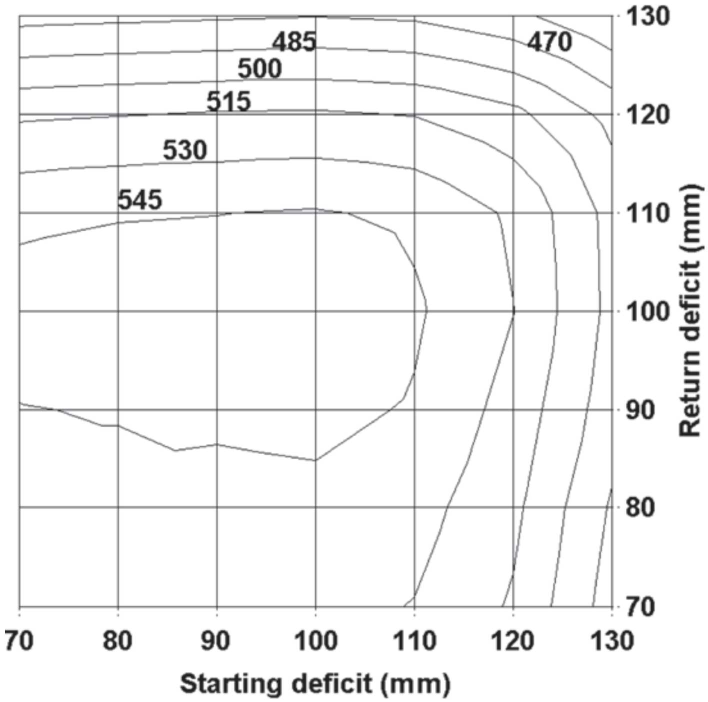


Figure 5. Net margin (€ ha^{-1}) iso-contours as a function of two decision rule parameters: soil water deficit for starting the first round of irrigation (starting deficit) and soil water deficit for subsequent rounds (return deficit).

to complete a round of irrigation. The previous results refer to the first sector, but MODERATO can also be used to analyze the variability between sectors. Bergez and Nolleau (2003) found a corn grain yield difference of 1.41 Mg ha^{-1} between the highest and lowest yielding sectors on the average over a series of climates. The greatest yield difference between sectors was 2.11 Mg ha^{-1}. Yield variability decreases as flow rate increases, as irrigation amount decreases, as soil depth increases or as gravimetric soil-available water capacity increases. The first two variations (increased flow rate, decreased amount) decrease the time necessary to finish a round of irrigation and therefore the delay between the first and last sectors. The last two variations (increased soil depth, increased water capacity) increase maximum soil water storage and therefore make the crop less dependent on irrigation timing.

4.6. Real-time use of MODERATO

The specific real-time use of MODERATO that is being developed concerns the use of short-term weather predictions. Five-day or even longer predictions are now available, and should certainly be useful in determining irrigation management. Real-time use of MODERATO means that one inputs actual climate and management decisions up to the day when MODERATO is run. MODERATO can then be used in real time to compare two different decisions, namely irrigate this day or not. This use does not involve decision rules, but simply looks at the consequences of each decision over a one-week period.

4.7. MODERATO as a diagnostic aid

The main objective of MODERATO is to test irrigation management strategies. However, the companion tools are also useful for analyzing the results of past management. In this case one inputs actual decisions rather than decision rules.

5. Conclusions

Coupling decision rules with a crop model allows one to evaluate management strategies and to propose improved decisions. As MODERATO shows, this is possible and useful even for complex management decisions such as those concerning irrigation. The structure of the decision rules is very important, since that determines the range of strategies that can be tested. In MODERATO, the structure of the decision rules is largely dictated by the specific irrigation context, but there is still quite a bit of latitude in the rules. Defining decision rules that are both flexible and applicable is an area where further research is required.

We have concentrated here on the decisional part of MODERATO. However, in using models as an aid in crop management, both the decision model and the crop model are essential.

References

Aubry, C., Papy, F., Capillon, A., 1998. Modelling decision-making processes for annual crop management. Agricultural Systems 56, 45–65.

Bergez, J.-E., Garcia, F., 2002. A hierarchical partitioning method for optimizing irrigation strategies. EWDA-02 Workshop on Sequential Decisions under Uncertainty in Agriculture and Natural Resources, Toulouse, France, September 19–20, pp. 39–44.

Bergez, J.-E., Nolleau, S., 2003. Maize grain yield variability between irrigation stands: a theoretical study. Agricultural Water Management 60, 43–57.

Bergez, J.-E., Debaeke, P., Deumier, J.-M., Lacroix, B., Leenhardt, D., Leroy, P., Wallach, D., 2001a. MODERATO: an object-oriented decision tool for designing maize irrigation schedules. Ecological Modelling 137, 43–60.

Bergez, J.-E., Eigenraam, M., Garcia, F., 2001b. Comparison between dynamic programming and reinforcement learning: a case study on maize irrigation management. EFITA 2001, Third European Conference of the European Federation for Information Technology in Agriculture, Food and the Environment, Montpellier, France, June 18–20, pp. 343–348.

Bergez, J.-E., Deumier, J.-M., Lacroix, B., Leroy, P., Wallach, D., 2002. Improving irrigation schedules by using a biophysical and a decisional model. European Journal of Agronomy 16, 123–135.

Deumier, J.-M., Bouthier, A., Bonnifet, J.-P., Mangin, M., Lacroix, B., Renoux, J.P., 2002. Irrinov® maïs : une méthode pour bien irriguer le maïs. Perspectives Agricoles 278, 76–83.

Kirda, C., Kanber, R., 1999. Water, no longer a plentiful resource, should be used sparingly in irrigated agriculture. In: Kirda, C., Moutonnet, P., Hera, C., Nielsen, D.R. (Eds), Crop Yield Responses to Deficit Irrigation, Kluwer, Dordrecht, pp. 1–20.

Leroy, P., Balas, B., Deumier, J.-M., Jacquin, C., Plauborg, F., 1996. Water management at farm level¸ CAMAR 8001-CT91 Report. The Management of Limited Resources in Water, pp. 90–150.

Leroy, P., Deumier, J.-M., Jacquin, C., 1997. IRMA: un simulateur de l'organisation des chantiers d'irrigation. Perspectives Agricoles 228, 76–83.

Shaffer, M.J., Brodahl, M.K., 1998. Rule-base management for simulation in agricultural decision support system. Computer and Electronics in Agriculture 21, 135–152.

Wallach, D., Goffinet, B., Bergez, J.-E., Debaeke, P., Leenhardt, D., Aubertot, J.N., 2001. Parameter estimation for crop models: a new approach and application to a corn model. Agronomy Journal 93, 757–766.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Chapter 20

Managing wheat for ethanol production: a multiple criteria approach

C. Loyce, J.P. Rellier and J.M. Meynard

1. Introduction 419
2. The strategies defined in BETHA 420
3. The crop model 422
4. Criteria for judging bio-fuel wheat production systems 422
5. Multiple criteria analysis with BETHA 423
6. Results 425
7. Discussion 425
References 426

1. Introduction

The European Union has an obligatory fallow policy, but since 1993 it has been legal to consider non-food crops as fulfilling the fallow requirement. Agricultural professionals promote such production as a way of increasing farmer revenue. Among non-food crops are crops for bio-fuels and in particular wheat for ethanol production.

Current production of bio-fuels is very low, despite the fact that they offer a diversification of both farmer income and energy sources. A number of problems with bio-fuels have been raised by critics. In France they are deemed too costly for society since unlike fossil fuels they are tax exempt. Without this advantage they would not be competitive. Environmentalists argue that although bio-fuels reduce CO_2 emissions compared to fossil fuels, this is offset by the fact that intensive wheat production causes water and air pollution. Finally, critics point out that bio-fuel production must be judged according to net energy production which is the energy in the ethanol minus the energy required for producing it.

Some of the criteria for judging wheat production for ethanol clearly are not the same as for wheat production for food. For example, the energy balance is essential for the former but is far less relevant for the latter. In addition, the economic context differs: the price of ethanol wheat is lower than that of regular wheat (about 76€/ton *versus* 122€/ton in 1995). This suggests that wheat production for ethanol may require very different production systems than those commonly used. The purpose of this study was to generate and evaluate a large number of production systems, in order to identify those that are adapted to wheat production for ethanol (Loyce et al., 2002a,b). The study is specifically adapted to wheat production for ethanol in the Champagne Crayeuse region in north-east France.

