Plant growth models and methodologies adapted to their parameterization for the analysis of phenotypes

Fenni Kang

ÉCOLE CENTRALE DES ARTS

ET MANUFACTURES

« ÉCOLE CENTRALE PARIS »

Thèse

présentée par Fenni Kang

pour l'obtention du GRADE DE DOCTEUR

Spécialité : Mathématiques Appliquées

Laboratoire d'accueil : Mathématiques Appliquées aux Systèmes (MAS)

Modèles de Croissance de Plantes et Méthodologies Adaptées à Leur Paramétrisation pour l'analyse des Phénotypes

Soutenue le 28/05/2013

Devant un jury composé de :

M. Paul-Henry COURNÈDE	(Directeur)
MME. Mengzhen KANG	(Rapporteur)
M. Jérémie LECOEUR	(Président, Rapporteur)
MME. Véronique LETORT	
M. Hubert VARELLA	

N° ordre : 2013 ECAP 0035

Plant Growth Models and Methodologies Adapted to Their Parameterization for the Analysis of Phenotypes

Plant growth models aim at describing the interaction between the growth of plants and their environment. Ideally, model parameters are designed to be stable for a wide range of environmental conditions, and thus to allow characterizing genotypes. They offer new tools to analyze the genotype × environment interaction and they open new perspectives in the process of genetic improvement.

Nevertheless, the construction of these models and their parameterization remain a challenge, in particular because of the cost of experimental data collection.

In this context, the first contribution of this thesis concerns the study of plant growth models. For sunflower (*Helianthus annuus* L.), the model SUNFLO [Lecoeur et al., 2011] is considered. It simulates the plant phenology, morphogenesis and photosynthesis under abiotic stresses. An extension of this model is proposed: this new SUNLAB model adapts into SUNFLO a module of biomass allocation to organs, using the source-sink concepts inspired by the GREENLAB model [De Reffye and Hu, 2003]. For maize (*Zea mays* L.), the CORNFLO model, based on the same principles as SUNFLO, was also studied. These models helps discriminating genotypes and analyzing their performances.

On the other hand, in order to parameterize these models, an original methodology is designed, adapted to the context of plant variety improvement by breeders. The MSPE methodology ("multi-scenario parameter estimation") uses a limited number of experimental traits but in a large number of environmental configurations for the parameter estimation by model inversion. Issues including identifiability, sensitivity analysis, and the choice of optimization methods are discussed. The influences of environmental scenarios amount on the model predictive ability and on estimation error are also studied.

Finally, it is demonstrated that selecting scenarios in different environmental classes (obtained by data clustering methods) allows to optimize the multi-scenario parameter estimation performances, by reducing the required number of scenarios.

Keywords: SUNFLO, CORNFLO, SUNLAB, MSPE, MSPEJ, MSPEE, Crop model, Plant growth model, Sunflower, Corn, Parameter estimation, Environment clustering

Modèles de Croissance de Plantes et Méthodologies Adaptées à Leur Paramétrisation pour l'analyse des Phénotypes

Les modèles de croissance de plantes cherchent à décrire la croissance de la plante en interaction avec son environnement. Idéalement, les paramètres du modèle ainsi défini doivent être stables pour une large gamme de conditions environnementales, et caractéristiques d'un génotype donné. Ils offrent ainsi des nouveaux outils d'analyse des interactions génotype × environnement et permettent d'envisager de nouvelles voies dans le processus d'amélioration génétique chez les semenciers.

Malgré tout, la construction de ces modèles et leur paramétrisation restent un challenge, en particulier à cause du coût d'acquisition des données expérimentales.

Dans ce contexte, le premier apport de cette thèse concerne l'étude de modèles de croissance. Pour le tournesol (*Helianthus annuus* L.), il s'agit du modèle SUNFLO [Lecoeur et al., 2011]. Il simule la phénologie de la plante, sa morphogenèse, sa photosynthése, sous les contraintes de stress abiotiques. Une amélioration de ce modèle a été proposée : il s'agit du modèle SUNLAB, implémentant dans le modèle SUNFLO les fonctions d'allocation de biomasse aux organes, en utilisant les concepts sources-puits du modèle GREENLAB [De Reffye et Hu, 2003]. Pour le maïs (*Zea mays* L.), le modèle CORNFLO, basé sur les mêmes principes que SUNFLO a également été étudié. Ces modèles permettent la différenciation entre génotypes.

D'autre part, afin de paramètrer ces modèles, une méthodologie originale est conçue, adaptée au contexte de l'amélioration variétale chez les semenciers : la méthode MSPE ("multi-scenario parameter estimation") qui utilise un nombre restreint de traits expérimentaux mais dans un grand nombre de configurations environnementales pour l'estimation paramétrique par inversion de modèles. Les questions d'identifiabilité, d'analyse de sensibilité, et du choix des mèthodes d'optimisation sont discutées. L'influence du nombre de scénarios sur la capacité de prévision du modèle, ainsi que sur l'erreur d'estimation est également étudiée.

Enfin, il est démontré que le choix des scénarios dans des classes environnementales différentes (définies par des méthodes de classification - clustering) permet d'optimiser le processus expérimental pour la paramétrisation du modèle, en réduisant le nombre de scénarios nécessaires.

Mots-clè : SUNFLO, CORNFLO, SUNLAB, MSPE, MSPEJ, MSPEE, Modèles de croissance de plantes, Tournesol, Maïs, l'estimation des paramètrique, Clustering environnemental

ACKNOWLEDGEMENT

The completion of this thesis owes much to numerous people. It is my pleasure here to thank all those who offered their support and encouragement during my research and study, and contributed in many ways to help me obtaining the PhD.

I would like to express my profoundest gratitude to my supervisor Paul-Henry Cournède. He opened the gate of the magic scientific world to me and he leaded me to enter in the world. He showed me the joy and the way of living in this world by teaching me how to address research questions and how to seek for scientific solutions. I have also many appreciations to his wise perspectives of our scientific paths, his compliments when we achieved good scientific fruits, and in particular his conviction to my ability when we met brambles in the researching path. He is the scientific father of me, who I should always appreciate for his orientation and care throughout the thesis.

I am deeply grateful to Véronique Letort - Le Chevalier. Her insightful ideas to my research largely boosted its progress. Her comforts motivated me getting though most stressful periods in my thesis. She has always been patient and encouraging. Without her guidance and her encouragement, this PhD would not have been achieved. I am mostly impressed by her serious and enthusiastic attitudes to science and her mild and amusing personality. She is the idol woman scientist I would myself become.

I greatly appreciate Jérémie Lecoeur. He depicted and laid out the blueprint of my research plan. His supports to my research and careers are pervading through my thesis until now. His points of views on plant growth modeling have profound and

steady influence on my understanding and interest of subjects in this field. His passions and motivations to science are contagiously stimulating me to give further efforts and further contribution to plant growth modeling.

I am especially thankful to Mengzhen Kang and Hubert Varella. They squeezed time from their busy schedules to attend my defense. Their valuable opinions and their unreserved help inspired me and brought me to a higher level of thinking.

I gratefully acknowledge the scholarships received from INRIA, the collaboration with Syngenta, and the aids from Laboratory of Mathematics Applied to Systems (Lab MAS) and Ecole Centrale Paris. Thanks to Frederic Abergel and Pascal Laurent of Lab MAS for their organizations of lab activities. I appreciate Sylvie Dervin, Annie Glomeron, Catherine Lhopital, Dany Kouoh-Etame of Ecole Centrale Paris, and Christine Biard of INRIA for their charge of my administrative affairs.

I would like to say thank you to all Digiplante past and current members, with whom I had enjoyable time discussing academic questions, sharing scientific opinions and methodologies, and going out together for benefiting amazing Parisian life. I am glad that my pleasant PhD time was spent with you. Zhongping Li, Yuting Chen, Qiongli Wu, Rui Qi, Xiujuan Wang, Li Song and Langshi Chen, the group of Chinese (mainly girls) brought so much fun to my lab life. Thank you for sharing my happiness and thank you for supporting me when I was in bad mood. Cedric Loi, the king of the group, has given me many help, especially at the beginning of my thesis, his smile helped me quickly integrated to the group. Thomas Guyard, Benoit Bayol, Chi-thanh Nguyen, Claire Mouton, Marion Carrier, thank you for offering robust technical solutions which fast boosted my work. Charlotte Baey, Robert Bayer, thank you for your nice accompaniments and kind encouragement during my thesis. Pauline Hezard, Samis Trevezas, Vincent Le Chevalier, Aurelie Cotillard, Corina Iovan, Blandine Romain, Margarita Pons-Salort, Amelie Mathieu, Octave Etard, Katarina Smolenova, I appreciate your sharing of scientific knowledge and the good time we spent going out

and in seminars.

I am indebted to Thomas Galinier, Benoit Pallas and Kamel Bezzou who provided me modeling methods and data materials. They invested their efforts on explaining me details of these tasks which very much facilitated my thesis. I appreciate Marion Moneuse and Hugo Magaldi, who are university students tutored by me for their projects in Digiplante. Thank you for your good contributions to our projects.

My time at the university and in Paris was happy and unforgettable thanks to my friends. Bo xiang is the most lovely and kindest girl I met. She gave me so many help, happiness and encouragements all the time. Olaf Torne helped me a lot for so many things: the correction of my papers, driving me around for knowing Paris, the suggestions for my career and so on. My good friends and my roommates, Honggui Wu and Yuanjing Zhou are so kind and so helpful. They took care of me and cooked for me when I was busy writing my thesis. Thanks a lot to other friends who were making an important part of my life in Paris: Abhijeet Gaikwad, Nicolas Millot, Mauro Politi, Rania Soussi, Wanping Lu, Beibei Hu, and Chaohui Wang.

Finally, my deep appreciation goes to my mom, dad, and boyfriend for their infinite love, encouragement and support. For my father who convinced me to persist in my PhD research when difficulties tortured me. For my mother who encouraged me to be brave of pursuing my dream. For my boyfriend who has been by my side to help me and amuse me. My heartfelt thanks to you all for your delight accompany in the last and in the future.

All sciences are now under the obligation to prepare the ground for the future task of the philosopher, which is to solve the problem of value, to determine the true hierarchy of values

Friedrich Nietzsche

The scientists of today think deeply instead of clearly. One must be sane to think clearly, but one can think deeply and be quite insane.

Nikola Tesla

I seem to have been only like a boy playing on the sea-shore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me.

Isaac Newton

CONTENTS

Acknowledgement 7

1. Introduction 17
1.1 Context: Breeding and Phenotyping 17
1.2 Crop Models Offer New Perspectives in Phenotyping and Breeding . . . 19
1.3 Parameterization of Plant Growth Models for Phenotyping 20
1.4 PhD objectives and outlines 21

Part I Materials and models 27

2. Field Experiments and Data 29
2.1 Experimental Data for Sunflower 30
2.1.1 Detailed Experimental Data for six Genotypes for two Years . . 30
2.1.2 Sparse Experimental Data for 90 Genotypes for three Years . . 31
2.2 Experimental Data for Corn 32
2.3 Additional Environmental Data 33
2.3.1 36 Years French Weather Data 33
2.3.2 Large Scale Environment Database 33

3. Plant Growth and Breeding Models 35
3.1 Principles of Crop Growth Modeling 35
3.1.1 Objectives and Constraints of Model Design in a Breeding Context 35
3.1.2 Methodologies and Mathematical Tools to Develop Crop Models 41
3.2 SUNFLO Model for Sunflower 44
3.2.1 Phenology Module 45
3.2.2 Architecture Module 46
3.2.3 Biomass Production Module 49
3.2.4 Biomass Allocation Module 52
3.2.5 Water Budget Module 52
3.2.6 Parameters 59
3.3 SUNLAB Model for Sunflower 60
3.3.1 Context and Objectives 61
3.3.2 Modeling: SUNLAB Modules 63

Organogenesis and Morphegenesis Module 64
Biomass Distribution Module 64
3.3.3 Related Datasets and Parameter analysis 67
3.3.4 Result: Sensitivity Analysis 69
3.3.5 Result: Model Calibration on Four Genotypes and Two Environmental Conditions 70
Parameter Estimation for the Four Genotypes 70
Model Performances: Reproducing Genotype-induced Variability 71
Model Performances: Reproducing Environment-induced Variability 72
3.3.6 Result: Model Validation 73
3.3.7 Result: Model Application: an Exploratory Study on Specific Leaf Area 75
3.3.8 Discussion 76

Part II Analysis 81

4. Multi-scenario Methodology of Parameter Estimation (MSPE) 83
4.1 MSPE Context and Objectives 83
4.2 The Feasibility of MSPE Methodology 86
4.2.1 Methodology 86
Proof of Concept Test 87
Statistical Test 89
4.2.2 Results 91
Proof of Concept Result 91
Statistical Test Result 92
4.2.3 Conclusion 92
4.3 Implementation and Optimization Issues in MSPE. 93
4.3.1 Parameter Selection 94
Continuity and Convexity 94
Parameter Sensitivity 97
4.3.2 Optimization Issues 100
Optimization Algorithms 101
Optimization Capacity 102
Computational Characteristics and Efficiency 104
4.3.3 Parameter Non-Estimability 105
4.3.4 Conclusion 108
4.4 Evaluation of MSPE Estimation Accuracy 108
4.5 Evaluation of MSPE Prediction Error 110
4.5.1 Number of Scenarios Increase Test 113
4.5.2 Cross Validation Test 115
4.6 Jackknife Based MSPE: an Extended Version 119

4.6.1 Delete-m Jackknife Estimator and MSPEJ Methodology 119
4.6.2 A Real Data Test 121
4.7 Conclusion 125

5. Environment clustering and Interaction with Crop Modeling 127
5.1 Data Clustering Methods 127
5.2 Environment Scenarios Clustering Based on Different Information Strategies 129
5.3 Multi-Scenario Parameter Estimation based on Scenario Selection . . . 132
5.3.1 Practical Identifiability 133
5.3.2 Improvement on MSPE Estimation Accuracy 135
5.3.3 Improvement on MSPE Prediction Ability 137
5.4 Conclusion 138

Part III Applications and Discussion 141

6. Model Simulations over Large Geographical Areas and Time Scales . . . 143
6.1 Context 144
6.2 Irrigation Demand Simulation 146
6.3 Simulation Experiments 147
6.4 Results 148
6.5 Discussion 151

7. Discussion 153

List of figures 161

List of figures 163

Bibliography 167

1. INTRODUCTION

1.1 Context: Breeding and Phenotyping

Increasing needs and requirements for food or raw materials pose grand challenges on plant derived products. The objective of breeding is to improve plant productivity across all scales from molecular to field applications by selecting or creating varieties with improved performance in agricultural environments. These crop improvement programmes, in particular, where breeding populations and cultivars are characterized by high genetic diversity and substantial genotype $\times$ environment interactions, are based on precise and efficient phenotyping.

Plant phenotype is the set of observable biophysical characteristics of a plant organism, as determined by both genes and environmental influences. Phenotyping is the construction, recording and analysis of phenotypes. It is the comprehensive assessment of plant complex traits such as growth, development, tolerance, resistance, architecture, physiology or yield. The phenotypic traits of interest can also include less integrated variables, for example to describe plant architecture or morphology (leaf surface area, plant height, stem diameter, internode length, leaf angle, seed number and size, tiller number) or phenology (flowering time, germination time). Phenotyping is thus a key step in the breeding process, by helping investigate the physiological principles involved in the control of basic plant functions [Walter et al., 2012].

However, the current limitations of phenotyping hamper the analysis of the existing genetic resources for their interaction with the environment. Progresses in plant phe-

notyping are key factors for the success of modern breeding and basic plant research. Early phenotyping by farmers or breeders to select crops with better yield or stronger resistance was mostly based on experience and intuition. Classical phenotyping tools are based on visual observations, measurements or biochemical analyses. A large set of different aspects led to the development of automated and high-throughput advanced plant phenotyping. For example, because the overall goal of phenotyping approaches with respect to plant breeding is to quantify or rank the success of a range of genotypes in certain environmental frameworks, which needs usually hundreds or thousands of genotypes to be compared with each other, this requires more rapid measurement procedures, a high degree of automation and access to appropriate, and well-conceived databases [Walter et al., 2012]. Different systems and initiatives were built for this purpose. For example, the ScanAlyzer platform by LemnaTec is a plant phenotyping system to extract and record plant phenotypic traits. It is capable to image plants in a greenhouse by automatically moving plants, placing them on beltways, and positioning them in front of a stereoscopic camera. Proprietary software analyzes the images to extract phenotypic-related information. Although fully developed and tested, this proprietary platform is very costly, requires a large investment in the appropriate infrastructure, and therefore its easy deployment and maintenance are in question [Tsaftaris and Noutsos, 2009]. Another initiative is PHENOPSIS, a custom growth chamber phenotyping system, developed by Optimalog, on contract by the Laboratory of Plant Ecophysiological responses to Environmental Stresses, in Montpellier France [Granier et al., 2005]. This proprietary system uses a robotic arm to position an array of sensors on top of a small plant within a growth chamber. As a custom-made proprietary solution there is limited information about its deployment cost. Many such systems have been built to facilitate the construction and record of phenotypes.