In order to screen management strategies we developed a decision support tool called BETHA (for "blé éthanol"). The steps involved in using this tool are as follows:

(1) Define the different management decisions that make up an overall management strategy for wheat, and list the possible modalities for each decision (Section 2).
(2) Define a series of climates representative of those that might be encountered. Here we used past climates (from 1978 to 1996) from the region in question.
(3) Define a crop model, which determines the results for each combination of management strategy and climate (Section 3).
(4) Generate all possible combinations of management decisions and eliminate those that are unrealistic. Associate with each management strategy and climate combination the results of the crop model. The algorithm used for this step is based on the Constraint Satisfaction Problem (CSP) approach (Mackworth, 1987; Schiex, 1993; Tsang, 1993). This approach has been proposed previously for the design of wheat crop management plans (Martin-Clouaire and Rellier, 1995).
(5) Define the criteria for evaluating a given management strategy for a given climate (Section 4).
(6) For each management strategy – climate combination, use a multiple criteria evaluation method in order to classify the results as "good" or "bad" (Section 5).
(7) Select management strategies suited to the requirements of ethanol wheat production. A management strategy is selected if a sufficient fraction of management strategy – climate combinations for that strategy is classified as "good" and not classified too often as "bad" (Section 5).

Some characteristics of wheat management strategies suited to bio-fuel production are presented in Section 6. Section 7 contains a discussion.

2. The strategies defined in BETHA

The management decisions and the possible modalities for each are shown in Table 1. The decisions concern sowing period and density, cultivar, fungicide and insecticide protection, nitrogen fertilization and the use of a growth regulator. We did not consider different possible weed control or tillage decisions, since these decisions should be determined at the crop rotation level and not on the basis of a single cropping season. The present version of BETHA assumes standard tillage and complete herbicide protection. It is thus assumed that weeds are under control and that soil structure does not hamper crop emergence. Finally, as winter wheat is normally grown without irrigation in France, only non-irrigated systems are considered. For some of the decisions, such as cultivar, the possibilities are completely specified. For others, such as fungicide protection, the possibilities are not completely defined and the actual value will be based on decision rules (see Chapter 6). Actual type of fungicide treatments will then depend on climatic conditions which determine the risk of disease. A management strategy is a combination of modalities, one for each decision.

Some combinations of management decisions are unrealistic. For example, one would not combine low density and nitrogen fertilization with the use of a growth regulator. The CSP algorithm eliminates such combinations.

Table 1. Management decisions and possible values.

Decision	Sowing density (SD)	Sowing period (SP)	Insecticide treatment (INS)	Cultivar (CULT)	Nitrogen fertilisation (FER) (Nitrogen dose Dn × Number of applications AP)	Fungicide protection (FUN)	Growth regulator (REG)
Modality							
	A: regional reference (the sowing density increases with delayed sowing period)	SP1: 1–10/10	If SP = SP1 or SP2 then treatment in autumn, else no treatment in autumn	Apollo	Dn	F0: No treatment	If (SD = B and SP = SP5) or if [SD = B and (SP = SP2 or SP3 or SP4) and AP = AP1] then growth regulator else no growth regulator
	B: reduced by 60% compared with A	SP2: 10–20/10		Arche	$X - 180$	F11: reduced protection, with one treatment (morpholine + triazole)	
		SP3: 20–31/10		Balthazar	$X - 160$	F12: reduced protection, with one treatment (triazole)	
		SP4: 1–10/11		Beaver	–	F2: reduced protection, with two treatments (morpholine + triazole)	
		SP5: 10–20/11		Charly	X	F31: preventive protection, with three treatments (morpholine + triazole)	
				Delfi	–	F32: preventive protection, with three treatment (morpholine + triazole + prochloraze)	
				Estica	$X + 180$		
				Euréka	X is calculated using the balance sheet method for a yield objective corresponding to the fourth quantile (upon five) of the distribution of potential yields		
				Forby	AP		
				Gaspard	AP1: one application (Dn $\leq X - 40$) or (Dn $> X - 40$ and Dn < 40)		
				Junior	AP2: two applications (Dn $\geq X + 40$) or (Dn $> X - 40$ and Dn < 80)		
				Renan	AP3: three applications (Dn $\geq X + 40$)		
				Rialto			
				Ritmo			
				Scipion			
				Soissons			
				Trémie			
				Tribun			

3. The crop model

The crop model is not dynamic but rather a static model that consists of simple relationships that link crop results (yield and quality) to crop characteristics, to the environment (soil, climate) and to crop management (Fig. 1). An example of a model equation is $P_0 = p_1(Q_n/Y_r) + p_2$, where P_0 is grain protein content, Q_n is nitrogen absorbed by the crop, Y_r is yield and p_1 and p_2 are parameters.

First, the choice of a static model was made because there exist in the literature static relationships that take into account the effects of the major factors (nitrogen and disease according to Meynard et al. (1981)) that influence wheat production in the north and center of France (Rémy and Viaux, 1982; Meynard, 1985, 1991; Spiertz and Vos, 1985; Bergström and Brink, 1986; Chaney, 1990; van Keulen and Stol, 1991; Gate, 1995; Richards et al., 1996; Meynard et al., 1997; Makowski et al., 1999). Second, the CSP algorithm is not adapted to the use of dynamic models.

4. Criteria for judging bio-fuel wheat production systems

Different stakeholders (farmers, industry, government, conservationists) have different criteria for judging management strategies and results for bio-fuel wheat. Overall, 8 different criteria were used to evaluate management strategies, based on discussions with all stakeholder groups.

(1) Semi-net profit to the farmer, defined as value of the crop minus the costs of inputs, manpower and equipment operations. Fixed expenses such as rents, taxes, salaries, etc. are not taken into account.

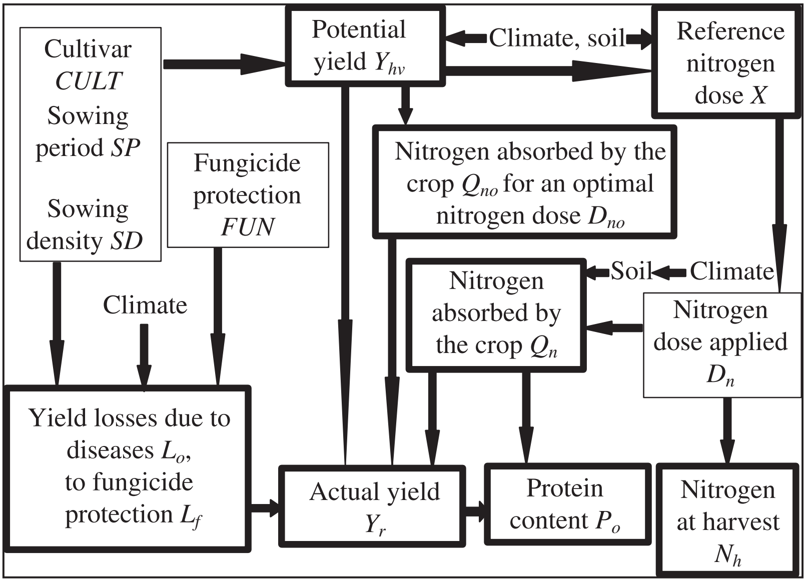

Cultivar $CULT$
Sowing period SP
Sowing density SD
Potential yield Y_{hv}
Climate, soil
Reference nitrogen dose X
Fungicide protection FUN
Nitrogen absorbed by the crop Q_{no} for an optimal nitrogen dose D_{no}
Soil
Climate
Climate
Nitrogen absorbed by the crop Q_n
Nitrogen dose applied D_n
Yield losses due to diseases L_o, to fungicide protection L_f
Actual yield Y_r
Protein content P_o
Nitrogen at harvest N_h

Figure 1. General framework of the crop model.

(2) Number of interventions. Since the value of the crop per hectare is low, the time devoted to the crop must be low.

(3) Production cost per ton. It is important to lower the costs so that government subsidies could be reduced.

(4) and (5) Impact of the crop on the environment. The two criteria we use are residual mineral N left in the soil after harvest and the volume of active ingredients in pesticides applied to the crop. The last criteria has been used by Wijnands (1997). The use of a simple pesticide criterion is reasonable because the production systems that we will compare use the same type of active ingredients and the same application methods for pesticides. If required to compare production systems with a variety of active ingredients or application methods, we need more complex criteria (Van der Werf and Zimmer, 1998).

(6) The net energy, defined as the energy produced (tons of wheat per hectare times ethanol yield per ton of wheat times energy content of ethanol) minus the energy used for production (including energy for crop production and transformation as well as for packaging by-products and transportation).

(7) and (8) Quality of the cakes for animal feed which are by-products of ethanol production. The quality determines the economic value of the by-product. Two specific criteria are used here, namely protein percentage (too low a value reduces the nutrient value of the cakes, while too high a value makes the cakes stick together too much) and grain hardness, which depends on the wheat variety. With a low hardness value it is easier to separate the bran and kernel and grinding into flour requires less energy (Abecassis, 1993).

5. Multiple criteria analysis with BETHA

A number of different methods for taking multiple criteria into account in decision making have been proposed (Schärlig, 1985). The simplest is to aggregate the different criteria into a single criterion. Often the combined criterion is a weighted sum of the various individual criteria (Charnes and Cooper, 1977; Brans, 1983; Morgan et al., 1989). In the present case however this approach would be difficult to apply. First of all, the different criteria are on different scales and it would be difficult to combine them. Second, different stakeholders are interested by different criteria, so that a high value for one criterion in many cases cannot compensate for a low value for another criterion. For example, revenue is important to the farmer while energy balance is important to the government and to conservationists. For neither group will a high value for one of these criteria compensate for a low value for the other.

The multiple criteria method we use here is based on the concepts of concordance and discordance (Roy, 1985; Roy and Bouyssou, 1993). A management strategy *a* is preferred to strategy *b* if a majority of the criteria are concordant with the proposition "*a* is preferred to *b*" and if there are no criteria that are so discordant as to invalidate this proposition.

In BETHA we use a version of this approach to classify management strategies into one of the two categories called "good" and "bad," for a given climate (Perny, 1998). Then, depending on the number of climates for which it is "good" and "bad," a management strategy is selected or not. The classification proceeds as follows:

(1) Define for each category K_i (K_1 = "good", K_2 = "bad") and for each criterion X_j (X_1 = semi-net profit, X_2 = number of interventions, etc.) the concordance and

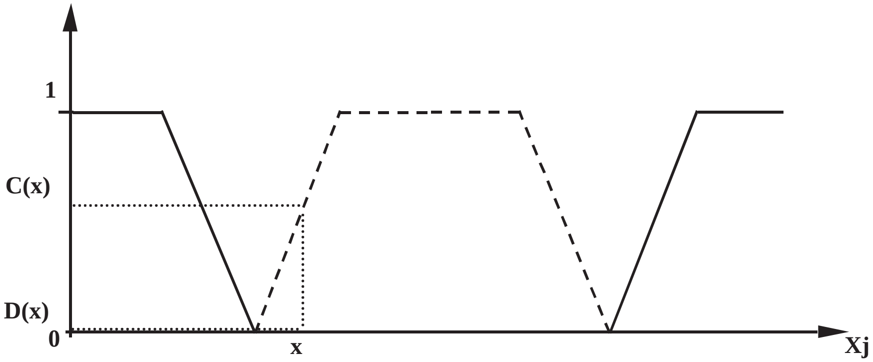


Figure 2. Concordance function (dashed line) and discordance function (solid line). The dotted vertical line indicates a particular result for this criterion, which translates into a concordance index $C(x)$ and a discordance index $D(x)$.

discordance functions $C_{i,j}(\cdot)$ and $D_{i,j}(\cdot)$ (Fig. 2). Suppose for example, that the value of criterion X_j for the management strategy a and for a given climate is $x_{j.a}$. Then the concordance value $C_{\text{good},j}(x_{j.a})$ measures how much we agree with the statement "a is a good strategy as far as criterion X_j is concerned" and the discordance value $D_{\text{good},j}(x_{j.a})$ measures how much we disagree with this statement. Similarly $C_{\text{bad},j}(x_{j.a})$ and $D_{\text{bad},j}(x_{j.a})$ indicate our agreement or disagreement with the proposition that a is "bad" with respect to criterion X_j. To simplify somewhat, we set $D_{\text{bad},j}(x_{j.a}) = C_{\text{good},j}(x_{j.a})$ and $C_{\text{bad},j}(x_{j.a}) = D_{\text{good},j}(x_{j.a})$. The concordance and discordance functions express our subjective ideas about what values are good or bad. A category K_i has the nature of a fuzzy set, i.e. to which strategies more or less belong. That reflects the imprecise nature of the stakeholders' criteria, which makes it unrealistic to set sharp boundaries between categories.

(2) Define for each category K_i and for each criterion X_j, the absolute veto value $v_{i,j}$, such that if the criterion value x goes beyond, for instance, $v_{\text{good},j}$ then $D_{\text{good}}, j(x) = 1$ resulting in the strategy in question having no chance to be a "good" strategy, regardless of the results for the other criteria. On the other side of $v_{i,j}$, the veto strength may decrease gradually. For example, a strategy that results in net energy production below 3000 MJ ha^{-1} cannot be a "good" strategy in the considered climate and may be more or less retained to be "good" for greater values.

(3) Define the set of weights W_j that expresses the relative importance of each criterion X_j. This again is a subjective statement. The overall concordance score for the strategy a relative to the category K_i is then the weighted sum over all criteria, $CT_i(a) = \Sigma W_j C_{i,j}(x_{j.a})$. The overall discordance score is given by $DT_i(a) = 1 - \Pi[1 - D_{i,k}(x_{j.a})]^{W_j}$. Notice that the weights are identical in the concordance and discordance score equations.

(4) Define the decision variable $R_{\text{good}}(a) = CT_{\text{good}}(a).\ [1 - DT_{\text{good}}(a)]$ and the threshold t_{good} such that if $R_{\text{good}}(a) > t_{\text{good}}$ then the strategy a is assigned to the "good" category. Similarly define $R_{\text{bad}}(a) = CT_{\text{bad}}(a).\ [1 - DT_{\text{bad}}(a)]$ and the threshold t_{bad} such

that if $R_{\text{bad}}(a) > t_{\text{bad}}$ then a is assigned to the "bad" category. Note that it is possible that a be assigned to both categories.

(5) Evaluate the results of each strategy for N different climates. Let n_{good} be the number of climates for which strategy a is considered as "good" and n_{bad} the number of contexts in which it is considered as "bad". Then the management strategy a is selected if $n_{\text{good}}/N \geq MIN_{\text{good}}$ and $n_{\text{bad}}/N \leq MAX_{\text{bad}}$, where MIN_{good} and MAX_{bad} are thresholds that will depend on the decision maker's aversion to risk.

6. Results

The management strategies selected by BETHA for ethanol wheat are very different than typical strategies for wheat for food. In particular, the ethanol-wheat strategies are much less intensive. Compared to management recommended for wheat for food, the most extensive strategy for ethanol wheat has 40% less nitrogen fertilizer, 100% less fungicide and growth regulators and a 50% lower sowing density.

The semi-net profit is higher for low levels of inputs due in large part to the low price of ethanol wheat (46€ per ton less than for wheat for food). Furthermore, low input strategies improve the production cost per ton and the environmental criteria compared to intensive strategies. However, such strategies do not improve the energy balance but remain acceptable because their energy balance is higher than the veto value (3000 MJ ha^{-1}).

7. Discussion

It is difficult to define criteria for judging management strategies. A major problem is that in many cases the relevant criteria really apply to other spatial or temporal scales than a single field and just the wheat growing season. For example, nitrogen pollution can be reduced by introducing a catch crop after harvesting wheat, so we would consider both wheat and the following catch crop (Machet and Mary, 1990). Also, wheat has an impact on the succeeding crop which we have not taken into account. For example, wheat diseases can be transmitted *via* crop residue from one season to another. We would need multi-year simulations to take this into account. We have not done this, because in any case we do not know how to quantify these effects, which furthermore depend on the succeeding crop.

Another difficulty is that there is often insufficient knowledge on which to base the criteria. Judging environmental impact in particular is difficult. No doubt the environmental criteria that we used are insufficient to completely characterize pollution risks. In future, we may consider replacing them with the nitrate content in drained water and indicators of the risk associated with the use of plant protection products.

It is also difficult to determine relevant concordance and discordance functions and weights W_j. We finally chose a set of weights that gives priority to economic and environmental objectives. However, identifying the different criteria and fixing the weights can be a rewarding learning experience. In our case it was the result of talking to the different stakeholders and trying to understand their objectives. The quality of the initial work done to formulate the problem affects to a large extent the quality of the solution.

The use of fuzzy logic hopefully means that small changes in the concordance and discordance functions will only lead to small changes in the results. To verify, we did a sensitivity analysis to investigate the effect of slightly changing weights. Five sets of weights other than the standard set, each giving priority to a particular coalition of criteria, were tested. Care was taken to ensure that the values of the weights did not deviate too much from the standard values. Results showed that the solutions were not very sensitive to limited variations in the weights.

Overall, the approach of BETHA seems to be a promising way of developing management strategies for agricultural systems. It makes it possible to comply with the complex requirements of agricultural production and to sort strategies using criteria with different units. We used BETHA for the specific case of ethanol wheat, but it could be used in general as a tool for identifying management strategies for winter wheat when a compromise has to be found between economic, environmental and quality considerations.

References

Abecassis, J., 1993. Nouvelles possibilités d'apprécier la valeur meunière et la valeur semoulière des blés. Industrie des céréales 81, 25–37.

Bergström, L., Brink, N., 1986. Effects of differentiated applications of fertilizer N on leaching losses and distribution of inorganic N in the soil. Plant and Soil 93, 333–345.