1.2 Crop Models Offer New Perspectives in Phenotyping and Breeding

While the progress of phenotypes recording has been boosted by the above introduced platforms, the analysis of the produced data still remains quite crude, generally based on classical statistical comparisons with actual genotyping information to correlate genotypes to phenotypes, see for instance Tsaftaris and Noutsos [2009]. Moreover, if the information recorded on plant descriptive variables is clearly enhanced by such systems, the technical implementation makes the range of environmental variations that can realistically be explored usually pretty limited, so that the statistical analysis performed generally lacks some predictive capacity in a wide range of environmental conditions.

In this context, Hammer et al. [2006] suggests that "while developing a predictive capacity that scales from genotype to phenotype is impeded by biological complexities associated with genetic controls, environmental effects and interactions among plant growth and development processes, organ-level plant growth model can help navigate a path through this complexity". The general idea is that plant growth models aim at describing the ecophysiological processes driving plant growth in interaction with the environment so that the parameters of the resulting model should be stable in a large range of environmental conditions and potentially characterize the genotype under study. As stated by Letort [2008]; Tardieu [2003], one genotype should be characterized by one set of model parameters. Such idea was declined in a few studies, on a submodel of maize leaf elongation in Reymond et al. [2003] or at whole plant level for sunflower in Casadebaig et al. [2011], Lecoeur et al. [2011].

As a consequence, well constructed plant growth models should be able to simulate phenotypic traits of various genotypes in diverse environments, and thus may provide an efficient help to analyze phenotype: it can predict crop performance over a range

of environmental conditions and help explaining the principle causes of phenotypic features from environment and genotypic factors. The consequence in breeding is potentially of great interest. 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 quantitative genetics, the phenotype is the result of the expression of the genotype, the environment and the interaction between the genotype and the environment [Messina et al., 2006]. Progress in breeding higher-yielding crop plants would be greatly accelerated if the phenotypic consequences of making changes to the genetic marker of an organism could be reliably predicted. Letort [2008] showed how plant growth models could be used as an intermediate in this process.

1.3 Parameterization of Plant Growth Models for Phenotyping

To improve the predictive capacity of plant growth models in various environments, the basic idea is to enrich the mechanistic description of plant ecophysiology [Yin and Struik, 2010]. However, the more complex the models are, the more troublesome their parameterization and the assessment of the estimate uncertainty [Chen and Cournède, 2012; Ford and Kennedy, 2011] are, specifically due to the costly experimentation and the great number of unknown parameters to consider. Likewise, local environmental conditions (in terms of climatic and soil variables, as well as biotic stresses) and initial conditions in specific fields are also very delicate to characterize. Consequently, the propagation of uncertainties and errors, which are related to parameters and inputs of these dynamic models, may result in unsatisfactory prediction concerning the plant-environment interaction in real situations.

In the context of breeding, in order to be able to discriminate between genotypes

based on their corresponding model parameters, the uncertainty on model parameters should be as small as possible. On the other hand, implementing the heavy experimental data collection necessary for model parameterization (see examples of experimental protocols for the parameterization of the GreenLab model for maize Guo et al. [2006], grapevine Pallas et al. [2010], rapeseed Jullien et al. [2011] or chrysanthemum Kang et al. [2012b]) is too costly in an industrial context implying the tests and characterization of large numbers of genotypes. For this purpose, as suggested by Jeuffroy et al. [2006b], it would be very useful if a methodology could be devised to take advantage of farmers' data (that are classically available at a reduced cost) for the parameterization of plant models. More generally, a well chosen panel of environmental conditions in which a few plant traits are measured should mathematically provide enough information for model identification.

1.4 PhD objectives and outlines

This thesis focuses on four issues: plant model design, parameter estimation, optimization of experimental protocol via environment classification, and model applications on producing phenotype analysis knowledge. Model analysis methods such as sensitivity analysis are also used for facilitating above issues.

Model design. The first objective of our research is the study of plant models adapted to the analyzis of the interaction between environment and genotypes. Chapter 3 is the chapter to introduce general crop modeling and model analysis theories, and involved models in this thesis. Two models SUNFLO [Casadebaig et al., 2011; Lecoeur et al., 2011] and SUNLAB [Kang et al., 2012a] are used for the sunflower crop (*Helianthus annuus*) and the CORNFLO model for the corn crop (*Zea mays L.*). These models are used for model simulation and prediction comparison, model analysis methodology testing, and model application.

SUNFLO, CORNFLO and SUNLAB all possess parameters with biological meaning that can potentially be characteristic of plant genotypes. These models have the advantage to predict complex plant or crop traits under diverse environmental conditions. Ecophysiological models are required to have more physiological feedback and accurate simulation of phenotypic features. SUNFLO, CORNFLO and SUNLAB simulate plant phenology, morphogenesis, photosynthesis, biomass production and biomass distribution under abiotic stress including temperature and drought stress. The water deficit, as an unbalance between soil water availability and evaporative demand, causes a set of decreased plant physiological functions.

The sunflower model SUNLAB is developed mainly to improve the biomass distribution module in SUNFLO, by adopting source and sink mechanism to determine organ biomasses. Parameters for four genotypes “Albena”, “Heliasol”, “Melody” and “Prodisol” are estimated based on two field experiments, one of which is under water deficit situation. SUNLAB computes more phenotypic traits than SUNFLO, such as all organ biomasses at a daily step. The model can be also used for the simulation of the specific leaf area variable. Specific leaf area (SLA) is the ratio of leaf area to dry leaf mass, which is usually an influent input variable often associated with large uncertainty ranges in most dynamic crop growth models [Rawson et al., 1987]. It is an important variable in plant growth modeling. In most dynamic models, it is usually used to determine blade surface area values from blade biomass, as in GREENLAB [Christophe et al., 2008] or in TOMSIM [Heuvelink, 1999]. Since blade area in turn determines the biomass production, accurate estimation of SLA is mentioned as a major source of error in models and implies difficulties in obtaining a reliable mechanistic computation of leaf area index, which is the main component of biomass production modules [Heuvelink, 1999; Marcelis et al., 1998]. It is however generally considered as constant, although it has been shown, for instance on wheat [Rawson et al., 1987], that SLA varies with genotypes, leaf ranks and leaf growing periods. Regarding sunflower,

the variations of SLA and the factors influencing them are still poorly known. As SUNLAB can simulate the dynamics of individual blade mass profiles independently from those of blade areas, the SLA can be computed as a model output, contrary to the classical situation in which it is taken as an input.

Parameter estimation. The second objective of our research is the conception of an original methodology for model parameter estimation, adapted to the context of plant variety improvement by breeders. In parameter estimation, an estimator takes the measured data as input and produces an estimate of the parameters, with an evaluation of the uncertainty on the parameter estimates (confidence or credible intervals). In crop models' parameter estimation, a specific problem is the large number of parameters compared to the amount of field data Makowski et al. [2006], which also causes data assimilation problem requiring expensive experiments for heavy data collections.

A methodology, multi-scenario parameter estimation methodology (MSPE), is designed to solve it. It uses a limited number of experimental traits but in a large number of environmental configurations for the parameter estimation by model inversion. While ecophysiological models of plant growth are widely researched to analyze genotype-by-environment interactions, the estimation of their parameters is a crucial issue in order to allow the discrimination between genotypes. In breeding programs, however, the amount of experimental trait data is usually not sufficient for accurate parameter estimation. MSPE takes advantage of the multi-environmental trials (potentially large amounts of environmental conditions available) set in place by breeders to evaluate the performances of their genotypes. The assumption is that such variety of discriminating scenarios should compensate for the little amount of information (data) for each scenario.

The methodology is tested on the SUNFLO model with theoretical data confirming its feasibility. Practical issues for carrying out MSPE are discussed. The first issue

is proposing priorities on the parameters to estimate with sensitivity analysis. The second includes ensuring the most appropriate numerical optimization methods for the model (Gauss-newton, Simulated Annealing and Particle Swarm Optimization methods are compared in the case study) and figuring out the best computation solution to coordinate with corresponding optimization methods. The use of the computing mesocenter of Centrale helps enhance the efficiency of the comutation for this purpose. The last issue is investigating parameters non-estimability problem under MSPE, resulting from model structural non-identifiability and scenario data's information insufficiency (practical identifiability). The hypothesis that "the increase of scenarios makes estimated parameters possessing better prediction ability" is proved by a simple test which increases scenario amount for parameter estimation to detect corresponding prediction error, and by a more rigorous test based on cross validation method, in which 20000 points are used to measure the prediction error of estimated parameters for a specific scenario amount. An extended version of MSPE, named MSPEJ, is the multi-scenario parameters estimation methodology based on delete-m Jackknife method. The interval estimator's feasibility is proved. Our tests indicate that parameter distribution variances are reduced along with the increase of scenario amount based on Jackknife samples. They are presented in Chapter 4.

Environmental protocols. Environment inputs are of course crucial determinants for the model and influence a lot model outputs. This obvious idea is used to improve the experimental protocol for parameter estimation by investigating the choice of the environmental configurations (the scenarios) for the MSPE methodology. In Chapter 5, environmental scenarios are clustered by hierarchical and centroid-based analysis respectively based on the environmental information including temperature, radiation, precipitation, and potential evapo-transpiration, based on their influences on plant growth features such as crop yield in this thesis, and based on the combination of both. The three clustering graphs are illustrated and the last strategy is recommended

since it clusters environmental scenarios taking into account both its environmental information and its influence on plant growth. Environment clusters can be used to optimize experiment design. A scenario in one cluster whose correlation is 0.9 can be recognized as the representative scenario for the cluster. Selecting only one representative scenario from each cluster in experiment design saves experimental cost and therefore the phenotype construction cost. MSPEE, a multi-scenario parameter estimation methodology based on environment clustering and scenario selection, is used to improve the efficiency of MSPE. With fewer scenario amounts, estimated parameters in MSPEE have the same good prediction ability than MSPE. Fewer scenario amounts in jackknife samples also produce the same variance than MSPE. Moreover, the system practical identifiability is improved.

Application. An illustration of how crop models can be applied for phenotypes analysis is also studied in this thesis. Jones et al. [2006] concluded that four most important applications of crop model are: prediction, the determination of optimal management, large spatial-scale applications, and the characterization of plant varieties and plant breeding. The project in Chapter 6 is an application involving all the four aspects. SUNFLO is used to produce large phenotypic traits of 20 sunflower genotypes across large geographies and over large time scales. In particular, crop water demand for irrigation and yield are investigated. The large geographies include 25 locations with diverse drought conditions in five European counties: France, Greece, Italy, Portugal and Spain which account for 12 million ha corresponding to 75% of the total area equipped for irrigation in EU. The large time scales include a real dataset from 1951 to 2011 and a prediction dataset from 2012 to 2100 (based climatic scenarios simulated by climatic models).

To sum up, this thesis aims at producing three types of knowledge. Firstly, it explores models which can well describe the genotype by environment interaction for a better understanding of phenotypes. Secondly, it researches on modeling analysis method-

ology to improve model parameterization. Thirdly, it is targeted to use developed models and methodology on real world phenotypes analysis and prediction.

Part I

MATERIALS AND MODELS

2. FIELD EXPERIMENTS AND DATA

Experimental data is essential in any modeling process, and especially for crop modeling. It is necessary for model design since a first step often consists in data analysis, before building a conceptual model. Models then need to be calibrated by confrontation to experimental data and model predictions should be evaluated against independent sets of data. Two kinds of data are considered in crop modeling: environmental data and crop data. Environmental data consists of weather data, soil features and crop management data. More precisely, the weather data considered in this thesis include maximum and minimum temperature, rainfall, relative humidity, solar radiation. Weather data are required at daily time steps to assess daily crop growth processes. Soil data include thickness of soil layer, soil texture, soil moisture, wilting point of soil, etc. Crop management data include date of crop sowing, irrigation, sowing density, etc. Crop data consist of the experimental measurements performed on the growing crop, such as leaf area, seeds biomass etc.

Data determines the effective boundaries of the model applicability. Models developed for a specific region may not be valid as such in another region. Proper parameter calibration and model validation is needed before using a model. For example, a sunflower model fitting to a 2001 French farming field may not work as as well in 2010, or in another farming location. Even for a model designed to fit general cases, it is necessary to know what data are used to support its universality. It is necessary to understand the data used to verify the hypothesis and to limit the models, theories and applications. In this chapter, we are going to introduce three databases of five datasets. They are used either in model design, model analysis, model validation, or

model application.

2.1 Experimental Data for Sunflower

2.1.1 Detailed Experimental Data for six Genotypes for two Years

This dataset is used for the calibration of the SUNLAB model in section 3.3. It includes three sub datasets, respectively called "2001", "2002a" and "2002b". They all come from field experiments conducted in 2001 and 2002 at SupAgro experimental station at Lavalette (43° 36'N, 3° 53' E, altitud 50 m) on a sandy loam soil for four genotypes "Albena", "Heliasol", "Melody" and "Prodisol". In "2001", Sunflowers were sown on 5 May 2001 at a density of about 6 plants m^{-2} and a row spacing of 0.6 m, in a randomized complete block design with four replications. Plots measured $5.5m \times 13.0m$. In the other two datasets, experiments were conducted with the same plant arrangement. But sunflowers were sown on 15 May 2002 with plots measured $8.0m \times 8.0m$. During the experiment, meteorological data such as temperatures and radiation were recorded. FTSW representing the available water in the soil was estimated. Organogenesis was described based on the phenomenological stages that are recorded every 2-3 days. Once a week, six plants per genotype were harvested. Individual leaf areas were estimated from blade lengths and widths. All the above-ground organs (leaves, stem, capitulum and seeds) were collected and then oven-dried at 80°C for 48 h. The dry weights of these organs were measured by compartments. Daily radiation interception efficiency $RIE(d)$ and daily radiation use efficiency $RUE(d)$ were respectively calculated and estimated based on field measurements [Lecoeur et al., 2011]. In all experiments, the crop was regularly irrigated and fertilized to avoid severe water deficits and mineral deficiency. But in practice, the three experiments

showed different water deficit conditions. The index $FTSW$ of the three experiments is represented in Fig.2.1. Since the experiment measurements were carried out every a few days, an interpolation on experimental data was drawn to better highlight the contrast.

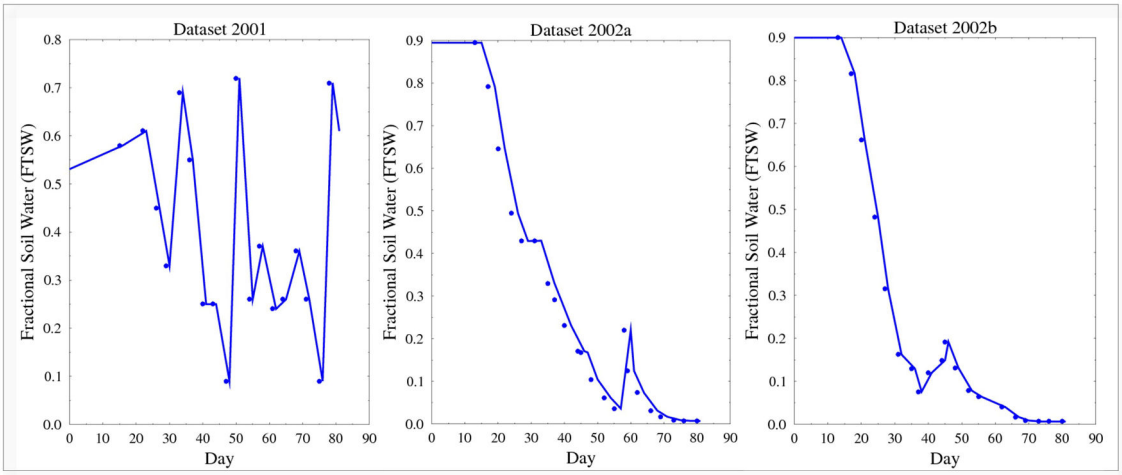

Fig. 2.1: $FTSW$ for three datasets "2001", "2002a", "2002b"

2.1.2 Sparse Experimental Data for 90 Genotypes for three Years

This dataset includes experimental data collected for 90 F1 hybrid sunflower genotypes. This F1 generation is the set of plant offsprings resulting cross matings of two parental lines. One of the 90 genotypes is the genotype "Melody" whose parameters has been estimated by the way of direct measurement in Lecoeur et al. [2011]. For each genotype, plants were grown under around 20 *scenarios*, chosen among a few locations and in three years 2008, 2009 and 2010. Each scenario contains environmental information including temperature, radiation, precipitation and evapo-traspirational reference. The corresponding crop data consists only of crop yield. This kind of experimental protocol, with sparse crop information but collected in many different environmental conditions (scenarios) is typically collected by breeding companies but is unusual for classical parameter estimation approaches. The multi-scenario parame-

ter estimation methodology we developed, MSPEJ, described in section 4.6.2, allows dealing with kind of information and was applied to one genotype, “Melody”, in order to test its feasibility. For time reasons, this result was not extended to all the 90 genotypes parameters, which should permit to analyse the linkage between quantitative trait loci and SUNFLO parameters. This future perspective is discussed in section 7.