Brans, J.P., 1983. Les mathématiques face aux problèmes de décision. In: Roy, B. (Ed), La Décision, Ses Disciplines, Ses Acteurs. Proceedings of a Symposium/Workshop, Colloque de Cerisy, Presses universitaires de Lyon, Lyon, pp. 47–68.

Chaney, K., 1990. Effect of nitrogen fertilizer rate on soil nitrate nitrogen content after harvesting winter wheat. Journal of Agricultural Science, Cambridge 114, 171–176.

Charnes, A., Cooper, W.W., 1977. Goal programming and multiple objective optimizations. European Journal of Operations Research 1, 39–54.

Gate, P., 1995. Ecophysiologie du Blé, de la Plante à la Culture," Lavoisier Tec et Doc, ITCF.

Loyce, C., Rellier, J.P., Meynard, J.M., 2002a. Management planning for winter wheat with multiple objectives (1): the BETHA system. Agricultural Systems 72, 9–31.

Loyce, C., Rellier, J.P., Meynard, J.M., 2002b. Management planning for winter wheat with multiple objectives (2): ethanol-wheat production. Agricultural Systems 72, 33–57.

Machet, J.M., Mary, B., 1990. Effet de différentes successions culturales sur les risques de pertes en nitrate en région de grande culture. In: Calvet, R. (Ed), Nitrates-Agriculture-Eau. Symposium International – INA P-G, Paris-La Défense, INRA, pp. 395–405.

Mackworth, A.K., 1987. Constraint satisfaction. In: Chapiro, S.C. (Ed), Encyclopedia of Artificial Intelligence. Wiley, New York, pp. 205–211.

Makowski, D., Wallach, D., Meynard, J.M., 1999. Models of yield, grain protein and residual mineral nitrogen responses to applied nitrogen for winter wheat. Agronomy Journal 91, 377–385.

Martin-Clouaire, R., Rellier, J.P., 1995. Making sequential crop management decisions: a constraint satisfaction approach. IFAC workshop on Artificial Intelligence in Agriculture, Wageningen, The Netherlands, pp. 179–184.

Meynard, J.M., 1985. "Construction d'itinéraires techniques pour la culture du blé d'hiver," Thèse, INA-PG, Paris.

Meynard, J.M., 1991. Pesticides et itinéraires techniques. In: Bye, P., Descoins, C., Deshayes, A. (Eds), Phytosanitaire, Protection des plantes, Biopesticides. INRA, Paris, pp. 85–100.

Meynard, J.M., Boiffin, J., Caneill, J., Sebillotte, M., 1981. Elaboration du rendement et fertilisation azotée du blé d'hiver en Champagne crayeuse. II. Types de réponse à la fumure azotée et application de la méthode du bilan prévisionnel. Agronomie 9, 795–806.

Meynard, J.M., Justes, E., Machet, J.M., Recous, S., 1997. Fertilisation azotée des cultures annuelles de plein champ. In: Lemaire, G., Nicolardot, B. (Eds), Maîtrise de l'azote dans les agrosystèmes. INRA, Reims, France, pp. 183–199.

Morgan, O.W., McGregor, M.J., Richards, M., Oskoui, K.E., 1989. SELECT: an expert system shell for selecting amongst decision or management alternatives. Agricultural Systems 31, 97–110.

Perny, P., 1998. Multicriteria filtering methods based on concordance and discordance principle. Annals of Operations Research 80, 137–166.

Rémy, J.C., Viaux, P., 1982. The use of nitrogen fertilisers in intensive wheat growing in France. Symposium on fertilisers and intensive wheat production in the EEC, The Fertiliser society of London, London, pp. 67–92.

Richards, I.R., Wallace, P.A., Paulson, G.A., 1996. Effects of applied nitrogen on soil nitrate-nitrogen content after harvest of winter barley. Fertilizer Research 45, 61–67.

Roy, B., 1985. Méthodologie multicritère d'aide à la décision. Economica, Paris.

Roy, B., Bouyssou, D., 1993. Aide multicritère à la décision: méthode et cas. Economica, Paris.

Schärlig, A., 1985. Dècider sur plusieurs critères. Panorama de l'aide à la décision multicritére, Vol. 1, pp. 304.

Schiex, T., 1993. Problèmes de satisfaction de contraintes. In: Alliot, J.M., Shiex, T. (Eds), Intelligence artificielle et informatique théorique Cepadues, Toulouse, pp. 245–280.

Spiertz, J.H.J., Vos, J., 1985. Grain growth of wheat and its limitation by carbohydrate and nitrogen supply. In: Day, W., Atkin, R.K. (Eds), Wheat Growth and Modelling. Plenum Press, New York, pp. 129–141.

Tsang, E., 1993. Foundations of Constraint Satisfaction. Academic Press, London.

van Keulen, H., Stol, W., 1991. Quantitative aspects of nitrogen nutrition in crops. Fertilizer Research 27, 151–160.

Van der Werf, H., Zimmer, C., 1998. An indicator of pesticide environmental impact based on a fuzzy expert system. Chemosphere 36, 2225–2249.

Wijnands, F.G., 1997. Integrated crop protection and environment exposure to pesticides: methods to reduce use and impact of pesticides in arable farming. European Journal of Agronomy 7, 251–260.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Appendix

Statistical notions

1. **Random variable** **429**
2. **Cumulative distribution and density functions** **430**
3. **Several random variables** **430**
4. **Expectation, variance, covariance, and correlation** **430**
5. **Some particular distributions** **431**
6. **Conditional distribution** **431**
7. **Regression** **432**
8. **Estimators and estimates** **433**
9. **Bayesian and frequentist statistics** **434**
References **435**

1. Random variable

For our purposes a random variable, say X, is a function defined on the set of all possible outcomes of an experiment involving some degree of randomness, that takes values in the set of integers or on the real line. A classic example is a throw of a die, with X being the number of dots on the upturned side. This X is a random variable because one does not know in advance what the result of throwing the die will be. It can be any number between 1 and 6. Another example would be $X =$ yield in a randomly chosen French corn field. This X is a random variable both because the choice of field is random and because yield in a particular field is not known in advance.

It is important to identify the range, or "population", of experiments or observations which give rise to a random variable. The die throwing experiment concerns some particular die and throws that make the die turn over several times before it lands. The population consists of possible throws. For the random variable $X =$ yield in a randomly chosen French corn field, the population can be thought of as all near future years and in each year all fields planted to corn. The notion of future years includes the diversity of possible climates, pest and disease levels, initial conditions, etc. If we restricted the population to fields with conventional practices (eliminating for example organic farmers), that would give rise to a different random variable.

2. Cumulative distribution and density functions

The cumulative distribution function of a random variable X is defined as $F_X(x) = P[X \leq x]$, the probability that X is less than or equal to the value x. For a fair die for example $F_X(2) = P[X = 1] + P[X = 2] = 2/6$.

For a discrete random variable, the probability density function (we will also call it simply the density function) $f_X(x)$ is the probability that $X = x$. For a fair die, the probability of each number from 1 to 6 is 1/6 so that $f_X(x) = 1/6$ for $x = 1, \ldots, 6$.

For a continuous random variable $f_X(x)dx$ is the probability that X is in the range x to $x + dx$ for infinitesimal dx. The relationship between the cumulative distribution function and the density function is $F_X(x) = \int_{-\infty}^{x} f_X(u)du$. Since the probability of having some value between minus and plus infinity is 1, we have $\int_{-\infty}^{\infty} f_X(u)du = 1$.

We will often refer to the "distribution" or "probability distribution" of a random variable. This could be defined by its cumulative distribution function or equivalently by its density function.

3. Several random variables

Let $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ be a random vector. The notation $(X_1, \ldots, X_n)^{\mathrm{T}}$ means to take the transpose of the vector, which converts the row vector (convenient for display), into a column vector (the form we work with).

The joint density function of the random variables $X_1, \ldots, X_n$, noted $f_{X_{1,\ldots,}X_n}(x_1, \ldots, x_n)$ is the probability that X_1 is in the range x_1 to $x_1 + dx_1$ <u>and</u> that X_2 is in the range x_2 to $x_2 + dx_2$, etc.

The marginal distribution of a random variable X_1 is defined by $f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_{1,\ldots,}X_n}(x_1, \ldots, x_n)dx_2, \ldots, dx_n$.

4. Expectation, variance, covariance, and correlation

The definition of the expectation of a continuous random variable is $E(X) = \int_{-\infty}^{\infty} x f_X(x)dx$. The expectation of a function of a random variable, say *g(X)*, is defined by $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx$. If X_1 and X_2 are random variables and c_1 and c_2 are constants then $E(c_1 X_1 + c_2 X_2) = c_1 E(X_1) + c_2 E(X_2)$.