2.2 Experimental Data for Corn

This experimental dataset, entitled as dataset 2.2, is build from experiments on 11 corn genotypes in around 10000 scenarios comprising around 1000 counties of around 60 American states in 10 years from 2001 to 2010. For each scenario, information about the environmental information including weather data and soil data, about the crop practices such as sowing density, date and harvest date, and about the crop yield are available. Among them, the 10 years daily weather data were obtained from the database of Syngenta Corporation. Soil data were extracted from the soil survey geographic database (SSURGO) produced by the Natural Resources Conservation Service of United States Department of Agriculture (USDA). They have diverse drought status. Crop practices are also obtained from USDA. Yield data are from National Agricultural Statistics Service of USDA. We used the 720 experimental scenarios with full irrigation for one genotype to test MSPE and MSPEE parameter estimation methodology in Part II. Then all the available scenarios are used to test the method based on environment clustering in Chapter 5.

2.3 Additional Environmental Data

2.3.1 36 Years French Weather Data

To provide meaningful information about the performance of plants in a certain environmental context, a set of environmental information needs to be recorded to analyze genotype × environment interactions. This dataset contains the record of environmental information, including minimal temperature, maximal temperature, radiation, precipitation, and evapo-transpiration reference on daily basis in Toulouse, France. The data is available whole year round from 1971 to 2007. It is used for the theoretical study of multi-scenario parameter estimation methodology (Chapter 4) to produce corresponding plant phenotypic traits simulations.

2.3.2 Large Scale Environment Database

To illustrate the model potentials for yield predictions at large scales and its ability to discriminate genotype performances, relevant climatic scenarios are required. These were obtained from an open source dataset "ENSEMBLES", which is funded by the EU FP6 Integrated Project (Contract number 505539). Its climate prediction system is based on the principal state-of-the-art, high resolution, global and regional Earth System models developed in Europe. It is validated against quality controlled, high resolution gridded datasets for Europe, to produce for the first time, an objective probabilistic estimate of uncertainty in future climate at the seasonal to decades and even longer timescales. For any European point coordinate of longitude and latitude, this climate prediction system produces a variety of possible weather information, such as wind, humidity, cloud cover, snow depth etc. In the context of our crop modeling research, only the variables of temperature, radiation, precipitation and evapo-transpirational reference are needed. We picked up 25 locations in five European

counties - France, Greece, Italy, Portugal and Spain - and large time scales - from 1951 to 2100 - for our phenotype analysis application in Chapter 6.

3. PLANT GROWTH AND BREEDING MODELS

Plant or crop models, that use systems approach to simulate the interaction between crops and environment, is an important tool to assimilate knowledge gained from field experiments, to promote the understanding of biological system behaviors and underlying eco-physiological functions, and to supply mathematical analysis for solving agriculture and biological problems. This chapter gives a description of the models considered or developped in the thesis. CORNFLO is a growth model for corn (*Zea mays* L.) and will be deeply analysed in Part II. SUNFLO is a sunflower growth model (*Helianthus annuus* L.): it will be analysed and compared with other sunflower models, and used for applications. SUNLAB is a new model developed in this thesis in order to expand SUNFLO abilities for phenotype analysis.

3.1 Principles of Crop Growth Modeling

3.1.1 Objectives and Constraints of Model Design in a Breeding Context

Context: models to guide the breeding process. Generating timely, robust, reliable and useful information about complex biological systems is the key to address many of the world's most pressing policy concerns in diverse areas: public health,

human and animal disease, food production, and ecological conservation [Tsaftaris and Noutsos, 2009]. Modeling is a modern approach to provide such information. A mathematical model is a description of a system using mathematical concepts and language. It can help to explain a system and to study the effects of different exponents, and to make predictions about behaviors. More specifically, crop models aim at describing and understanding one of the most important biological cycles: the interaction between crop genotypes and agricultural environment.

Assessments of genotype performances in *in situ* experimental trials hamper the breeding process by temporal, logistic and economical difficulties. Indeed, genotypes perform differently depending on the environmental conditions (soil, climate, etc.) and the management practices (sowing date, nitrogen inputs, irrigation, etc.). Therefore a large number of trials are needed to explore a sufficiently diverse set of genotypes x environment x management (GxExM) combinations in order to characterize these complex interactions. The emerging approach to overcome these difficulties relies on the use of models that determine the plant phenotype in response to environmental inputs. These models should simulate the phenotypic traits of interest (e.g. yield) with good robustness and predictive capacity. They should also present a trade-off between mechanistic aspect and complexity: Chapman et al. [2003] state that, for such use, a growth model should include 'principles of response and feedbacks' to 'handle perturbations to any process an self-correct, as do plants under hormonal control when growing in the field' and to 'express complex behavior even given simple operational rules at a functional crop physiological level'. For the analysis of phenotypes, it is expected that crop models can faithfully enough reproduce the wide range of phenotypic responses for various genotypes in various environments. They should provide insights on the causal chain of processes that produce a given phenotypic trait and help deciphering the relative contributions of different environmental factors. Once properly calibrated and validated, these models could be used to guide future studies

on improving agricultural practices and breeding, or to examine the adaptation of given genotypes to some target environments and to provide recommendations. Consequently, an important question is to identify what kind of models can be used in that context.

Choice of Model Class. Crop models can be defined as a system representation of crops. Several classifications can be proposed:

- Static vs. dynamic: Static models do not account for the time variable and describe a system at an equilibrium or steady state (or at least at a given time point); on the contrary in dynamic models, states and outputs of the described system can change with respect to time.
- Discrete vs. Continuous: models can be written under discrete or continuous formulations depending on the set of definition of their space-time variables. In particular, for dynamic models, this sub-classification is defined by the time variable that can be $(t_n)_{n\in\mathbb{N}}$ or $t \in \mathbb{R}$. This choice leads to writing the model under the form of recurrence or differential equations.
- Deterministic vs. stochastic: deterministic models produce the same outputs for a given set of inputs, while stochastic models include some random variables that introduce some non-predictable effects (variability of the outputs can be described through various statistics, e.g. probability distribution, mean, variance).
- Empirical vs. mechanistic: empirical models (or descriptive models) are derived on direct descriptions of observed data. They are usually regression based and provide a quantitative summary of the observed relationships among a set of measured variables [course of V. A. Bokil, Department of Mathematics, Oregon State University, MTH 323: Spring 2009]. Mechanistic models (or explanatory models) generally arise from approaches relating to the complex system theory: they consider the individual components of the system and their interactions,

and what emergent properties appear.

With the different characteristics come different advantages and drawbacks. For example, deterministic models do not allow to represent the different sources of residual variability, which are actually inherent to biological and agricultural systems [Brockington, 1979]. This might look unsatisfactory when variability is an important component of model outputs, e.g. in rainfall prediction, or if the degree of uncertainties or of unexplained variations reaches a high level. However, stochastic models tend to be technically difficult to handle and can quickly become complex. Moreover, they can lack some explanatory properties, if random variables are introduced in place of more mechanistic modules, to describe some processes whose internal mechanisms are unknown or voluntarily ignored in the modeling work. Therefore, in certain cases, deterministic models may be adequate despite the intrinsic variability of biological phenomena. Regarding the choice between empirical or mechanistic models, it is obvious that most models are in fact made of a mixing of these two approaches. Pure empirical models are mere interpolations of observation data and should be used only in the range of conditions over which they have been derived [SINCLAIR and SELIGMAN, 1996]: it is advisable to avoid extrapolation. For instance, under contrasting conditions, the above water use efficiency-cane yield relationship may not hold [Keating et al., 1999]. In general, mechanistic models are often more useful, as they consist of a quantitative formulation of a set of hypotheses [Wells, 1992] and as they can be used out of their calibration interval (provided that the model predictive capacities have been preliminarily checked). However, the consequences of using an inappropriate mechanistic model are worse than for empirical models because the parameters in mechanistic models provide information about the quantities and properties of real system components. Thus, the appropriateness of mechanistic models needs close scrutiny [Christopoulos and Michael, 2000]. For applications, the choice of a model class is complicated and depends on the project objectives. In this thesis, the studied

models are all dynamic, discrete, mechanistic and deterministic. We choose them to be dynamic and discrete because we are interested in daily simulation of crop traits. They are mainly deterministic because we want to well simulate the statistical average of crop features and the uncertainty of system and environmental data are comparably not important for the current phase. We aim to improve the understanding of crop system mechanisms which are also the most crucial for the analysis of phenotype, genotype, and environment interactions.

Global flowchart of a discrete dynamic deterministic crop model including a genetic module. For applications in the breeding context, the adequate models should be able to take into account a representation of the genomic regions associated with variability in the complex traits of interest [Hammer et al., 2006]. In this thesis, the studied crop models consider plants as dynamic systems and aim at simulating the relative contributions of its genotype and of the environmental conditions in the constitution of its phenotype. These models should prove their ability to discriminate different genotypes by different parameter sets that should be shown to remain stable under varying environmental conditions. They should simulate plant genotypic responses to environmental variations by describing crop eco-physiological functions with mathematical equations. The generic formulation of a dynamic deterministic plant system model in discrete time can write as:

$$X(t + \Delta t) = g(X(t), U(t), \theta) \tag{3.1}$$

where t is time, Δt is some time increment, $X(t) = [X_1(t), \cdots, X_s(t)]$ is the vector of state variables, representing the plant phenotypic characteristics at at time t, $U(t)$ is the vector of explanatory variables, representing environmental information as input to dynamic plant system at time t, θ is the vector of parameters, representing biophysical parameters of plant genotypes and g is the set of system functions, representing the interaction of plant genotypes confronted to environmental input to produce the

output of plant phenotypic performances.

This can be illustrated through the diagram in Fig. 3.1: $U(t)$ is the environmental input including weather, density, soil information, etc. An example of the set of biophysical functions, g, is represented through three modules including organogenesis, biomass production and partitioning. There are multiple ways to construct each module, depending on the considered crop models and output scales. A set of model parameters θ is representing a genotype: the values of these parameters can be the output of a 'genetic' model whose inputs are genetic information such as quantitative trait loci (QTL). This kind of modeling approach can simulate the responses of virtual plants carrying diverse combinations of alleles under different scenarios of abiotic stress.

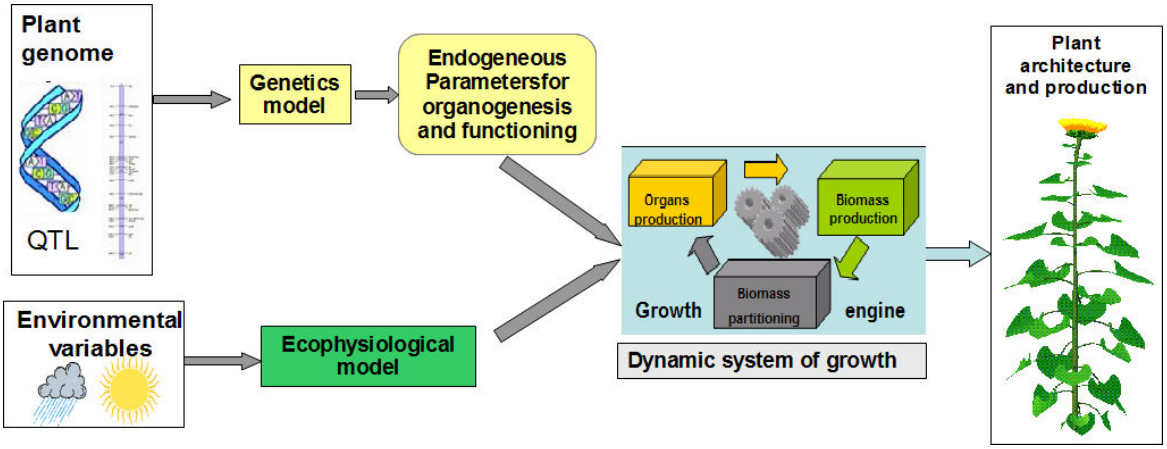

Fig. 3.1: Flowchart of plant growth modeling.

The main difficulty is to mathematically express the genetic variability of responses to environmental conditions. Modeling via gene regulatory networks is not feasible for such complex systems, but plants can be modeled using response curves to environmental conditions that are 'meta mechanism' at plant level. Each genotype is represented by a set of response parameters that are valid under a wide range of conditions. Transgenesis of one function experimentally affected one response parameter only. Transgenic plants or plants carrying any combination of quantitative trait

loci might therefore be simulated and tested under different climatic scenarios, before genetic manipulations are performed [Tardieu, 2003].

3.1.2 Methodologies and Mathematical Tools to Develop Crop Models

The basic procedures for developing any model involve model design, model calibration, and model validation. Following these procedures, models in this chapter are designed to fulfill specific objectives, are calibrated to confront to experimental data and are validated to assess their performances or robustness and to define their usage scope.

Model Design. Strategies for model design depend on the modeler's objectives. The models considered in this thesis are designed to fulfill different combinations of objectives.

Three models are presented and studied in this chapter: two existing models, CORNFLO and SUNFLO [Lecoeur et al., 2011], are analysed, and a new model, SUNLAB [Kang et al., 2012a], is developed. These models are designed to simulate plant growth and physiological functions under drought stress. To this end, a current approach consists in building functional-structural plant models (FSPM), which combine two traditional perspectives (emphasizing either plant function or plant architecture). FSPMs have several advantages, such as their ability of capturing subtle differences in resource allocation or structural growth and their consequences for future seedling performance [Sievanen et al., 2000]. FSPMs also realistically represent the spatial distribution of plant organs, which is an important aspect in whole-plant resource uptake [Kellomaki et al., 1985]. Yin and Struik [2010] recommend that crop models should be upgraded based on understanding at lower organizational levels for complicated phenomena such as sink feedback on source activity. In this chapter, we will present how the SUNLAB

model extends SUNFLO by introducing a feedback of allocation processes on the production. The modules describing the modelling of water deficit effects are detailed in section 3.2.5.

Another of our modeling objectives is related to the ability of these models to discriminate genotypes through variations in their parameter values [Jeuffroy et al., 2006a]. This property was analysed for SUNFLO in [Lecoeur et al., 2011] and is explored for the new model developed in this thesis, SUNLAB, in section 3.3. When designing a model, a key question is to determine whether parameters should have a biological meaning. According to Yin and Struik [2010], parameters of many current crop models only have little biological meaning: they suggest building less empirical models by exploiting the existing physiological understanding of the growth processes and by employing mathematical tools. The recent advances of functional genomics and systems biology enables the elucidation of the molecular genetics bases of different processes and of the link between so-called "genetic coefficients" and model parameters, thus showing the promises of using models in analyzing genotype-phenotype relationships of some crop traits. Most parameters of CORNFLO, SUNFLO, and SUNLAB possess biological meaning, which should facilitate these models future uses on finding parameters x QTL linkage to narrow genotype-phenotype gaps.

Model Calibration. Model calibration is the process of setting the values of the model parameters. It is based on experimental data that are collected at given time points and under particular sets of environmental conditions. The collected data are expected to be in adequacy with the modeller's choice in terms of variables and modelling scales. For plant growth models, two popular approaches coexist for model calibration: estimation through direct measurements (for the parameters having biological meanings and that are directly observable) and estimation through mathematical methods (for the so-called hidden 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. However, this method often requires specific trials and measurements, and 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 [Jeuffroy et al., 2006a]. CORNFLO and SUNFLO parameters were estimated in this way: SUNFLO parameters' values are given in section 3.2.6.

The indirect method, involving mathematical and statistical methods, estimates one or more parameters by confronting observed data to simulation results. Its main advantage is that it can be experimentally less costly and time-consuming than the direct measurement of parameters [Jeuffroy et al., 2006a]. For instance in most dynamic models, the direct measurement method would often require frequent measurement points (e.g. daily), while with the indirect method, data can be collected only at some given time points and still allow the modellers to retrieve the past growth of the crop. Parameters can even be estimated from very limited sets of data, as shown in this thesis with the estimation methodology (MSPE) we developed. Here, a frequentist approach is adopted (*i.e.*, it is assumed that parameters are not random variables and that there exist 'true' fixed values. No *a priori* information is taken into account, except possibly by adding constraints or boundaries, in contrast with the Bayesian approach). Technically speaking, it includes several different methods, such as computing the least square error estimators or maximum likelihood estimators. Particular attention must be paid to possible correlations existing between parameters, which may produce estimator values which are satisfactory for prediction under a limited range of conditions only. Least square estimation is used for SUNLAB parameters, with the Gauss-Newton algorithm for optimization of the cost functions (section 3.3.5). In Part II, two other optimization algorithms are also used for the

tests: Simulated Annealing and Particle Swarm Optimization, as described in section 4.3.2.

Model evaluation and validation. Model evaluation and validation are important steps since non-validated models may lead to wrong decisions. These include several aspects. One of them relies on sensitivity analysis: in that context, it can be used to detect over-parameterization, for selecting the order of priority for parameters to be estimated, or for analysing the model behaviour (sections 3.3.4 and 4.3.1).

The process of parameter estimation raises the problem of the continuity and convexity of the objective function to the model parameters. Model identifiability and continuity analysis are presented in section 4.3.1 and are used for selecting the adequate optimization methods to use for given parameters. Finally, an important aspect of model evaluation consists in testing its predictive ability. To this end, a set of experimental data, distinct from the one used as target for parameter estimation, should be collected. In this thesis, SUNLAB validation is performed in section 3.3.6) and then in Part II, squared residuals are examined to produce prediction squared error. We adopted two methods: the classical method based on an independent sample validation data from the sample population as the training data, and the cross validation method, explained in section 4.5.2. The details of these methodologies will be elaborated when it is used in corresponding sections.

3.2 SUNFLO Model for Sunflower

The SUNFLO model for sunflower (*Helianthus annuus* L.) consists of five modules: Phenology Module, Architecture Module, Biomass Production Module, Biomass Allocation Module, and Water Budget Module [Lecoeur et al., 2011]. It estimates the biomass production for the crop sunflower under environmental inputs, mainly temperature, precipitation, and evapotranspiration reference. It simulates the plant phe-

nology and development, the accumulation and distribution of biomass, and the production of seeds. It takes into consideration the plant water budget which determines whether the available water quantity is enough for the plant to grow up in good conditions. A table of parameters can be found in section 3.2.6. In this thesis, it is used in Part II for model analysis and in Part III for model applications.

The CORNFLO model is a functional plant growth model simulating the growth and yield of maize (*Zea mays* L.). It is developed by Jérémie Lecoeur in Syngenta Seeds Corp. This model is used in all chapters in Part II for testing parameter estimation strategies. It has five same modules as SUNFLO model and its module formulas have many similarities with SUNFLO. are not elaborated in this thesis.

3.2.1 Phenology Module

The Phenology Module simulates the timing of the plant growth stages and how these are influenced by seasonal and interannual variations in climate. Daily average temperature $Tmoy(d)$ (℃ days) is transformed into daily effective temperature $Teff(d)$ by subtracting a base temperature $Tbase$ which is 4.8 ℃ for sunflower genotypes. It therefore models the effect of thermal stress on plant development and functions:

$$Teff(d) = Tmoy(d) - Tbase \tag{3.2}$$

A variable defined as 'phenology accelarator', $AP(d)$ (Eq. 3.3), depends on $Teff(d)$ and it is dampened by water stress constraint $FHTR(d)$ (Eq. 3.43) on day d:

$$AP(d) = 0.1 * Teff(d) * (1 - FHTR(d-1)) \tag{3.3}$$

This variable $AP(d)$ together with $Teff(d)$ intervenes in the calculation of the accu-

mulated thermal time $TT(d)$ (Eq. 3.4):

$$TT(d) = \sum_{t=0}^{d} Teff(t) + AP(t) \tag{3.4}$$

$TT(d)$ is a significant variable determining four plant key physiological stages, expressed as genotype dependent thermal dates: flower bud appearance (TT_E1), beginning of flowering (TT_F1), beginning of grain filling (early maturation, TT_M0) and physiological maturity (TT_M3). These thermal dates trigger some variations of plant functions through plant growth periods, such as the emergences of leaf, capitule, seed etc. (e.g. eq. 3.5) and biological efficiency in different periods (Eq. 3.17).

3.2.2 Architecture Module

The thermal time of blade emergence $TI(i)$ at rank i depends on two parameters: phyllochron $Phy2$ and LAI_a.

$$TI(i) = (i - 5) * Phy2 + LAI\_a \tag{3.5}$$

The thermal time of capitulum emergence is denoted TT_E1. The thermal time of seed initialization is denoted TT_M0. When $TT(d)$ reaches an organ initialization thermal time, the organ emerges.

To calculate the leaf area expansion curve $GRe(i, d)$ for leaf at rank i on day d, we need to calculate three variables: the maximal expansion speed $Ae(i)$, the spread of leaf area expansion curve $Ke(i)$, and the thermal time at which this maximal expansion rate is reached for each rank $Te(i)$. The total number of leaves is denoted as NFF. The leaf which has the maximal potential leaf area $SFiMax$ is located at rank $position\_SFiMax$.

$Ae(i)$ is calculated based on the three parameters.

$$\begin{aligned} b &= 1.5 - 0.22 * position\_SFiMax - 0.0035 * SFiMax + 0.08 * NFF \\ a &= -2.3 + 0.019 * position\_SFiMax - 0.0016 * SFiMax + 0.02 * NFF + b * 0.92 \\ Ae(i) &= SFiMax * \exp\left(a * (\tfrac{i - position\_SFiMax}{position\_SFiMax - 1})^2 + b * (\tfrac{i - position\_SFiMax}{position\_SFiMax - 1})^3\right) \end{aligned} \tag{3.6}$$

The spread of leaf area expansion curve $Ke(i)$ is defined as:

$$Ke(i) = \begin{cases} 0.01 & i < 7 \\ LAI\_Kei & i \geq 7 \end{cases} \tag{3.7}$$

where LAI_Kei is a parameter. And the thermal time at which the maximal expansion rate is reached for each rank, $Te(i)$, depends on $Ke(i)$ as follows:

$$Te(i) = \begin{cases} TI(i) + 70 & i < 7 \\ TI(i) + LAI\_b/Ke(i) & i \geq 7 \end{cases} \tag{3.8}$$

where LAI_b is a parameter. The illustrations of leaf area expansion curves $GRe(i, d)$ for leaves at different ranks for genotypes "Melody" and "Albena" are given in Fig. 3.2.

$$GRe(i,d) = Teff(d) * Ae(i) * Ke(i) * \exp\left(\frac{-Ke(i) * (TT(d) - Te(i))}{1 + \exp(-Ke(i) * (TT - Te(i)))^2}\right) \tag{3.9}$$

Dampened by water stress constraint $FHLE$ (Eq. 3.41) and radiative constraint FLe (Eq. 3.16) that will be defined later, the accumulation of $GRe(i, d)$ constructs the leaf surface $SFe(i, d)$ of leaf i on day d:

$$SFe(i,d) = \sum_{t=0}^{d} (GRe(i,t) * FHLE(t) * FLe(t)) \tag{3.10}$$

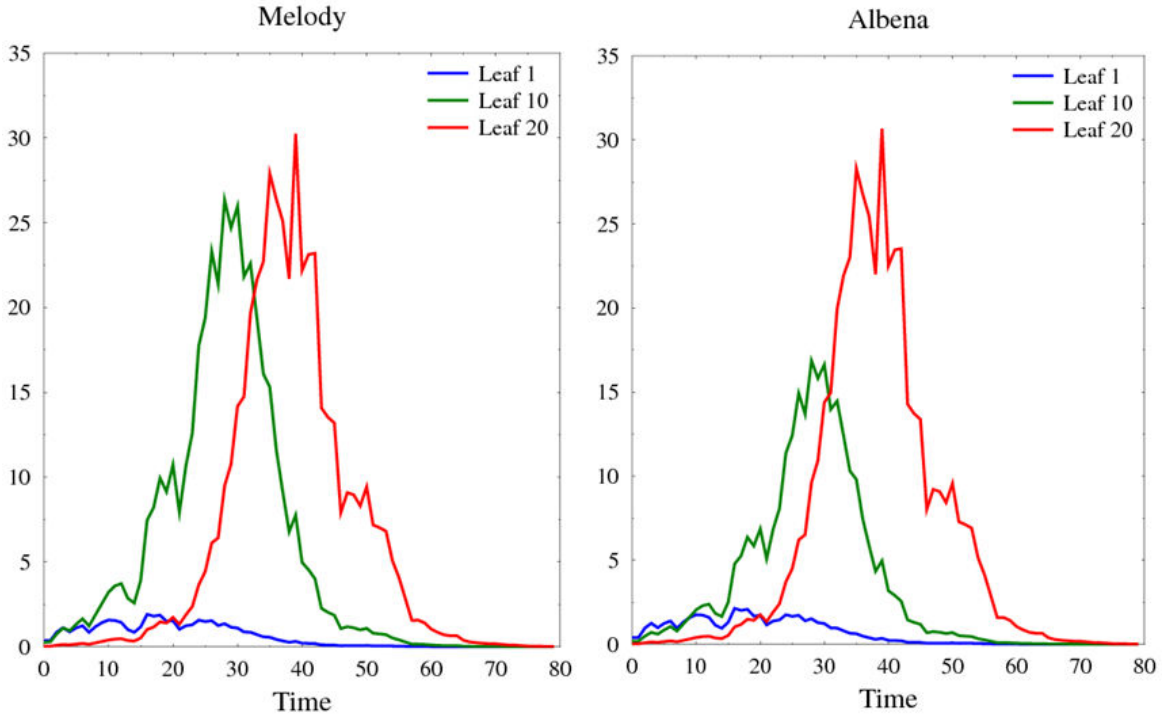

Fig. 3.2: Leaf area expansion curve (GRe) for leaf ranks 1, 10, 20 of genotypes "Melody" and "Albena".

The number of dead leaves $NFmortes$ is a linear function of thermal time $TT(d)$:

$$NFmortes(d) = NFF * \frac{TT(d) - TT\_F1}{TT\_M3 - TT\_F1} \tag{3.11}$$

Therefore the active leaves surfaces $SF(i,d)$ include only leaves that are not yet senescent:

$$SF(i,d) = \begin{cases} 0 & i \leq NFmortes(d) \\ SFe(i,d) & i \succ NFmortes(d) \end{cases} \tag{3.12}$$

The sum of living leaf areas $SF(i,d)$ gives the total efficient plant area $SFp(d)$:

$$SFp(d) = \sum_{i=0}^{NFF} SFe(i,d) \tag{3.13}$$

3.2.3 Biomass Production Module

As in Cornflo, the plant daily dry biomass production $DBP(d)$ is estimated by the energetic approach of Monteith Monteith [1977] as a multiplicative function of radiations $Rg(d)$, radiation absorption efficiency $Ea(d)$, radiation use efficiency $Eb(d)$, and a climatic efficiency which is taken equal to 0.48:

$$DBP(d) = 0.48 * Rg(d) * Ea(d) * Eb(d) \tag{3.14}$$

$Ea(d)$ is simulated from the Beer-Lambert law as a function of the leaf area index $LAI(d)$, which is calculated from total active leave surface area $SFp(d)$ and plant density to simulate the plant capacity to intercept radiation, and an extinction coefficient $coEff$ determined for each genotype:

$$\begin{aligned} LAI(d) &= SFP(d) * density \\ Ea(d) &= \left(1 - e^{-coEff * LAI(d)}\right) * 0.95 \end{aligned} \tag{3.15}$$

Daily incident photosynthetically radiation $PARi(d)$ is used to determine the radiation constraint $FLe(d)$, which influences the leaf surface expansion (see above in equation 3.10). These variables are calculated as:

$$\begin{aligned} PARi(d) &= 0.48 * Rg(d) * Ea(d)/(SFP(d) * density) \\ FLe(d) &= -0.139 + \frac{2.82}{1 + \exp\left(-\frac{PARi(d) - 4.134}{2.093}\right)} \end{aligned} \tag{3.16}$$

The potential radiation use efficiency $Ebp(d)$ represents the plant potential ability to use the radiation after absorption. It varies depending on the phenological stages:

$$Ebp = \begin{cases} 0 & TT(d) = 0 \\ Eb\_0 & 0 \leq TT(d) \prec 300 \\ Eb\_0 + \frac{(TT(d)-300)*2}{TT\_F1-300} & 300 \leq TT(d) \prec TT\_F1 \\ Eb\_Max & TT\_F1 \leq TT(d) \prec TT\_M0 \\ Eb\_fin * \exp\left(Eb\_c * (1 - \frac{TT(d)-TT\_M0}{TT\_M3-TT\_M0})\right) & TT\_M0 \leq TT(d) \prec TT\_M3 \\ 0 & TT\_M3 \leq TT(d) \end{cases} \tag{3.17}$$

where Eb_0, Eb_c, Eb_Max, Eb_fin are parameters estimated in Lecoeur et al. [2011]. Fig. 3.3 is an illustration of Ebp for genotypes "Melody" and "Albena". Eb

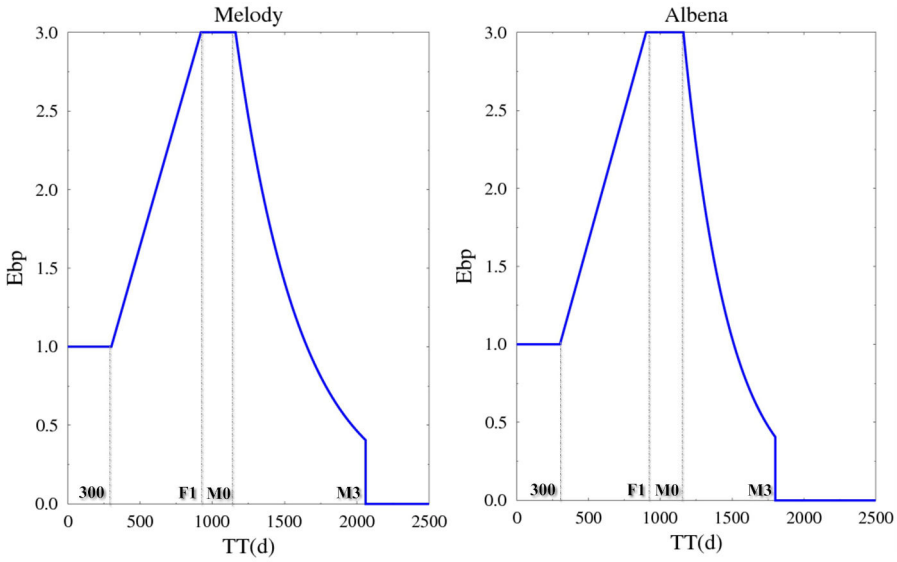

Fig. 3.3: Potential radiation usage efficiency(Ebp) for genotypes "Melody" and "Albena".

(Eq. 3.19) is calculated based on $Ebp(d)$, water constraint on radiation use efficiency $FHRUE$ (Eq. 3.42), PHS which is the genotypic parameter of the photosynthesis capacity compared with the genotype "Melody", and a thermal factor $FT(d)$ (Eq. 3.2.3).

$$FT(d) = \begin{cases} 0 & Tmoy(d) \leq Tbase \\ \dfrac{Tmoy(d)}{Topt1PHS - Tbase} - \dfrac{Tbase}{Topt1PHS - Tbase} & Tbase \prec Tmoy(d) \leq Topt1PHS \\ 1 & Topt1PHS \prec Tmoy(d) \leq Topt2PHS \\ \dfrac{Tmoy}{Topt2PHS - TmaxPHS} - \dfrac{TmaxPHS}{Topt2PHS - TmaxPHS} & Topt2PHS \prec Tmoy(d) \leq TmaxPHS \\ 0 & Tmoy(d) \geq Topt2PHS \end{cases} \quad (3.18)$$

where $Topt1PHS$, $Topt2PHS$, $TmaxPHS$ are parameters estimated in Lecoeur et al. [2011]. Fig. 3.4 is the illustration of FT for sunflower genotypes.