The expectation of a function of several random variables, say $g(X_1, \ldots, X_n)$, is $E[g(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f_{X_{1,\ldots,}X_n}(x_1, \ldots, x_n)dx_1, \ldots, dx_n$. In this book, unless explicitly noted otherwise, operators are assumed to refer to all random variables in the expression that is operated upon. Thus an expression like $E[g(X_1, X_2)]$ means to take the expectation over both random variables. If the intention is to take the expectation over say only X_1 then we write either $E_{X_1}[g(X_1, X_2)]$ or $E[g(X_1, X_2)|X_2]$ which means to treat X_2 as a fixed value and to take the expectation only over X_1 (see the section "Conditional Distribution" below). Note that $E_{X_2}\{E_{X_1}[g(X_1, X_2)]\} = E[g(X_1, X_2)]$, i.e. if we first take the expectation over X_1 treating X_2 as fixed, and then take the expectation over X_2, this is the same as taking the expectation over both random variables.

The definition of the variance of a continuous random variable is $\mathrm{var}(X) = \sigma_X^2 = E[X - E(X)]^2$. If X is a random variable and c a constant, then $\mathrm{var}(cX) = c^2\mathrm{var}(X)$. The square root of the variance is called the standard deviation and noted as σ_X.

The covariance between two random variables X_1 and X_2 is defined as $\mathrm{cov}(X_1, X_2) = \sigma_{X_1,X_2} = E\{[X_1 - E(X_1)][X_2 - E(X_2)]\}$. Note that $\mathrm{cov}(X_1, X_2) = \mathrm{cov}(X_2, X_1)$. The variance–covariance matrix of a random vector $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ has the variances of the random variables as diagonal elements and has covariances as off-diagonal elements:

$$\Sigma_X = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots & \mathrm{var}(X_n) \end{pmatrix}$$

The correlation coefficient of two random variables X_1 and X_2 is defined by $\rho(X_1, X_2) = \frac{\mathrm{cov}(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}}$. The correlation coefficient has the property $-1 \leq \rho(X_1, X_2) \leq 1$.

5. Some particular distributions

A very simple probability distribution is the uniform distribution, whose probability density function is noted as $U(a, b)$. A random variable whose density function is $U(a, b)$ has equal probability of taking any value in the range (a, b) and zero probability of being outside this range. The notation $X \sim U(a, b)$ means that X has the probability density function $U(a, b)$. It is equivalent to $f_X(x) = 1/(b - a)$ for $a \leq x \leq b$ and $f_X(x) = 0$ otherwise.

Another very commonly encountered distribution is the normal distribution whose probability density function is noted as $N(\mu, \sigma^2)$. The two parameters are the expectation μ and the variance σ^2. If X has a normal distribution with expectation μ and variance σ^2 then $f_X(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$. We will also write this as $X \sim N(\mu, \sigma^2)$.

If a random vector $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ has a multivariate normal distribution then the density function is

$$f_X(x_1, \ldots, x_n) = N_n(\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}\,|\Sigma|^{1/2}}\mathrm{e}^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}$$

where $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ is the vector of expectations, $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ is the vector of the values of $X = (X_1, \ldots, X_n)^{\mathrm{T}}$, Σ is the variance–covariance matrix and $|\Sigma|$ is the determinant of Σ.

6. Conditional distribution

The conditional density function of X_1 given X_2, noted as $f_{X_1|X_2}(x_1|x_2)$ is the probability that X_1 is in the range x_1 to $x_1 + dx_1$ given that X_2 has the value x_2. The conditional density function is related to the joint and marginal density functions by

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \tag{1}$$

The random variable that has the above density function, say $X_3 = (X_1|X_2)$, is a new random variable which may have a different distribution for each possible value of X_2. When we want the random variable that corresponds to a particular value of X_2 we write $X_3 = (X_1|X_2 = x_2)$. To help understand the difference, note that $E(X_1|X_2)$ is a random variable because X_2 is a random variable, whereas $E(X_1|X_2 = x_2)$ is a number.

In the case of the multivariate normal distribution, all the conditional distributions are also normal. Suppose that we partition a random vector X into two sub-vectors noted as X_A (with expectation vector μ_A) and X_B (with expectation vector μ_B). The variance–covariance matrix of X is correspondingly partitioned as $\Sigma_X = \begin{pmatrix} \Sigma_\mathrm{A} & \Sigma_\mathrm{A,B} \\ \Sigma_\mathrm{B,A} & \Sigma_\mathrm{B} \end{pmatrix}$. The conditional distribution of X_A given that $X_\mathrm{B} = b$, where b is a vector of constants, is then a normal distribution with expectation and conditional variance

$$E(X_\mathrm{A}|X_\mathrm{B} = b) = E(X_\mathrm{A}) + \Sigma_\mathrm{A,B}\Sigma_\mathrm{B}^{-1}(b - \mu_\mathrm{B})$$

$$\mathrm{var}(X_\mathrm{A}|X_\mathrm{B} = b) = \Sigma_\mathrm{A,B}\Sigma_\mathrm{B}^{-1}\Sigma_\mathrm{B,A}.$$

An often useful formula is $\mathrm{var}(X_1) = \mathrm{var}\,[E(X_1|X_2)] + E\,[\mathrm{var}(X_1|X_2)]$. In words, the overall variance is equal to the variance of the conditional expectation plus the expectation of the conditional variance.

Two random variables are defined to be independent if $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$. This has consequences for the conditional distributions. Substituting into Eq. (1) gives

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} = \frac{f_{X_1}(x_1)f_{X_2}(x_2)}{f_{X_2}(x_2)} = f_{X_1}(x_1)$$

This shows that in the case of independence the distribution of X_1 given that X_2 has some particular value is exactly the same as the distribution of X_1 which completely ignores X_2. In the case of independence then, knowledge of the value of X_2 contains no information about the value of X_1.

7. Regression

The major purpose of regression is to study the dependence of one set of variables on another. In simple linear regression the relationship is assumed to be of the form

$$E(Y|X = x) = \alpha + \beta x$$

where Y and X are random variables and α and β are parameters which are fixed but unknown. The more general non-linear regression equation is

$$E(Y|X = x) = k(x; \theta) \tag{2}$$

where k is an arbitrary function, X is a random vector, x is a corresponding vector of fixed values and θ is a vector of parameters. It is often further assumed that $\mathrm{var}(Y|X = x) = \sigma^2$ so that the variance is independent of the value of x.

The response form of a crop model is usually written as $Y = f(x; \theta)$, but in fact is usually treated like a regression equation (for example, for parameter estimation) and so it would be more logical to write it in the form of Eq. (2). The assumption then is that the model represents an expectation, and the true responses vary around that.

8. Estimators and estimates

In essentially all cases of interest here, the random variables are associated with an infinite "population" of experiments and so the expectation, variance or any other function of the random variables cannot be obtained by observing all the individuals in the population. One can however estimate those values, based on a random sample from the population in question.

Suppose that one has a random sample of n individuals from the population to which the random variable X refers, and that one performs on each the same experiment or observation that gives rise to X. This gives a collection of n random variables noted as $X_1, X_2, \ldots, X_n$. X_1 for example is the random variable associated with the first individual in the sample. By definition of a random sample this could be any individual in the population with equal probability. Thus X_1 has exactly the same probability distribution as X, and so do $X_2, X_3, \ldots, X_n$.

Unbiased estimators of the expectation and variance of X are

$$\begin{aligned} \hat{E}(X) &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\mathrm{var}}(X) &= \frac{1}{n-1}\sum_{i=1}^{n} [X_i - \hat{E}(X)]^2. \end{aligned} \tag{3}$$

Suppose now that X is a vector with n components. We now use the notation X_j to denote the jth component of X and $X_{i,j}$ to note the jth component of X_j, the random variable associated with the ith individual. An unbiased estimator of the covariance of X_j and $X_{j'}$ is

$$\hat{\mathrm{cov}}(X_j, X_{j'}) = \frac{1}{n-1}\sum_{i=1}^{n} [X_{i,j} - \hat{E}(X_j)][X_{i,j'} - \hat{E}(X_{j'})]$$

Here and throughout we indicate estimators with a hat.

Equation (3) says that we can obtain an estimated value for $E(X)$ by drawing a random sample of size n, measuring the random variable in question for each individual and then taking the average of the measured values. Note that this is a recipe and not a numerical value. The recipe is itself a random variable since it is a function of random variables. Such a recipe is called an estimator. The distribution of an estimator represents how it would vary if the entire experiment were repeated. The variation arises because in each repetition different individuals would be chosen for the random sample, and in addition the results for a given individual have a random element. The expressions for $\widehat{\text{var}}(X)$ and for $\widehat{\text{cov}}(X_j, X_{j'})$ are also estimators. The value of an estimator obtained for a given sample of data is called an estimate. Contrary to an estimator, an estimate is a numerical value.