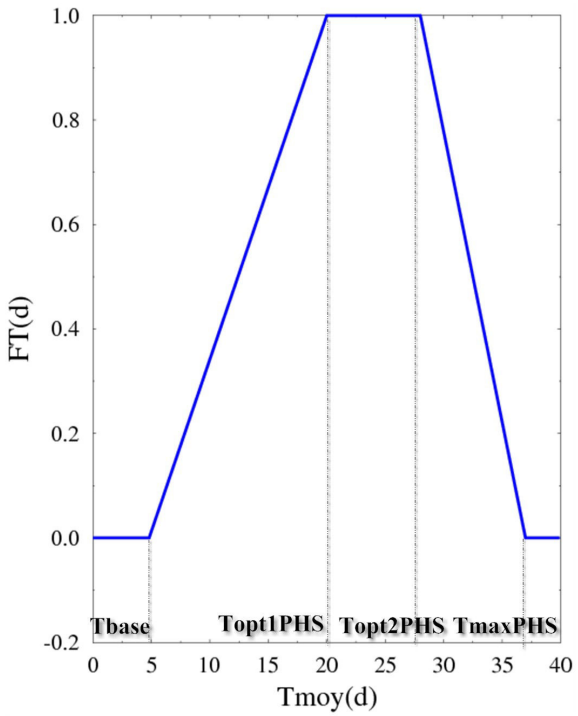

Fig. 3.4: Thermal factor for sunflower genotypes.

$$Eb(d) = Ebp(d) * FHRUE(d) * FT(d) * PHS \tag{3.19}$$

Based on $DBP(d)$, we get plant total dry biomass $TDM(d)$:

$$TDM(d) = \sum_{t=0}^{d} DBP(t) \tag{3.20}$$

3.2.4 Biomass Allocation Module

The total dry biomass $TDM(d)$ is allocated to the capitule by a linear relationship with the harvest index $HI_{capitule}(d)$ to get the capitulum biomass $MS_{capitule}(d)$:

$$HI_{capitule}(d) = \frac{0.632}{1 + (\frac{TT(d) - TT\_E1}{774})^{-2.827}} \tag{3.21}$$

$$MS_{capitule}(d) = HI_{capitule}(d) * TDM(d) \tag{3.22}$$

3.2.5 Water Budget Module

The water cycle of sunflower is mainly modeled through processes of root water absorption and transpiration from the plant side, and precipitations, irrigation, and soil evaporation from the environment side (see Fig. 3.5a).

To model water stress, an index is defined as the fraction of transpirable soil water $FTSW(d)$, taking values from 0 (no water stress) to 1 (severe water stress). It depends on the interaction of the root system with the environmental factors that include soil characteristics (namely particle size on each horizontal layer, humidity capacity and soil density), soil evaporation, precipitations and irrigation. Evaporation and plant transpiration decreases the available amount of water in soil. The calculation of $FTSW$ is done through the following steps:

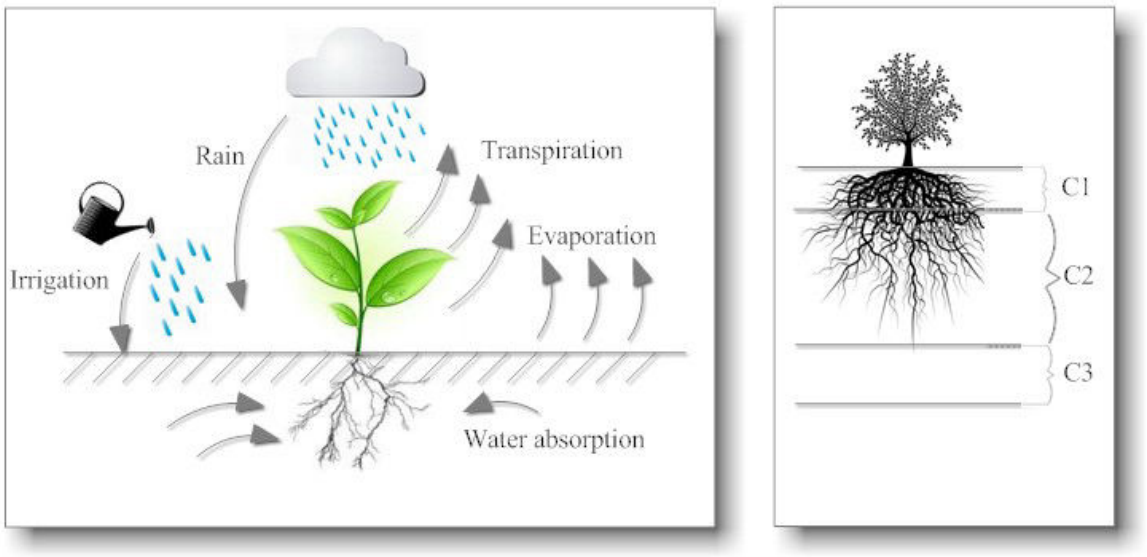

Fig. 3.5: Water budget module in SUNLAB: (a) Left: processes considered in the water cycle model; (b) Right: the three soil layers C1, C2 and C3.

1. Root elongation Root depth $zRoot(d)$ (cm) increases with a daily ratio $dRoot(d)$ as in equation 3.23:

$$\begin{aligned} dRoot(d) &= 0.7 \times Tmoy(d) \\ zRoot(d) &= zRoot(d-1) \times dRoot(d) \end{aligned} \tag{3.23}$$

2. Definition of three soil layer Soil is modeled into three layers: $C1$, $C2$ and $C3$, as shown in Fig.3.5b. The depth $dC1$ (mm) of $C1$ is fixed to 300 mm. The thickness $dC2$ of layer $C2$ is determined by root length: it is initialized at $1mm$ and equals root depth $zRoot(d) - dC1$ once $zRoot(d)$ becomes larger than $dC1$. For the last layer, the maximal soil depth that needs to be considered for modelling the elongation of sunflower root system is assumed equal to $1800mm$ so that $dC3 = 1800 - dC1 - dC2$. Effective soil depth $z(cm)$ is equating $zRoot(d)$.

3. Maximal water content at depth z The maximal soil water content at depth z, expressed in $g.cm^{-1}$, is denoted $MSW(z)$ and is defined as:

$$MSW(z) = (Hcc - (Hpf * IEgen))/100 \times da \times z \tag{3.24}$$

where $MSW(z)$ depends the maximal soil water content per soil depth, which is determined by soil humidity capability Hcc (%), the humidity at permanent wilting Hpf (%), bulk density da ($g.cm^{-3}$) and an index of water extraction by the plant $IEgen$.

4. Available water content in each soil layer The available soil water $ASWCi(d, z)$ ($g.cm^{-1}$) is computed for each soil layer Ci, $i = 1, ..., 3$. Their calculations depend on the calculations of evaporation $EV(d)$ ($g.cm^{-3}$) in the first layer, and transpiration in the first $TRC1(d)$ ($g.cm^{-3}$) and second layer $TRC2(d)$ ($g.cm^{-3}$). Evapotranspiration is the loss of water from a vegetated surface through the combined processes of soil evaporation and plant transpiration. Water lost through soil evaporation passes directly from the soil to the atmosphere. But water lost by transpiration must enter the plant via the roots, then pass to the foliage where it is vaporized and lost to the atmosphere through leaf stomata. The evapotranspiration process is influenced by multiple factors such as plant type, plant development stages and weather.

In the simulation of soil evaporation $EV(d)$, $CumEVjDebut(d)$ is the cumulated water lost through soil evaporation. Its value is cleared out if the daily precipitation is big enough. The threshold value is determined by a soil-dependent parameter $Q0$ (in an environment scenario equating 9 for example):

$$CumEVjDebut(d) = \begin{cases} EV(d) & if \quad Rain(d) > CumEVjDebut(d-1) \\ & or \quad Rain(d) > CumEVjFin(d-1) + Q0 \\ CumEVjDebut(d-1) + EV(d) - Rain(d) & \\ & if \quad Rain(d) \leq CumEVjDebut(d-1) \end{cases} \quad (3.25)$$

The duration of $CumEVjDebut(d)$ from 0 to $Q0$ is called a “plateau”.

$CumEVjDebut(d)$ records the water lost from the beginning of a plateau and $CumEVjFin(d)$ records the water lost from the end of a plateau. It means $CumEVjFin(d)$ begins to count when $CumEVjDebut(d)$ reaches $Q0$, and $CumEVjFin(d)$ is cleared out when daily precipitation is as big as the $CumEVjFin(d)$ value.

$$CumEVjFin(d) = \begin{cases} 0 & if \quad ksEVj(d) = 1 \\ CumEVjFin(d-1) + EV(d) - Rain(d) & otherwise \end{cases} \quad (3.26)$$

The bigger $CumEVjFin(d)$ is, the lower $EV(d)$ is: this results from the influence of $CumEVjFin(d)$ on variables $DSW(d)$ and $ksEVj(d)$. $ksEVj(d)$ is the evaporation coefficient depending on the value of $DSW(d)$ which records the day without water supply from the beginning of the plateau. The evaporation coefficient $ksEVj(d)$ is reducing as $DSW(d)$ grows. In the end, $EV(d)$ (Eq. 3.29) is obtained by evaporation coefficient $ksEVj(d)$, radiation interception efficiency $Ea(d)$, and reference crop evapotranspiration $ETref(d)$.

$$DSW(d) = \begin{cases} 0 & if \quad Rain(d) > CumEVjFin(d-1) \\ DSW(d\ -\ 1)\ +\ 1 & otherwise \end{cases} \quad (3.27)$$

$$ksEVj(d) = \begin{cases} 1 & if \quad CumEVjDebut(d-1) \prec Q0 \\ \sqrt{DSW(t)+1} - \sqrt{DSW(t)} & \\ & if \quad Rain(d) \leq CumEVjFin(d-1) \\ 1 & otherwise \end{cases} \quad (3.28)$$

$ETref(d)$ is the estimation of the evapotranspiration from a reference surface, namely an extensive, hypothetical grass reference crop with specific characteristics [de Bruin et al., 2010]. It is a day-by-day environmental input in the model.

$$EV(d) = ksEVj(d) * ETref(d) * (1 - Ea(d)) \quad (3.29)$$

To calculate the transpiration $TRC1(d)$ and $TRC2(d)$, we need to calculate a variable $partC1(d)$ which is the proportion of the depth of the first layer $dC1$ to those of layer1 and layer2:

$$partC1(d) = \begin{cases} 1 & zRoot(d) \prec 300 \\ \frac{dC1}{dC1+dC2} & otherwise \end{cases} \quad (3.30)$$

The transpiration potential speed $vTRp(d)$ is derived from radiation interception efficiency $Ea(d)$ and reference crop evapotranspiration($ETref(d)$):

$$vTRp(d) = 1.2 * ETref(d) * Ea(d) \quad (3.31)$$

Accordingly, $TRC1(d)$ and $TRC2(d)$ are determined by $partC1(d)$, $vTRp(d)$ and the

constrain $FHTR(d)$ (Eq. 3.43) of water stress on transpiration:

$$\begin{aligned} TRC1(d) &= partC1(d) * vTRp(d) * FHTR(d) \\ TRC2(d) &= (1 - partC1(d)) * vTRp(d) * FHTR(d) \end{aligned} \tag{3.32}$$

The soil water available in the first layer $ASWC1(d, z)$ depends on precipitation $Rain(d)$ $(g.cm^{-3})$, irrigation $Irr(d)$ $(g.cm^{-3})$, evaporation $EV(d)$ $(g.cm^{-3})$ and transpiration $TRC1(d)$ $(g.cm^{-3})$.

$$\begin{aligned} ASWC1(d, z) = min\{&MSW(z), ASWC1(d-1, zRoot(d-1)) \\ &+ \frac{(Rain(d) + Irr(d) - TRC1(d) - EV(d)) \times z}{dC1}\} \end{aligned} \tag{3.33}$$

The extra available soil water in layer $C1$, non-zero if the soil capacity $MSW(z)$ is exceeded, is denoted $D1(d)$ and is drained to layer $C2$:

$$D1(d) = \begin{cases} 0 & ASWC1(d) \leq MSWC1 \\ ASWC1(d) - MSWC1 & ASWC1(d) > MSWC1 \end{cases} \tag{3.34}$$

Thus, $ASWC2(d, z)$ depends on $D1(d)$, transpiration $TRC2(t)$ $(g.cm^{-3})$ and available usable water $UWC3(d)$ $(g.cm^{-3})$ from $C3$:

$$\begin{aligned} ASWC2(d, z) = min\{&MSW(z) - MSW(dC1), ASWC2(d-1, zRoot(d-1)) \\ &+ D1(d) - TRC2(d) + UWC3(d)\}) \end{aligned} \tag{3.35}$$

where $UWC3(d)$ represents the influx of water coming from soil layer deeper than the root length, $C3$:

$$UWC3(d) = \frac{ASWC3(d-1) \times dRoot(d)}{dC3} \tag{3.36}$$

Drainage from $C2$ at day d is denoted $D2(d)$:

$$D2(d) = \begin{cases} 0 & ASWC2(d) \leq MSWC2 \\ ASWC2(d) - MSWC2 & ASWC2(d) > MSWC2 \end{cases} \tag{3.37}$$

$D2(t)$ is transferred to the available soil water in $C3$:

$$ASWC3(d,z) = min\{MSW(dC3) - MSW(z), ASWC3(d-1, zRoot(d-1)) + D2(d) - UWC3(d)\} \tag{3.38}$$

5. Daily available water content and fraction of transpirable soil water If $zRoot(d)$ is less than or equal to $dC1$, $ASW(d)$ only accounts for the available soil water content in layer $C1$. Otherwise, it is the sum of available soil water in both layer $C1$ and $C2$:

$$ASW(d) = \begin{cases} ASWC1(d, zRoot(d)) & zRoot(d) \leq dC1 \\ ASWC1(d, dC1) + ASWC2(d, zRoot(d)) & z > dC1 \end{cases} \tag{3.39}$$

Then, $FTSW(d)$ is the ratio between $ASW(d)$ and $MSW(z)$:

$$FTSW(d) = ASW(d)/MSW(zRoot(d)) \tag{3.40}$$

The water stress index $FTSW(d)$ has effects on three processes in this model: leaf expansion $FHLE$, radiation use efficiency $FHRUE$ and plant transpiration $FHTR$. Depending on genotypes and plant functions, critical values RT and RE regulates the plant drought tolerance. While $FTSW(d)$ is less than its respective critical value in any of the three processes, the influential effects of drought to dampen the processes are as in bellowing equations:

$$FHLE(d) = FTSW(d)/RE \tag{3.41}$$

$$FHRUE(d) = FTSW(d)/RT \tag{3.42}$$

$$FHTR(d) = FTSW(d)/RT \tag{3.43}$$

3.2.6 Parameters

The SUNFLO parameters were estimated for 20 genotypes in Lecoeur et al. [2011] using the approach of direct experimental measurements and statistical analysis. Our studies in the following chapters are based on four of these genotypes: "Albena", "Melody", "Heliasol" and "Prodisol". These parameters can be classified into two types: non-genotypic parameters and genotypic parameters. Non-genotypic parameters are parameters that are constant within the species: they take common values for all the genotypes. By contrast, genotypic parameters take different values for each genotype. The class to which each parameter is belonging was determined after analysis based on experimental observations in [Lecoeur et al., 2011]. Their names and units are shown in tables 3.1 and 3.2. Parameter values for genotypes will be given when they are used in the corresponding chapters.

Tab. 3.1: SUNFLO model: non-genotypic parameters

Parameter	Unit	Meaning
$Phy2$	℃ days	Phyllochrone for leaves above rank 6 The parameters determining leaf emergence time.
LAI_a	℃ days	Calculated as the thermal time of the third pairs of leaves' expansion termination
LAI_b	℃ days	Constant for thermal time of leaf maximal expansion rate
LAI_Kei	℃ days	Expansion speed of each leaf ranking above 6
Eb_0 Eb_c Eb_Max Eb_fin	#	The parameter determining the potential radiation use efficiency $Ebp(d)$ for the different phenology stages
$TmaxPHS$ $Topt1PHS$ $Topt2PHS$	#	The parameter determining a thermal factor $FT(d)$, which regulates the potential radiation use efficiency to get the actual one $Eb(d)$

Tab. 3.2: SUNFLO model: genotypic parameters

Parameter	Unit	Meaning
$E1$	℃ days	Thermal time of flower bud appearance
$F1$	℃ days	Thermal time of the beginning of flowering
$M0$	℃ days	Thermal time of the beginning of grain filling
$M3$	℃ days	Thermal time of physiological maturity
pos_SFiMax	#	Rank of the leaf with largest area
$SFiMax$	cm^2	Largest leaf area
NFF	#	Total number of leaves
$coEff$	#	Extinction coefficient
PHS	#	The parameter quantifying the photosynthesis capacity difference between each genotype and Melody
RE	#	Threshold value determining fractional soil water influence on leaf expansion under water stress
RT	#	Threshold value determining fractional soil water's influence on plant transpiration and radiation use efficiency under water stress
HI	#	Proportion of capitulum biomass in total dry biomass

3.3 SUNLAB Model for Sunflower

Note: Most of this chapter content is from Kang et al. [2012a].

A new functional-structural model SUNLAB for the crop sunflower (*Helianthus annuus* L.) was developed in this thesis. It is dedicated to simulate the sunflower organogenesis, morphogenesis, biomass accumulation and biomass partitioning to organs (section 3.3.2). It is adapted to model phenotypic responses of different genotypic variants to diverse environmental factors including temperature stress and water deficit. A sensitivity analysis was conducted to quantify the relative parameter influence on the main trait of interest, the yield (section 3.3.4). The model was calibrated for four genotypes on two experimental datasets collected on plants grown under standard non-limiting conditions and moderate water stress (section 3.3.5). Its predictive ability was then tested on an additional dataset in section 3.3.6. The four considered genotypes - "Albena", "Melody", "Heliasol" and "Prodisol" - are the products of more than 30 years of breeding effort. Comparing the values found for the four parameter

sets associated to each variant allows identifying genotype-specific parameters. Since SUNLAB parameters seem to show genotypic variability, it potentially makes the model an interesting intermediate to discriminate between genotypes. SUNLAB simulates individual leaf area and biomass as two state variables: an interesting corollary is that it also simulates dynamically the specific leaf area (SLA) variable, as shown in section 3.3.7.

3.3.1 Context and Objectives

As one of the major oilseed crops worldwide, sunflower production has to face the growing social demand in a context of strong ecological and economical constraints: growers are confronted to the challenge of increasing sunflower productivity under changing climatic conditions while maintaining low-input levels and reduced costs. A partial response to this challenge could be found by breeding new genotypes and by identifying the best genotype, among a set of existing ones, for a given location and for given management practices; see for instance Allinne et al. [2009]. An emerging approach for the assessment of genotype performances in *in situ* experimental trials is the use of models represented as a set of biophysical functions that determine the plant phenotype in response to environmental inputs. Models can help in breeding strategies and management by dissecting physiological traits into their constitutive components and thus allow shifting from highly integrated traits to more gene-related traits that should reveal more stable under varying environmental conditions [Hammer et al., 2006; Yin et al., 2004]. Consequently, an important question to examine is how to design models that can be used in that context. The models should simulate the phenotypic traits of interest (e.g. yield) with good robustness and predictive capacity. The models should also present a trade-off between mechanistic aspect and complexity. Casadebaig et al. [2011] discuss that question in the case of their model SUNFLO [Lecoeur et al., 2011], that was presented in section 3.2. It has shown good

performances to identify, quantify, and model phenotypic variability of sunflower at the individual level in response to the main abiotic stresses occurring at field level but also in the expression of genotypic variability [Casadebaig et al., 2011]. The authors mixed mechanistic and statistical approaches to deal with highly integrative variables such as harvest index (HI). HI is determined by a simple statistical relationship dependent on covariables previously simulated by the mechanistic part of the crop model throughout the growing season. Although this statistical solution and the large datasets used for its parametrization conferred good robustness to the prediction of HI and thereby crop harvest, feedback effects of biomass partitioning on other processes cannot be taken into account. Moreover, it was shown in Lecoeur et al. [2011] that HI is the parameter that contributes the most to the coefficient of variation of the potential yield (14.3%). It was also shown that when ranking the processes in terms of their impact on yield variability, the first one was biomass allocation (before light interception according to plant architecture, plant phenology and far behind photosynthesis). Therefore, Lecoeur et al. [2011] suggest that a better formalisation of the trophic competition between organs could be a way to improve our understanding of genotypic variation for biomass harvest index. In order to face this challenge, a new sunflower model, named SUNLAB, was derived from SUNFLO. The representation of plant topological development and allocation process at individual organ scale were inspired by the functional-structural plant model GREENLAB, that has been designed as a "source-sink solver" [Christophe et al., 2008] and is accompanied with the appropriate mathematical tools for its identification [Cournède et al., 2011]. SUNLAB thus inherits the flexible rules of sink competition for biomass partitioning at organ scale (blade, petiole, internode and capitulum) from GREENLAB, together with the more detailed representation of ecophysiological processes and environmental stress effects on biomass production and yield from SUNFLO.

This section presents in detail the mechanisms of SUNLAB and parameter estimation

procedure based on field experimental data. A sensitivity analysis is performed on the model parameters, using the Sobol method, to investigate the relative contribution of each parameter and their interactions to the model output uncertainty. The output that we consider is the main trait of interest in most breeding procedures, that is the final yield. The potentials of SUNLAB for genotypic characterization are illustrated by comparing the parameters obtained after the estimation process for four genotypes, namely "Albena", "Heliasol", "Melody" and "Prodisol". The performances of SUNLAB to reproduce phenotypic variability coming either from genotypic or from environmental influences are tested against experimental datasets used for calibration. An additional dataset is then used for model validation. An interesting and uncommon output of SUNLAB is the specific leaf area (SLA, $cm^2.g^{-1}$), *i.e.* the ratio of leaf area to dry leaf mass, which is usually an influential input variable often associated with large uncertainty ranges in most dynamic crop growth models [Rawson et al., 1987]. We finally discuss the potential benefits of integrating two modelling approaches: that of SUNFLO, an ecophysiological model whose parameters can be assessed by direct field measurements, and that of GREENLAB, a mechanistic dynamic model whose parameters are estimated by optimization methods from experimental data. After further tests and improvements, this new SUNLAB model should present robust enough predictive capacities and ability to differentiate between genotypes in order to be proposed as a proper tool for the understanding of gene $\times$ environment interactions.

3.3.2 Modeling: SUNLAB Modules

SUNLAB consists of five modules: phenology, water budget, organogenesis and morphogenesis, biomass accumulation, and biomass partitioning. Phenology, water budget, and biomass accumulation modules are directly inherited from the SUNFLO model. The organogenesis and morphogenesis module is modified from the corresponding SUNFLO module by defining for each organ the dates, expressed in thermal

time, of initialization and termination of its growth. The biomass partition module is an entirely new module. We describe here equations of these modules, briefly for those inherited from SUNFLO - we refer to section 3.2 for an exhaustive description - and in detail for the new contributions. Model parameters, which are mentioned in the following equations, will be listed in section 3.3.3.

Organogenesis and Morphegenesis Module

From the emergence and senescence blades numbers obtained from SUNFLO module functions, the thermal times of initiation $bladeInitTT(i)$ and senescence $bladeSeneTT(i)$ of each blade of rank i can be computed:

$$\begin{aligned} bladeInitTT(i) &= (i-1)/R \\ bladeSeneTT(i) &= M3 - \frac{i \times (M3 - M1)}{Ntotal} \end{aligned} \tag{3.44}$$

The petiole i and the internode i from the same metamer of blade i have the same value of initiation thermal time. While petiole i has the same value of senescence time as $bladeSeneTT(i)$, senescence thermal time of internode i is the same as the accumulative thermal time in the end of the plant life. Capitulum initialization thermal time equates $M0$ and it grows until the end. With all the information of initialization thermal time and senescence thermal time of every organ, a general sunflower structure can be constructed. For every organ, besides their appearance and senescence thermal time, their expansion thermal time are also calculated, explained in section 3.3.3: parameter analysis.

Biomass Distribution Module

As in GREENLAB, the biomass produced by leaves is distributed to all organs proportionally. The mechanism is to describe the total above-ground biomass $CDM(d)$

as a biomass common pool, which is the total biomass of all blades, petioles, internodes and the capitulum. Blades are "sources" to add the pool's biomass. Blades, petioles, internodes, and capitulum are "sinks" to partition biomass of the pool. The calculation of each organ's biomass on day d is done through three steps.

1. Sink competition degree Sink ability $SAP(d)$ (equation (3.45)) represents each organ's potential sink competition ability on day d.

$$SAP(d,t,i) = \frac{(\frac{CTT(d)-initTT(t,i)}{epdTT(t,i)})^{sinkA-1} \times (1-\frac{CTT(d)-initTT(t,i)}{epdTT(t,i)})^{sinkB-1}}{(\frac{sinkA-1}{sinkA+sinkB-2})^{sinkA-1} \times (1-\frac{sinkA-1}{sinkA+sinkB-2})^{sinkB-1}} \tag{3.45}$$

It varies with different organ type t (blade, petiole, internode or capitulum) and organ rank i (blade ranking i in the blade organ type for example). This function is simulated by the density function of beta distribution. Two organ type specific parameters $sinkA$ and $sinkB$ take charge of the curve shape, as illustrated in the result section Fig. 3.7. Organ rank affects the function by two variables: $initTT(t,i)$ and $epdTT(t,i)$. For each individual organ, the duration of sink activity is equal to the organ's expansion duration $epdTT(t,i)$ (°C $days$), started from its initialization thermal time $initTT(t,i)$ (°C $days$). For example, the blade ranking i germinates around 400 °C $days$ earlier than blade ranking $i+1$. Therefore blade $i+1$ has $initTT(t,i)$ 400 bigger than blade i. The detailed calculations are elaborated in parameter analysis section 3.3.3 because they are related to our strategy to determine SUNLAB parameters. $SAP(d,t,i)$ changes according to time and its value ranges from 0 to 1.

The individual organ's sink competition degree $SA(d,t,i)$ (equation (3.46)) is the organ i's actual sink demand at time d, calculated by multiplying its sink ability $SAP(d,t,i)$ and an organ type specific parameter: sink ratio parameter SR. Organ type "capitulum" has normally hundreds of times bigger sink ratio SR than organ type "blade".

$$SA(d,t,i) = \begin{cases} SAP(d,t,i) \times SR & initTT \leq CTT(d) \leq initTT + epdTT \\ 0 & otherwise \end{cases} \tag{3.46}$$

2. Total sink demand The plant's total sink demand $sumSink(d)$ is computed as the scalar product of the number of appeared organs to their organ sink demand $SA(d,t,i)$:

$$sumSink(d) = \sum_t \sum_i SA(d,t,i) \tag{3.47}$$

3. Individual organ's biomass distribution Total dry biomass $CDM(d)$ allocated to a single organ is calculated as the proportion of the organ's sink demand $SA(d,t,i)$ to total sink demand $sumSink(d)$. For example the biomass allocated to individual blade $indBladeMS(d,i)$ $(g.m^{-2})$ of blade ranking i is:

$$indBladeMS(d,i) = \frac{CDM(d) \times SA_{blade}(d,i)}{sumSink(d)} \tag{3.48}$$

Total blade biomass $bladeMS(d)$ $(g.m^{-2})$ is the sum of all individual blade biomass:

$$bladeMS(d) = \sum_i indBladeMS(d,i) \tag{3.49}$$

In total, SUNLAB simulates the individual blade biomass $indBladeMS(d,i)$ and total blade biomass $bladeMS(d)$, individual and total petiole biomass ($petioleMS(d)$, $g.m^{-2}$), individual and total internode biomass ($internodeMS(d)$, $g.m^{-2}$), and capitulum biomass($capMS(d)$, $g.m^{-2}$).

3.3.3 Related Datasets and Parameter analysis

Experiments and measurements for designing and constructing modules and parameters which are directly inherited from SUNFLO are not presented here, as they are described in detail in Lecoeur et al. [2011]. Data used for SUNLAB parameters estimation, simulation and application include three datasets "2001", "2002a"and "2002b" (2.1.1)."2001" and "2002a" as two datasets in discriminated environment are used to calibrate SUNLAB model and "2002b" is used for model validation.

Four genotypes "Albena", "Melody", "Heliasol" and "Prodisol" are considered in this project. These genotypes have been characterized by a large study of genetic improvement of sunflower over the last 30 years, and they are four of those most widely grown varieties in France. SUNLAB parameters can be decomposed in two subsets. One subset contains the parameters inherited from SUNFLO which keep the same values in SUNLAB (Table 3.3).

Tab. 3.3: Main SUNFLO inherited parameters values.

Parameter Name	Parameter values			
	Albena	Melody	Heliasol	Prodisol
$E1$ (°Cd)	510	540	480	510
$F1$ (°Cd)	900	920	880	900
$M0$ (°Cd)	1160	1160	1150	1120
$M3$ (°Cd)	1800	2060	1940	1840
NFF (#)	31	26	24	25
$position\_SFiMax$(#)	18.9	15.4	15.3	15.9
$SFiMax$ (cm^2)	488	613	670	498
$coEff$ (#)	0.78	0.96	0.88	0.87

The other subset contains 17 additional parameters of SUNLAB that needs to be estimated from experimental datasets. They include 12 parameters that drive the sink competition (SR, $sinkA$, $sinkB$ for four types of organs) and 5 parameters, which are used to adjust or define initial and final organ expansion thermal times: $initTTAdjust$ (°C $days$), $epdTTA$ (°C $days$), $epdTTB$ (°C $days$), $internodeEpdTT$ (°C $days$), and $capitulumEpdTT$ (°C $days$). Thermal time of blade growth initialization is calculated

by subtracting $initTTAdjust$ from $bladeInitTT(i)$ (°C $days$) which is the thermal time of blade emergence. The adjustment parameter $initTTAdjust$ is added to the model because according to the experimental criterion: leaves are recorded when lengths of their central vein are bigger than 4cm [Lecoeur et al., 2011], $bladeInitTT(i)$ is the thermal time when the leaf size could be measured, but at then this leaf has already received a small amount of biomass. The thermal time of blade expansion end is also measured. The thermal times of blade initialization and end of expansion can vary with their ranks: the variation is linear and depends on two parameters, $epdTTA$ and $epdTTB$. For example, the expansion duration of blade at rank i, expressed in thermal time, is:

$$\begin{aligned} bladeEpdTT(i) = \quad & bladeSeneTT(i) - (epdTTB - epdTTA \times i) \\ & -(bladeInitTT(i) - initTTAdjust) \end{aligned} \tag{3.50}$$

where bladeSeneTT (°C $days$) is the thermal time of leaf beginning of senescence. Petioles share the same initial, expansion and end biomass thermal time as the blades in the same metamers. Internodes have the same initial biomass thermal time as blades of the same metamers, but they have parameter $internodeEpdTT$ to define their expansion duration. Capitulum begins its sink competition at plant age $M0$, and expands in the thermal time $capitulumEpdTT$. Regarding the target data for parameter estimation, only blade areas were measured at organ scale. All other organs were only weighted at compartment scale. In particular, independent blade mass data was not available, while these data are required for a better estimation of SUNLAB parameters. Therefore, profiles of individual blade mass were estimated as follows: at each date where total blade mass and total blade areas were measured at compartment level, a virtual SLA value was computed and was used to generate a set of individual blade mass. The model can thus be viewed as a dynamic interpolation solver that generates both blade areas and mass between those fixed measurement dates. This

will be detailed in the SLA study in section 3.3.7.

A sensitivity analysis is performed on SUNLAB parameters to understand their relative influence on determining the main model output, the yield Y. A global method was used, the Sobol method [Saltelli et al., 2000; WU and Cournède, 2010]. In this method, parameters are considered as random variables that are drawn from predefined distributions, chosen here as uniform distributions since no *a priori* information was available for the SUNLAB parameters. This allows computing an estimator of the output variance, $V(Y)$. The first-order sensitivity index of a given parameter X_i can thus be defined as:

$$S_i = \frac{V_{X_i}(E_{\sim X_i}(Y|X_i)}{V(Y)} \tag{3.51}$$

where the inner expectation operator is the mean of Y taken over the possible values of all other parameters except $X_i\,(\sim X_i)$ while keeping X_i fixed. Then outer variance is taken over all possible values of X_i. Similarly, higher order sensitivity indices can be defined to characterize the effects of interactions between parameters on the output variance. Sensitivity indices are normalized thanks to the well-known formula of variance decomposition. The non-linear generalized least squares method with Gauss Newton method for optimization [Cournède et al., 2011] is used for estimating the parameters using field data including total blade biomass, total petiole biomass, total internodes biomass, capitulum biomass and individual blade biomass. The simulations, sensitivity analysis and estimation procedure were performed on a plant modeling assistant platform, named PYGMALION, developed in Digiplante team in Ecole Centrale Paris, France.

3.3.4 Result: Sensitivity Analysis

A sensitivity analysis was performed on the 15 parameters of SUNLAB for the yield, using the Sobol method of variance decomposition. Results are gathered in Table

3.4 for the most influential parameters. The sum of all first order indices was 0.87, which means that the part of variance due to parameter interactions is less than 15%: this justifies that the sensitivity analysis of this model can be grounded on first-order indices of parameters. The most influential parameters are those driving the dynamics of capitulum sink variations, *sinkAcap* and *sinkBcap*, accounting for 51% and 12% respectively of the yield variance. The only other parameter with significant sensibility index is a parameter of internode sink variation, *sinkAintern*. All other parameters account for less than 5% of the yield variance. This result suggests that dynamics of biomass allocation to the capitulum, more than the value of its sink, are important for yield determination.