Two important notions for estimators are bias and variance. Suppose that an estimator of some quantity is denoted $\hat{\mu}$ and the true value μ. Remember that an estimator is a random variable. The bias of the estimator is defined as $\text{bias}(\hat{\mu}) = E(\hat{\mu}) - \mu$ and the variance is $\text{var}(\hat{\mu}) = E\{[\hat{\mu} - E(\hat{\mu})]^2\}$. The mean squared error of the estimator is $\text{MSE}(\hat{\mu}) = E[(\hat{\mu} - \mu)^2] = E[(\hat{\mu} - E(\hat{\mu}) + E(\hat{\mu}) - \mu)^2] = \text{var}(\hat{\mu}) + [\text{bias}(\hat{\mu})]^2$.

Suppose that the variance of X is σ^2 (where X is a random variable with only one component). Then the variance of the estimator $\hat{E}(X)$ in Eq. (3) is

$$\text{var}[\hat{E}(X)] = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{\sigma^2}{n}$$

Once we actually do the measurements on the n individuals in the random sample, we have n measured values $x_1, x_2, \ldots, x_n$. These can be used to obtain estimates of the mean and variance of X, where, as we said, an estimate (as opposed to an estimator) is a numerical quantity. Given the n measured values the estimate of the expectation of X is $\tilde{E}(X) = (1/n)\sum_{i=1}^{n} x_i$ and the estimate of the variance is $\tilde{\text{var}}(X) = (1/(n-1))\sum_{i=1}^{n}[x_i - \tilde{E}(X)]^2$.

The parameters in the regression equation (Eq. (2)) can be estimated if one has a sample of n pairs (Y_i, X_i). A simple approach, with desirable properties if the Y_i are uncorrelated and the assumptions of the regression equation are correct, is that of ordinary least squares. The least squares estimator is defined by

$$\hat{\theta} = \arg\min_{\vartheta}\sum_{i=1}^{n}[Y_i - f(X_i, \vartheta)]^2 \tag{4}$$

9. Bayesian and frequentist statistics

The major difference between these two schools of statistics is not in how they do calculations but in what they choose to calculate. The difference applies particularly to parameter values.

In frequentist statistics, a parameter is treated as a fixed quantity and one concentrates on obtaining an estimator $\hat{\theta}$ of the parameter, based on the data in a sample. The least squares estimator of Eq. (4) is an example. This estimator will be, like the estimators introduced above, a random variable because different samples would give different results.

Frequentist statistics concentrate on the distribution of the estimator, in particular its bias and variance. The important point here is that the random variability concerns how the estimated value would vary if the sampling, and the experiment for each individual in the sample, were repeated.

Bayesian statistics on the other hand treat the parameter as a random variable whose distribution represents our knowledge about the parameter. This knowledge is assumed to have two sources. First, there is prior information available independently of the measured sample. Before the measured data are available, our information about θ is given by the prior distribution noted as $\pi(\theta)$. (We use here the notation that is commonly used in Bayesian statistics, which is somewhat different than the notation introduced above. In our previous notation the prior distribution might be noted $f_{\Theta}^{\text{prior}}(\theta)$, where Θ is the parameter treated as a random variable). The prior distribution gives the probability that the parameter value is in the range θ to $\theta + d\theta$ based on our prior information. The second source of information is measurements of a random variable Y. After the measured values y are available, our information about the parameter is given by the posterior distribution noted as $f_{\Theta|Y}(\theta|Y = y)$. This is the probability that the parameter value is in the range θ to $\theta + d\theta$ based on both the prior information and on the measured values. It is this distribution that is of primary interest. It can be calculated from the fundamental equation of Bayesian statistics which is

$$f_{\Theta|Y}(\theta|Y = y) = \frac{f_{Y|\Theta}(y|\theta)\pi(\theta)}{m_Y(y)},$$

where $m_Y(y)$ is the marginal distribution of Y. Unlike frequentist statistics, there is no notion here of how quantities would vary if the experiment were repeated. The posterior distribution is conditional on the results actually obtained from the experiment that was performed.

References

There are a very large number of basic statistical texts. We mention just two of them.

Casella, G., Berger, R.L., 1990. Statistical Inference. Wadsworth, Belmont, CA.

Mood, A.M., Graybill, F.A., Boes, D.C., 1974. Introduction to the Theory of Statistics, 3rd Edition. McGraw-Hill, Kogakusha, Tokyo.

Working with Dynamic Crop Models
Edited by Daniel Wallach, David Makowski and James W. Jones
© 2006 Elsevier B.V. All rights reserved.

Answers to Exercises

Chapter 2

1. (b) $Bias = 0.27$, $MSE = 0.122$, $RMSE = 0.349$, $MAE = 0.283$, $RRMSE = 0.235$, $RMAE = 0.180$, $EF = 0.490$, $r = 0.899$, $\text{index} = 0.856$, $\rho_C = 0.735$, $TDI(25\%) = 0.46$, $CP(0.2) = 33\%$, $Bias^2 = 0.073$, $SDSD = 0.011$, $LCS = 0.038$.
2. $Bias = 0.0$, $MSE = 0.239$, $RMSE = 0.488$, $MAE = 0.427$, $RRMSE = 0.329$, $RMAE = 0.405$, $EF = 0.0$, $r = 0.0$, $index = 0.0$, $\rho_C = 0.0$, $TDI(25\%) = 0.53$, $CP(0.2) = 17\%$ $Bias^2 = 0.0$, $SDSD = 0.239$ $LCS = 0.0$.
3. (a) $\Lambda = 0.04$, $\Delta = 1.05$, $MSEP(\hat{\theta}) = 1.09$
4. (a) $\Lambda = 0.01$, $\Delta = 0.123$, $MSEP(\hat{\theta}) = 0.133$
5. (a) $intercept = 0.667$, $slope = 0.172$ (b) $MSE = 0.045$. (c) $\hat{M}SEP_{\mathrm{CV}}(\hat{\theta}) = 0.14$
6. (a) 1.8
7. (a) 52 (b) 26
8. (a) 0.27 (c) 0.049 (e) 0.070
9. (a) 80

Chapter 3

Only the numerical results are supplied.

1. (b) The expected value and standard deviation are equal to 81.15 and 27, respectively.
1. (c) About 0.6.
2. (d) 0.96 for c_1 and 0.51 for c_2.
3. (a) 9.
3. (c) 68.82.
3. (d) 1545.9.
3. (e) 33.15, 82.89 and 90.42.
3. (f) 40.18, 67.76 and 98.52.

3. (g) $\dfrac{\mathrm{var}\left\{E[\hat{Y}(2300)|c_1]\right\}}{\mathrm{var}[\hat{Y}(2300)]} = 0.42.$

3. (h) $\dfrac{\mathrm{var}\left\{E[\hat{Y}(2300)|c_2]\right\}}{\mathrm{var}[\hat{Y}(2300)]} = 0.37.$

3. (j) 2925.01, 855.51 and 271.1.

3. (k) 1333.82, 3060.91 and 6.57.

3. (l) $\dfrac{E\left\{\mathrm{var}[\hat{Y}(2300)|c_1]\right\}}{\mathrm{var}[\hat{Y}(2300)]} = 0.88.$

3. (m) $\dfrac{E\left\{\mathrm{var}[\hat{Y}(2300)|c_2]\right\}}{\mathrm{var}[\hat{Y}(2300)]} = 0.48.$

Chapter 4

Only the numerical results are supplied.

2. (d) If $Y = 9$ and $\sigma = 1$, $E(\theta|Y) = 8.2$ and $\mathrm{var}(\theta|Y) = 0.8$. If $Y = 9$ and $\sigma = 2$, $E(\theta|Y) = 7 = (5 + 9)/2$ and $\mathrm{var}(\theta|Y) = 2$.

3. (b) The smallest *MSEP* possible value is equal to 0.36.

Chapter 5

Only the numerical results are supplied.

1. (d) $E(Z|M) = 0.718$ and $\mathrm{var}(Z|M) = 6.49 \times 10^{-4}$.

2. (f) $E(Z_{10}) = 0.188$, $\mathrm{var}(Z_{10}) = 0.009$, $E(Z_{10}|M) = 0.7097$, and $\mathrm{var}(Z_{10}|M) = 6.45 \times 10^{-4}$.

3. (h) $\begin{pmatrix} Z_{10}^{(1)} \\ Z_{10}^{(2)} \end{pmatrix} \sim N\left[\begin{pmatrix} 1.67 \\ 408 \end{pmatrix}, \begin{pmatrix} 0.00452 & 0.00324 \\ 0.00324 & 0.81 \end{pmatrix}\right],$

Kalman Gain $= \begin{pmatrix} 0.722 \\ 0.518 \end{pmatrix}$, $\begin{pmatrix} Z_{10}^{(1)} \\ Z_{10}^{(2)} \end{pmatrix}|M \sim N\left[\begin{pmatrix} 0.873 \\ 407.43 \end{pmatrix}, \begin{pmatrix} 0.0006 & 0.000432 \\ 0.000432 & 0.808 \end{pmatrix}\right]$

Chapter 7

1. (a)

Horizon	(i)		(ii)	
	θ_{100}	$\theta_{15\,000}$	θ_{100}	$\theta_{15\,000}$
Ap	31.96	14.85	38.88	16.57
E	30.92	15.49	35.09	14.50
BT	32.68	20.77	35.70	20.44

(b)

Horizon	AWC (mm per cm of soil)	
	(i)	(ii)
Ap	1.71	2.23
E	1.54	2.06
BT	1.19	1.53

(c) The AWC predicted for the entire soil with the PTFs (i) and (ii) is 130.8 and 171.0 mm of water, respectively.

2. The AWC measured for the entire soil is 135.1 mm of water. This value of AWC shows that with the PTFs selected, the PTFs predicting the water content at particular points of the water retention curve are more accurate than the PTFs predicting the parameters of the water retention curve. The error with the latter is essentially related to overestimation of θ_{100}.

3. (b) Median values

	St Bonnet-le-Froid	Villesèque	Mideulville
Precipitation decision rule	11/5	13/5	10/5
Water balance	10/5	11/5	10/5

(c) Areas sown in sunflower at Mideulville (ha).

Date	(i) Precipitation decision rule	(ii) Water balance
April 24	0	735
April 26	0	1021
April 27	0	481
April 28	735	0
May 4	0	3186
May 5	0	2207
May 6	1021	0
May 7	481	0
May 10	3186	0
May 11	2207	0

4. (a) Case 1: $\hat{M}SEP = 1.06 \times 10^6$ tons^2. Case 2: $\hat{M}SEP = 0.31 \times 10^6$ tons^2.

(b) Case 1 and 2: $MSE = 00208$ (t/ha)^2.

Index

above-ground biomass, 258
accumulated biomass, 259
agreement index, 21–2
agricultural statistics, 223–4
agro-environmental criterion, 391
Akaike Information Criterion (AIC), 119
allocation, 266
analyses of variance, 383
analysis of variance (ANOVA), 73, 237, 252, 381
architecture, 300
automated optimization algorithms, 415
available water capacity, 216
average model, 42, 44
aversion to risk, 425
AZODYN, 261, 285, 291, 294–9

balance method, 174–5
bas, 105, 119–20
Bayesian, 101, 104, 116, 128, 132, 139
Bayesian Information Criterion (BIC), 119
Beer's law, 264
Bellman's optimality equations, 196–7
BETHA, 283, 285, 293–4, 303
bias, 17, 22
biofuels, 419
biomass, 230, 286
bootstrap sampling, 360, 362, 363
bootstrap, 37, 113
bottom marginal variance, 70
branching methods, 190
breeding, 284, 311–3
  multi-environment trial (MET), 312–3
  program, 311, 313
  strategies, 313, 322, 328, 331–2
  target environments, 311
  target population of environments (TPE), 313–4, 319, 321

C^{++}, 274
canopy, 264
  temperature, 264
canopy structure, 228
  wheat, 229
capillary rise, 262
categorical factor, 383
CERES, 302
CERES-EGC, 261
Cholesky decomposition, 63
classification, 223
clay, 272
climate, 414
  change, 214
  scenarios, 214
climatic potential demand, 269
CO_2, 267
companion tools, 413
complete enumeration methods, 187
complex requirements of agricultural production, 426
computer experiments, 55
concordance, 423
concordance correlation coefficient, 22, 24
conditional mean, 402–3
constrained estimation, 124
constraint satisfaction problem (CSP), 420
corn, 252, 337, 409
  irrigation, 337
correlation, 126–7
correlation coefficient, 20, 22, 24
correlations between parameters, 360, 362, 365
covariance matrix, 309, 401
coverage probability, 22, 25
critical curves, 272
crop, 379, 382
crop biomass, 392
crop coefficient, 270
crop development, 257
crop management, 409
crop model, 391, 410, 420

crop nitrogen uptake, 392
crop residues, 272
crop species, 222
crop statistics/crop acreage, 223
crop succession, 380
crop water needs, 227
CROPGRO, 261, 285–7, 292, 300, 302
cropping systems, 377–390
CROPSYST, 261
cross validation, 35, 118, 253, 338–9
cultivar-specific parameters, 274
cultivation techniques, 379, 382

daily time step, 272
data assimilation, 8, 151, 229, 252–3, 391
  Kalman filters, 252
data mining, 223
data splitting
  split the data, 35
database, 284–5, 293
decision-making process, 177
decision model, 409
decision optimization, 252
  dynamic programming, 252
  multiple criteria analysis, 252
  reinforcement learning, 256
decision quality, 44
  criterion, 45
  estimating C, 45
decision rule(s), 173–4, 177, 183, 225, 227, 254, 409
  irrigation strategy, 227
  judging, 184
  mathematical form, 184
  uses, 175
decision strategy, 177
decision support systems, 391
decision support tool, 420
decisions, 175
  reproducing past decisions, 175
  predicting, 176
  optimizing, 177
deconvolute, 233
degree-day, 264
denitrification, 359–365
density, 346
deterministic optimization problem, 186
diagnosis, 175, 210
diagnostic data, 260
discordance, 423
discrete difference equations, 403
discrete time step, 405
disease, 283, 285, 293–4
dynamic equations, 178
dynamic system model, 3

effective porosity, 220
energy balance, 270
engineering tools, 7
Ensemble Kalman filter (ENKF), 163–4, 407
ensemble members, 401
ensemble realization, 401
error propagation, 55, 238
estimate parameters, 274
estimating input variables, 229
estimator, 104
evaluation, 11, 175, 252–3
  cross validation, 252
  graphical representation(s), 13, 342
  importance of, 11
  MSEP, 252
  output is a distribution, 26, 253
  regression based, 23–4
  role of, 12
  time-dependent output variable, 15, 25, 27
  validation, 11
evaluating heterogeneity, 415
evaluating strategy, 413
expected utility, 179
experimental error, 286–7
experimental plan, 385
explanatory variables, 5, 410
Extended Kalman Filter (EKF), 162–3, 402, 407
extinction coefficient, 269

factorial design, 73, 93
  fractional, 78
  resolution, 79
farmer behavior, 176–7
fertilization, 272
field capacity, 270
first-order sensitivity coefficient, 364
first-order sensitivity index, 70
flowering, 262
  thermal time, 292
flux gradient, 272
forcing state variables, 232
formulate the problem, 425
FORTRAN, 274

forward regression, 338
Fourier amplitude sensitivity test (FAST), 88
  extended, 91
Fourier coefficient, 90
fraction cover, 266
fraction of transpirable soil water (FTSW), 270
frequentist, 101
fruit size distribution, 351–5
future climate, 178
fuzzy set, 424

Gauss–Marquardt, 109
Gauss–Newton, 108
gene based approaches, 323–4, 327
gene based model(s), 313
gene flow, 252, 377–390
general circulation models (GCMs), 214
generalized least squares, 110, 127, 132, 339
generalized likelihood uncertainty estimation (GLUE), 134
GeneSys, 377–390
genetically modified crops, 254
genetically modified organisms (GMO), 286
genotype × environment, 282
genotype × environment × management, 302
genotypes, 380
GIS, 274
global methods, 193
global optimization, 313
gradient search algorithm, 191
graphical user interface, 413

harvest index (HI), 268, 342, 348
  threshold, 268
harvest pollution, 383
holding capacity, 270
HRI WELL-N, 56
humus, 272
hydrological model, 274

ICECREAM, 56
identifiability, 116–17
ideotype, 284, 286, 302–3
impact of the crop on the environment, 423
importance measure, 85
indeterminate growth, 267
indicator variables, 174, 184, 410
initial conditions, 222, 258, 403
  soil nitrogen, 222
  soil water, 222
initial status, 258
input factors, 380
interaction plot, 69
interactions, 233
  between fields, 233
interpolation, 211–12
irrigation, 252, 270
irrigation scheduling, 409
irrigation strategy, 409
IRRINOV$^{®}$MAIS, 414