Tab. 3.4: Sensitivity analysis of SUNLAB parameters: first-order indices of the most influential parameters (with index > 1%).

sinkAcap	sinkBcap	sinkAintern	SRcapitulum	SRintern	sinkBintern	internEpdTT
0.51	0.12	0.12	0.05	0.03	0.02	0.02

3.3.5 Result: Model Calibration on Four Genotypes and Two Environmental Conditions

Parameter Estimation for the Four Genotypes

The SUNLAB parameters were estimated for the four different genotypes ("Albena", "Melody","Heliasol", and "Prodisol") using experimental datasets of "2001" (non-limiting conditions) and "2002a" (with water deficit). Their values are shown in Table 3.5 with the associated standard deviation. Since the sink competition model is chosen to be proportional [Heuvelink, 1996], *i.e.* all the daily produced biomass is allocated and there are no reserves, a reference sink value has to be set: conventionally, the sink of blades *SRblade* is set to 1.

Parameter values are independently estimated for each genotype, *i.e.* no *a priori* geno-

typic correlations are imposed. This allows comparing the genotypes according to their parameter values. The standard error could allow testing the significance of differences between two parameter values, but this would only be an approximate result since the number of observations that directly influence the estimation of each parameter is unknown. Qualitative observations can nevertheless be done. For example, blade parameter *sinkAblade* in the sink variation function of blades appears significantly different between four genotypes, while no clear evidence of genotypic variability was found for capitulum sink ratio *SRcapitulum* (see also Fig. 3.7). The internode sink ratio, *SRinternode*, is found different for genotypes "Albena" and "Melody", but takes similar values for "Heliasol" and "Prodisol".

Tab. 3.5: Estimated parameter values of SUNLAB for four genotypes.

Parameter Name	Param. values (with associated standard error)			
	Albena	Melody	Heliasol	Prodisol
sinkAblade	8.4 (0.22)	2.8 (0.12)	2 (0.1)	4 (0.16)
sinkApetiole	3.4 (0.33)	1.5 (0.22)	1.5 (0.7)	4.3 (0.76)
sinkAintern	2.2 (0.12)	3.5 (0.05)	2.2 (0.07)	3.8 (0.08)
sinkAcap	5.6 (0.12)	4.3 (0.17)	6.5 (0.3)	6.5 (0.28)
sinkBblade	14.8 (0.4)	2.3 (0.16)	2.1 (0.18)	3.6 (0.26)
sinkBpetiole	16.8 (1.8)	4.1 (6.4)	2.7(0.76)	4.2 (0.5)
sinkBintern	13.8 (3.9)	7.7 (0.29)	1.7(0.07)	12.2 (0.44)
sinkBcap	3.4 (0.22)	2.5 (0.23)	6.1(0.44)	5.8(0.52)
SRpetiole	0.5 (0.04)	0.2 (0.03)	0.24(0.03)	0.43 (0.04)
SRintern	1 (0.06)	3 (0.19)	1.6(0.08)	1.8(0.09)
SRcap	1000 (253)	600 (126)	350(54)	500(144)

Model Performances: Reproducing Genotype-induced Variability

Even when grown under non-limiting controlled conditions, the four studied varieties present some phenotypic variability, that might be intrinsically regulated by genotypic influences. This phenotypic variability is in particular observed on daily radiation interception efficiency $RIE(d)$, total blade area $AA(d)$, leaf number $N(d)$, accumulated dry biomass $CDM(d)$ and biomass partitioning. This is illustrated in Fig. 3.6 for dry

mass compartments (blade, internode and capitulum) with the "2001" experimental dataset. This figure also illustrates the model ability to reproduce this (presumably) genotypic variability.

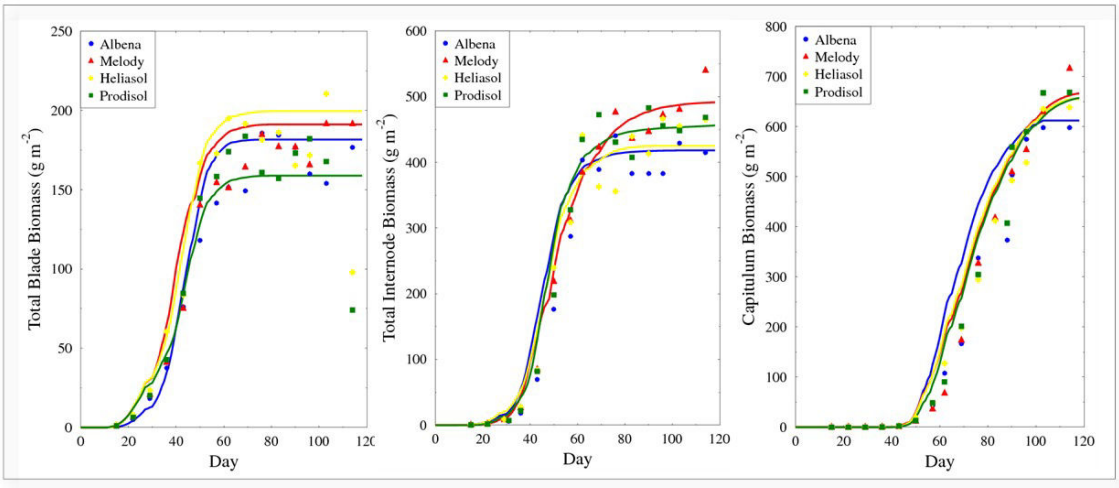

Fig. 3.6: Experimental data (dots) and simulation (lines) comparisons of blade dry mass, internode dry mass, and capitulum dry mass for the four genotypes - "Albena", "Melody", "Heliasol", and "Prodisol" - and for dataset "2001"(blue)

The estimated parameter values (Table 3.5) allow tracking back the dynamics of biomass allocation and analysing the internal mechanisms underlying sink competition. For instance, compared to "Prodisol", blades of "Albena" enter earlier in the competition for biomass but the capitulum reaches its maximum demand later (Fig. 3.7): this may explain that in the end "Albena" has bigger total blade biomass but smaller capitulum biomass than "Prodisol"(Fig. 3.6). Genotype performance can also come from the biomass accumulation module: "Melody" has larger internode and capitulum biomass than "Heliasol", and they have similar blade biomass, as can be seen in Fig. 3.6. This is due to a higher radiation use efficiency of the "Melody" genotype.

Model Performances: Reproducing Environment-induced Variability

The SUNLAB model was calibrated using "2001" and "2002a" experimental datasets that included data for plants grown under water deficit. The calibrated SUNLAB

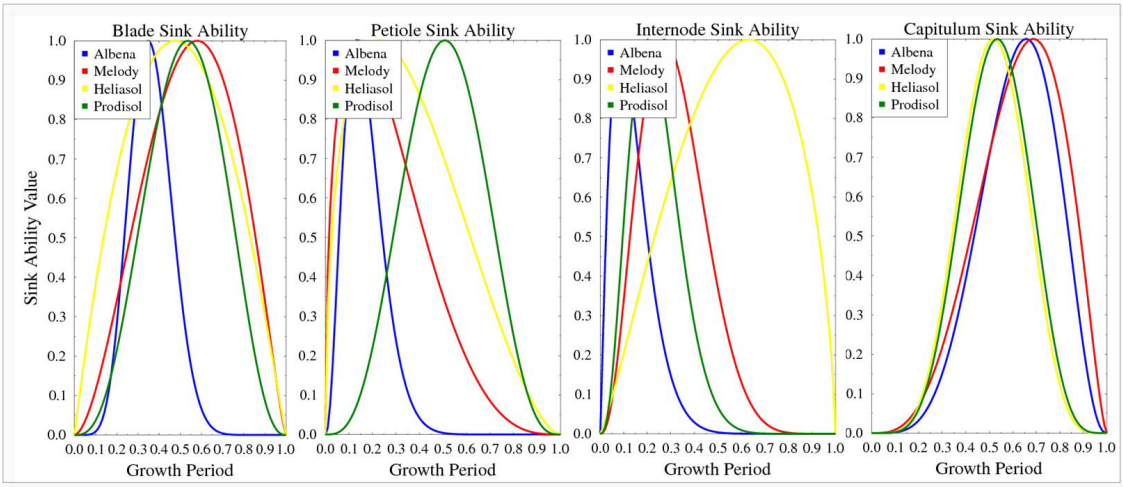

Fig. 3.7: Sink ability based on SUNLAB parameters and sink competition theory in biomass distribution module

model ably simulate the phenotypic variability induced by the two contrasted environmental conditions of "2001" and "2002a" datasets. This is illustrated in Fig.3.8 that shows experimental data and simulations of radiation interception efficiency $RIE(d)$, total blade area $AA(d)$, leaf number $N(d)$, accumulated above-ground dry biomass $CDM(d)$ and biomass compartments (capitulum, blades, petioles, internodes) for the "Melody" genotype. It can be noticed that "Melody" is not very sensitive to water stress since the dry mass accumulation does not significantly vary. The last two graphs of this figure present some details on two other genotypes: biomass compartments of "Prodisol" and individual blade mass profile for "Heliasol". Water stress induces a decrease in the capitulum biomass of "Prodisol" plants, despite a slight increase in blade biomass. The effect of water stress can also be observed on the individual blade mass profile of "Heliasol" plants: blades on the last ranks grow less in water deficit conditions ("2002a") than in standard conditions ("2001").

3.3.6 Result: Model Validation

In order to test the model predictive ability, it was confronted to an additional experimental dataset "2002b", that was not used for the calibration step. Fig. 3.9 presents

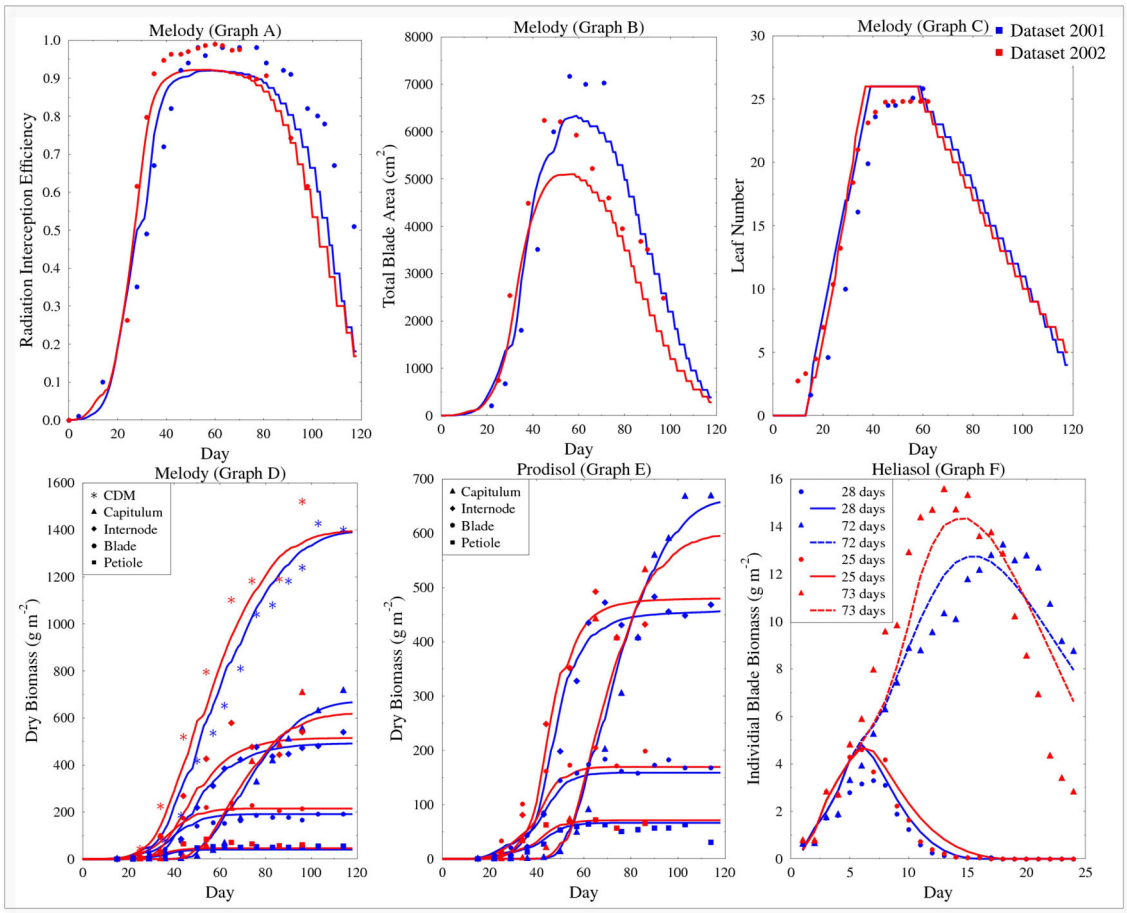

Fig. 3.8: Graphs A to D: Experimental data (dots) and simulation (lines) comparisons for the "2001" (blue) and "2002a" (red) conditions of the radiation interception efficiency $RIE(d)$, total blade area $AA(d)$, leaf number $N(d)$, accumulated aboveground dry biomass $CDM(d)$ and biomass compartments (capitulum, blades, petioles, internodes) for the "Melody" genotype. Graphs E and F: biomass compartments of "Prodisol" and individual leaf mass profile for "Heliasol".

some phenotypic traits for the "Albena" genotype: for total blade areas and radiation interception efficiency, data are underestimated by model predictions, but the results are reasonable for the biomass compartment dynamics. It has to be noted that this validation process is still a preliminary step since our additional experimental dataset was measured on plants growth in conditions similar to those of the "2002a" dataset which was used to calibrate the model.

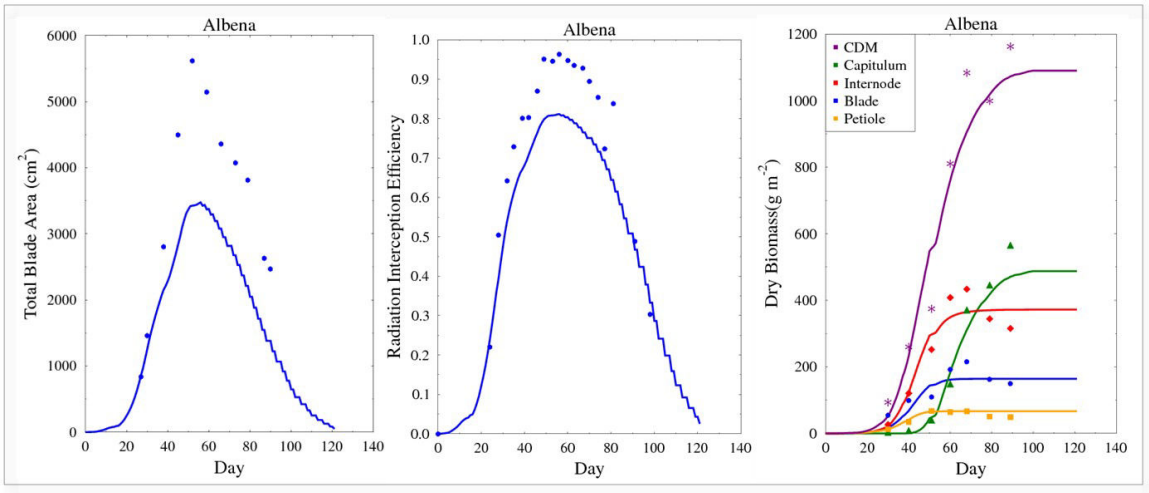

Fig. 3.9: Model validation for genotype "Albena" using an additional experimental dataset: "2002b"

3.3.7 Result: Model Application: an Exploratory Study on Specific Leaf Area

Specific leaf area (SLA) is an important variable in plant growth modeling. In most dynamic models, it is usually used to determine blade surface area values from blade biomass, as in GREENLAB [Christophe et al., 2008] or in TOMSIM [Heuvelink, 1999]. Since blade area in turn determines the biomass production, accurate estimation of SLA is mentioned as a major source of error in models and implies difficulties in obtaining a reliable computation of leaf area index, which is the main component of biomass production modules [Heuvelink, 1999; Marcelis et al., 1998]. It is however generally considered as constant, although it has been shown, for instance on wheat [Rawson et al., 1987], that SLA varies according to genotypes, leaf ranks and leaf growing periods. Regarding sunflower, the variations of SLA and the factors influencing them are still poorly known. As SUNLAB can simulate dynamics of individual blade mass profiles independently from those of blade areas, the SLA can be computed as a model output, contrary to the classical situation where it is taken as input.