Kalman filter, 151–2
Kalman Gain Matrix, 403
Kiefer–Wolfowitz method, 193
Kiwifruit, 252, 349–51, 354–5
Klomogorov–Smirnov test, 26
Kolmogorov–Smirnov statistic, 252, 354

land use, 223
  sampling, 224
Latin hypercube, 63
Latin Hypercube sampling, 362
leaching
  nitrate, 258, 262
leaf area index (LAI), 257–8, 392
leaf surface, 290, 300
least squares, 101, 116
lifespan, 264
limestone, 272
limiting factor, 258
linear model, 158, 383, 385, 387
linearization, 163
LINTUL-POTATO, 302
local methods, 191
local search methods, 188
low input strategies, 425
low resolution satellite data, 233

main effect, 70
management, 253, 262, 282–3
  decision, 273
management decisions, 174
management practices, 222
management strategy, 420
management variable, 173
Markov chain Monte Carlo (MCMC), 134
Markov control problem, 195
maximin approach, 182
maximum available water content (MAWC), 270
maximum filling, 268
maximum likelihood, 101–11, 114, 116, 198

maximum rooting depth, 231
mean absolute error, 18, 22
mean squared error (MSE), 339
  decomposition, 21, 23–4, 341
  systematic part, 24
  unsystematic part, 24
mean squared error of prediction (MSEP), 26, 105, 116, 118–19, 132, 151–2, 235, 237, 253, 339
  decomposition, 30, 42–3
  estimation, 33, 236
  measurement errors in X, 39
  measurement errors in Y, 38
  multiple fields, 235, 237
  parameter uncertainty, 40
  relation to MSE, 27, 34
  time-dependent output variable, 27
measurement error, 235, 400
measurements, 258
Metropolis–Hastings, 133–7, 146–8
mineralization, 269
mixed-effect model, 127–8
model
  generic, 261
model applications, 252
model complexity
  complex model, 34
  model, 30
model efficiency, 371–5
model error, 101–2, 110, 153, 155, 400
model structure, 303
modeling efficiency, 19, 22, 341
Monte Carlo, 133
Monte Carlo sampling, 62, 86, 379
Monte Carlo sampling techniques, 399
morphological traits, 312
  plant traits, 312
Morris' method, 73, 93
multi-criteria methods, 182
multiple criteria analysis, 423
multiple decision, 46
multiple fields, 254
multivariate regression model, 7

Nelder–Mead (simplex) method, 193
net energy, 423
neuro-dynamic programming, 199
nitrogen, 271
  diffusion-convection, 272
  nitrate, 272
  transformation, 252
  transport, 252
nitrogen absorption, 284, 287, 290, 294–5
nitrogen fertilization, 391
nitrogen fixation, 262
nitrogen nutrition index (NNI), 258, 271
nitrogen remobilization, 291
nitrogen stress, 267
nitrogen transformation model, 367
non-food crops, 419
non-linear model, 161
non-linear regression, 101, 106
nonparametric, 113
nonparametric regression, 46
number of grains, 268
numerical weather prediction, 212

objective function, 45, 178–9
optimal decision(s), 45, 173
optimal management strategy, 183
optimistic, 37
optimization, 8, 253, 311
optimization domain, 178, 182
optimization problem, 178
optimization problems with continuous domains, 191
optimization problems with discrete domains, 187
optimizing decision rules, 415
ordinary least squares, 107
organic matter, 272
  C/N ratio, 272
output variable, 258
overparametrization, 116

parameter adjustment, 303
parameter estimation, 8, 101, 151, 175, 252, 281, 284–95, 297, 303, 338, 391
  GLUE, 252
parameter(s), 5
  genotype, 281, 284
partial correlation coefficient (PCC), 81
partial derivatives, 402
partial least squares, 94
particle filter, 164
Pearson Product Moment Correlation Coefficient (PEAR), 81
pedotransfer function, 216
Penman-Montieth, 269
phenological stages, 258

phenology, 266
photoperiod, 264
photosynthesis, 262
photosynthetically active radiation (PAR), 264
plant breeding, 255
  varieties, 255
plant development, 258
plant growth, 258
pleiotropic effects, 312
pollen transfer, 254
population variance, 43
porosity, 220
posterior distribution, 104
preceding crop, 258
precision agriculture, 252, 391
prediction, 210, 253
prediction error, 404
predictive quality, 26, 283, 294–5, 303
Priestely-Taylor, 271
principal component analysis (PCA), 94, 313
prior distribution, 104
production cost per ton, 423
propagated, 402
protein content, 284, 287, 291

Q-learning, 199
qualitative evaluation, 350–1
quality, 258
  grain, 262
quality of the cakes for animal feed, 423
quantitative trait loci (QTI), 311

r^2, 383
  partial, 383, 385
radiation, 267
radiation interception efficiency, 300–1
radiation use efficiency (RUE), 258, 287, 294–6, 298, 300–1
radiative transfer model, 229
random, 350, 355
  randomness, 350, 355
random search algorithm, 190
random weather generators, 187
rank tests, 287
rate of development, 264
real-time use, 416
recommendations for corn irrigation, 414
reflectance, 229
regression equation, 7
regression model, 81
reinforcement learning, 199
relative mean absolute error, 19, 22
relative root mean squared error, 19, 22
remobilization, 268
remote sensing, 210, 223, 234, 252, 391
replications, 381
residuals, 344
residues, 16
respiration, 266
response model, 6, 178
response surface methodology, 80
response variables, 6
risk analysis, 55
risk aversion/inclination toward risk, 181
risks, 180
robustness, 287
root density, 270
root growth rate, 231
root mean squared error (RMSE), 18, 22, 123, 166
root mean squared error of prediction, 27
root system, 258
row spacing, 264

sample path optimization, 186
SAS QC, 94
saturated hydraulic conductivity, 220
scenario testing, 176, 210
scientific hypothesis, 7, 47
  test, 47
  validate, 48
seed bank, 380
selection, 284
selection bias, 340
selection of parameters, 116
semi-net profit, 422
senescence, 264
sensitivity analysis, 7, 55, 67, 118, 252–3, 290, 361–2, 426
  global, 61, 68
  input factors, 58
  input scenario, 59
  Latin Hypercube, 252
  local, 61, 67
  model independent, 57
  one-at-a-time, 71
  spatial, 378
  temporal, 378

sensitivity analysis response profiles, 373
sensitivity index, 372
  interaction, 75
  main effects, 75
  total, 75
simple relationships, 422
simulated annealing, 189
simulation
  duration, 381, 388
simulation units, 234
simulation-based control, 195
simulation-based optimization, 185–6
sink, 267
sink strength, 268
Sobol decomposition, 84
soil
  carbon, 399
  hydrodynamic properties, 273
  nitrogen, 258
  optimal estimation, 405
  water, 258
    balance, 257
    transpirable fraction, 258
soil carbon, 252
soil denitrification, 252
soil moisture, 346
soil physiochemical characteristics, 392
soil properties, 216
source, 267
sowing, 176
sowing date, 225, 229
sowing density, 264
spatial variability, 359–362, 365
Spearman correlation coefficient, 81
specific leaf area (SLA), 264
stakeholders, 422
standardized regression coefficient (SRC), 82
state variable, 4, 151–2, 258, 337, 410
static model, 422
statistical cartography, 223
statistical environment, 274
statistical package, 94
STICS, 56, 261
stochastic approximation, 193
stochastic dynamic equation, 5
stochastic dynamic programming, 198
stochastic models, 26
stochastic optimization problem, 185
stress indices, 258
sucrose crop model, 230
sugar beet, 230
  yield, 231
sunflower, 292, 300–1
support unit, 233
supra-optimum temperature, 264

target distribution, 27, 45
target environments, 311–12
Taylor series, 402
temperature, 346
tensiometer-based irrigation, 414
top marginal variance, 70
total deviation index, 22
total sensitivity index, 70
traits, 311
  complex, 311–2
  epistasis, 312
transfers, 233
  pathogen, 233
  pollen, 233
  water, 233
  weed, 233
transpiration, 267, 269
trial and error, 415
trophic competition, 268
typology, 225–6
  farm, 225–6
  irrigation strategies, 228

uncertainty, 128–9, 137, 400
uncertainty analysis, 55, 61–2, 361
  cumulative probability distribution, 65
uncertainty in input factors, 60
  climate generation, 60
  coding, 60
  control scenario, 60
  correlations between input parameters, 63
  nominal value, 60
  probability distributions, 60, 62
  uncertainty range, 60
unconditional mean, 405
unconstrained optimal decisions, 182
uptake, 270

value iteration, 198
vapor pressure deficit (VPD), 267
variability, 212
  spatial, 212
variance based measures of sensitivity, 84
variance–covariance matrix, 110–12, 127

variety, 344
  transgenic/GM or transgenic/genetically modified, 379
vernalization, 264
veto value, 424
virtual experiment, 380
volunteer density, 384

water retention, 216
water stress, 258
water use efficiency (WUE), 266
weather data, 211
weather generator, 213
weighted least squares, 110, 118, 121, 123, 252
weights, 424
wheat for ethanol, 252
wheat, multi-resistant, 283, 286
wheat production for ethanol, 419
wilting point, 270
winding stairs sampling scheme, 87
winter wheat dry matter, 58
WTGROWS, 56

yield
  decomposition of error, 340
yield components, 285, 290
yield gap analysis, 175
yield loss due to disease, 285
yield variability, 416

zoning, 211