In Fig.3.10, the simulated and observed values for individual blade areas and masses of "Melody" in the "2001" dataset are displayed for each blade rank and six different growth stages. The SLA was computed at the time when individual blades have

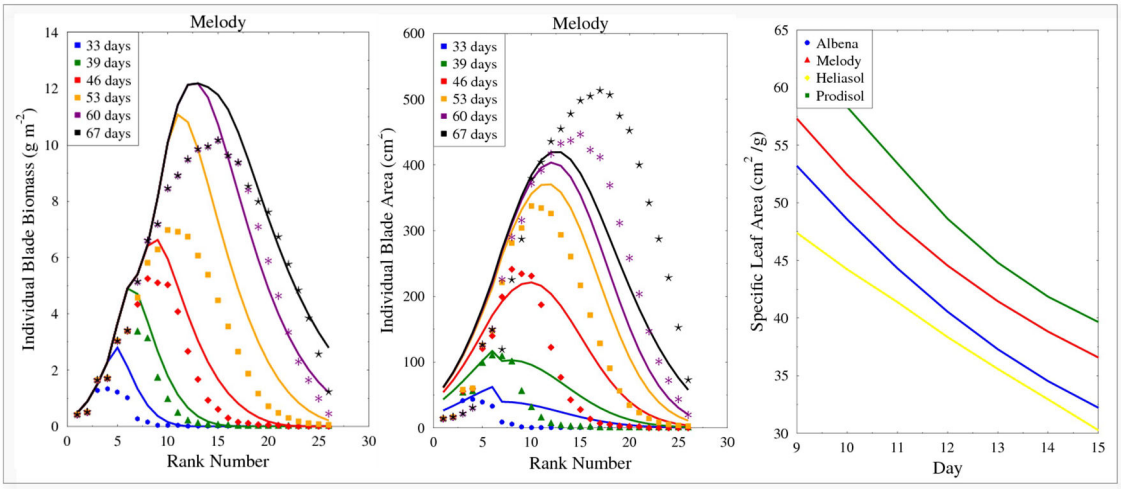

Fig. 3.10: Comparison of simulation and field data for individual blade area and biomass of genotype "Melody"; the right graph is the simulation of specific leaf area for the four genotypes

reached their higher mass. It was done only for blades ranking from 9 to 15 which are those showing the best accordance to the field data (Fig. 3.10). The computed SLA shows some variability among the four genotypes. Since the current SUNLAB parameters come from the reconstructed individual blade masses, the simulated SLA results will need to be improved with better experimental data in the future for more accuracy on this result.

3.3.8 Discussion

A functional-structural model SUNLAB was developed. It describes the sunflower topology and morphogenesis at organ level with blades, petioles, internodes, and capitulum. Coordination of the expansion dynamics of these organs are ruled by their initiation and senescence time, expressed with respect to thermal time. Ecophysiological processes work together with plant structural dynamics to affect biomass

accumulation and partitioning to organs. The model was applied on data of four genotypes “Albena”, “Melody”, “Heliasol” and “Prodisol” to evaluate the ability of this newly-developed model to reproduce observed data of sunflower growth.

As a joint concept between SUNFLO and GREENLAB, SUNLAB has better structural features than SUNFLO and it succeeds to deal with the biomass distribution at organ level. Compared to GREENLAB, SUNLAB inherits the ecophysiological functions of SUNFLO that have been validated in different environmental conditions for 26 genotypes [Casadebaig et al., 2011; Lecoeur et al., 2011] and possesses SUNFLO’s following merits. Firstly, SUNFLO contains more genotype-specific parameters. It could predict well large phenotypic variability of complex genotypic traits. These genotypic traits, represented as genotypic parameters in the model, have enough genotypic variability to discriminate between genotypes. In the construction process of SUNFLO, the authors used the approach of linking a complex phenotype to a set of accessible genotypic traits. Each genotype is defined by chosen traits which were transcribed into a set of genotype-specific parameters. These genotypic parameters are thus under certain genetic control. With the reason of improving the model parameters update ability for yearly cultivar releases, parameters number is limited while a useful predictive capacity is maintained. Meanwhile, as most SUNFLO parameters could be estimated by direct measures, it allows parameter values to be more representative of crop physiology than those that are estimated indirectly with optimization algorithms. Secondly, SUNFLO and SUNLAB have better ecophysiological functions. GREENLAB over-simplifies a number of processes, such as photosynthesis and assimilate conversion to biomass [Guo et al., 2006; Ma et al., 2008], and it is still in its preliminary stage to include water source influence and root system [Li et al., 2009]. In SUNFLO and SUNLAB, the radiation use efficiency is taken into account for photosynthesis. Many environmental stresses to phenotypic plasticity are considered, such as temperature and water. The included root sub-model induces water stress, which

affects crop processes such as leaf expansion, plant transpiration, and biomass production. This consideration enriches environment discrimination by taking into account the effect of soil texture, apparent soil density or stone content.

Modeling crop growth and breeding through empirical experimental analysis and the parametrization from direct parameter measurements, such as SUNFLO model, bears clear advantages in terms of ecophysiological relevance and parameter accuracy. The genotypic variability may also be easier to characterize by considering directly elementary ecophysiological processes. This perspective has led to automated and high-throughput advanced plant phenotyping (see for example Granier et al. [2005], Sotirios A and Christos [2009]). However, direct and accurate measures on elementary processes do not necessarily imply that the combination of these processes will provide the same accuracy at plant scale. The nonlinear interaction between processes as well as the necessary simplifications in terms of the number of ecophysiological processes considered in the model make the whole plant model not a simple combination of the elementary models well calibrated by experiments, plants are complex systems whose description of elementary process interactions, plasticity and robustness remains an open issue [Yin and Struik, 2010]. Therefore, parametrization methods relying on model inversion to estimate parameters from experimental data [Cournède et al., 2011; Guo et al., 2006], at whole plant level offers an interesting alternative. The parameters thus obtained are less ecophysiologically relevant and contain a part of empiricism, but are more representative from the point of view of the plant global behavior, and it may still be possible to use these parameters to differentiate between genotypes [Letort, 2008]. Moreover, some processes like biomass allocation at organ level can be difficult to observe experimentally and using inverse methods for parameter estimation may be necessary. While it is hard to find a balance for a model design, SUNLAB model is an interesting trial.

SUNLAB benefits from both strategies: direct measurements of ecophysiological pa-

rameters when it is possible, and parameter estimation by inverse methods for others, and preserves a good capacity for genotypic differentiation. SUNLAB has good calibration results of discriminating different genotypic and environment scenarios to simulate multiple phenotypic traits. The genotype "Melody"and "Heliasol" (Fig. 3.6) were shown to have better drought tolerant ability than the other two genotypes. They had almost no influence on their yields while the other two had slight reduction (around 15% of 2001 harvest). While SUNLAB well simulated genotypic variance, the next step is to investigate the genetic determinism on the model's genotype specific parameters which account for the feature as illustrated for example in [Buck-Sorlin et al., 2005]. SUNLAB is designed to simulate drought stress on the crop sunflower. In this project, two environment scenarios were used to calibrate the model. 2002a has a stronger water deficit than 2001 (Fig. 2.1), particularly after beginning of grain filling $M0$. With some variation according to plant species, certain stages such as germination, seedling or flowering could be the most critical stages vulnerable to water stress[Hadi et al., 2012]. Seed germination is the first critical stage and the most sensitive in the life cycle of plants [Ahmad et al., 2009] and seeds exposed to unfavorable environmental conditions, such as water stress at this stage may have seedling establishment compromised [Albuquerque and Carvalho, 2003]. However our simulation and field data suggested that the drought stress on crops was very small. It is possibly because since sunflower is categorized as a low to medium drought sensitive crop [Turhan and Baser, 2004], the water deficit level is not strong enough to result in severe influences. An environmental scenario with stronger water deficiency is required to explore the model's simulation and predictive capacity.

Finally, in functional structural models, the mechanistic description of ecophysiological processes is a key step to improve their predictive capacities and their ability to differentiate between genotypes, making them proper candidates for the understanding of gene $\times$ environment interactions (see some efforts in this direction in

[Allen et al., 2005; Bertheloot et al., 2011; Minchin and Lacointe, 2005]). However, the parametrization effort of these more and more complex models should always be taken into account when improving their mechanistic description, to prevent from a high level of uncertainty in the parameters which may hinder the original purposes of the model in terms of prediction and genotypic differentiation.

Conclusion. This new model provides a novel way of investigating genotype performances under different environmental conditions. These promising results are a first step towards the potential use of the model as a support tool to design sunflower ideotypes adapted to the current worldwide ecological and economical challenges and to assist the breeding procedure.

Part II

ANALYSIS

4. MULTI-SCENARIO METHODOLOGY OF PARAMETER ESTIMATION (MSPE)

In this Chapter, we propose an innovative parameter estimation methodology adapted to breeding programs: the Multi-scenario Methodology of Parameter Estimation (MSPE). The methodology takes advantage of the multi-scenario trials (potentially large amounts of environmental conditions available, but with the availability of only a small quantity of experimental plant traits information for each scenario) set in place by breeders to evaluate the performances of their genotypes. Four research questions are investigated: the feasibility of MSPE (mostly identifiability), the practical implementation of the method (with sensitivity analysis and numerical optimization issues), the effect of the number of scenarios on the estimation error and on model prediction ability.

4.1 MSPE Context and Objectives

The role of ecophysiological models of plant growth to analyze genotype-by-environment interactions is now well acknowledged [Hammer et al., 2006; Yin et al., 2004]. The identification of QTL for model input traits opens new perspectives for breeding [Letort et al., 2008; Yin and Struik, 2010]. Ideally, ecophysiological models involve biophysical parameters that are stable for a given genotype in a range of environmental conditions: "one parameter set, one genotype" [Tardieu, 2003]. A crucial issue is, however, the

estimation of these parameters. In order to allow the discrimination between genotypes, model parameter estimation should be accurate enough; this is usually ensured through heavy experimental work to measure trait data [Reymond et al., 2003]. For example, for SUNFLO model, detailed leaf surface area every day in sunflower growth periods are needed to estimate the parameters of the architecture model [Lecoeur et al., 2011]. Besides, to estimate all significant parameters, other heavy traits data are needed, such as radiation absorption efficiency and total dry biomass day-by-day.

Such type of experiments is difficult to implement in breeding programs, for which a lot of genotypes and large ranges of environmental conditions are considered [Jeuffroy et al., 2006b]. Reducing the amount of data collection to save experimental cost normally results in sacrificing parameters' accuracy. In reality, in the field of experimental agronomy or breeding, there are plant agronomic data in many different environmental conditions (farmers statistics for example), but for each of them, only a small quantity of experimental plant traits information are available. We propose and test an original strategy built to take advantage of the mathematical formulation of plant growth models as dynamical systems and of the multi-environmental trials (potentially large amounts of environmental conditions available) set in place by breeders to evaluate the performances of their genotypes. The Multi-Scenario methodology of Parameter Estimation (MSPE) is designed to take large scenarios' simple data to estimate parameters, instead of using large collection of detail plant growth data, as shown in Fig. 4.1. More generally, this methodology aims at proposing an alternative solution for the strategy of crop model parameterization, when it is hampered by the heavy cost of experiments and data collection. Four research questions are investigated.

1. The feasibility of MSPE methodology. The hypothesis is that such variety of discriminating scenarios should compensate for the little amount of information (data) for each scenario. Therefore the collection of those limited data in many environmental scenarios is sufficient to estimate model parameters accurately. Our first objective

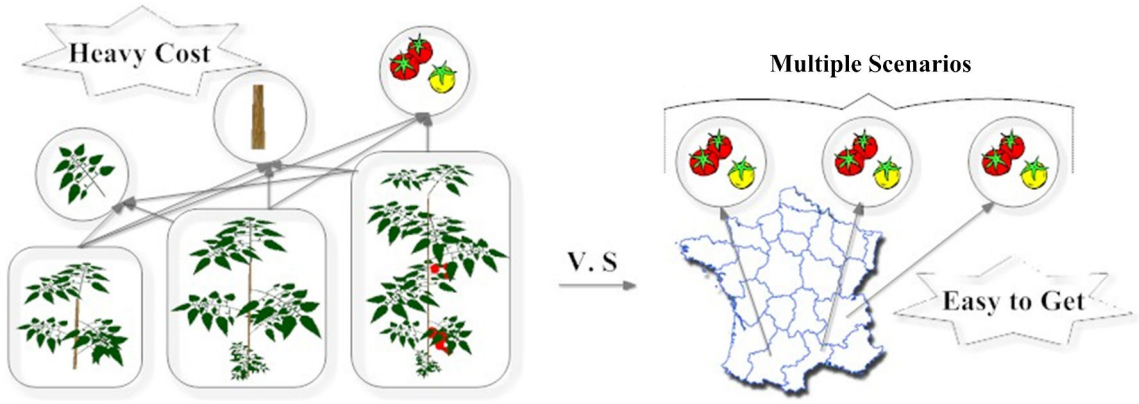

Fig. 4.1: Illustration of the idea of MSPE methodology

is to validate this hypothesis and thus check the system identifiability, both structural (which model parameters can potentially be estimated), and practical (is the level of statistical information available from experimental data sufficient for model inversion). For this purpose, two kinds of tests, "proof of concept" test and statistical property test, are conducted for virtual data with the SUNFLO model. The calibration accuracy with this method is also investigated.

2. Implementation and optimization issues in MSPE. Even if MSPE has been proved to be feasible in the first research question, the practical usage of MSPE raises several issues. It is essential to know what significant parameters should be chosen for MSPE. For this purpose, sensitivity analysis is used. Moreover, the optimization problem arising from the maximization of likelihood (or minimization of generalized least squares) is not a simple one that can be solved in the frame of convex optimization. We also study the practical identifiability from a concrete point of view by investigating how many scenarios are sufficient for model inversion. Several optimization algorithms are tested and discussed. Observation data sufficiency is assumed to be related to the number of scenarios. A presumption is that the insufficiency (and thus non-identifiability) can be weakened by increasing the number of scenarios.

3. Evaluation of estimation accuracy. The effect of the number of scenarios on estimation accuracy is investigated by the Jackknife method: we analyze parameter

distributions, particularly their means and variances, with respect to the number of scenarios. From an applicative point of view, this may help by fixing the number of experimental situations necessary for parameter estimation for a desired level of accuracy.

4. Evaluation of prediction error. It is assumed that the increase in the number of scenarios in MSPE should result in a better prediction ability of the calibrated parameter values. A rough illustration based on the test of the number of scenarios increase and a strict proof based on cross-validation will be presented.

4.2 The Feasibility of MSPE Methodology

4.2.1 Methodology

To validate the hypothesis, and to test the calibration accuracy of the method, two kinds of tests, "proof of concept" test and statistical property test are conducted, first based on virtual data.

SUNFLO model, as detailed in Chapter 3, consists of five interacting sub-models: Phenology, Architecture, Biomass Production, Biomass Allocation, and Water Budget. From environmental inputs, the model predicts the sunflower yield. There are around 30 parameters and 50 variables in the model. Ten parameters are shown to be genotype-dependent in table 3.2, while the other parameters are supposed genotype-independent and are therefore supposed constant for all genotypes. Some of these parameters are relatively easy to measure, like the maximal number of leaves of a sunflower for one genotype. On the other hand, some of them are difficult to get because their computation relies on very heavy collection of field data. SUNFLO model is used for genotype characterization in this study. Cournède et al. [2011] described a formalization of the system observation vector adapted to the irregular and

composite observations often characterizing experiments on living systems, particularly on plants. Such formalization was implemented in the PYGMALION software, developed at Ecole Centrale Paris, which offers a framework for the implementation of dynamic systems of plant growth and their mathematical and statistical analysis, including the model inversion allowing the consideration of very heterogeneous types of experimental data. Thanks to this platform, a 2-stage Aitken estimator [Taylor, 1977] adapted to this composite data is used for parameter estimation in these tests. Datasets in 2.3.1 and 2.1.2 are used as experimental scenarios, and the observation function of the model is composed of all the experimental data resulting from a family of scenarios. The square error of this observation function conditional to the parameter vector is then minimized with respect to the parameter vector via a Gauss-Newton descent method [Walter and Pronzato, 2006].

Proof of Concept Test

The "proof of concept" test is adopted to demonstrate the principle of the method. Its general strategy is (1) to generate a virtual experimental data set with SUNFLO from a given set of parameters P_1 and a family of experimental scenarios, and (2) test given an initial parameter set P_2, whether the estimation algorithm manages to retrieve the original vector P_1. The strategy is described in Fig. 4.2.

For every experimental scenario, M trait data $(Trait_1, \cdots, Trait_M)$ are generated with P_1. $(Obs_1, \cdots, Obs_N)$ describe the observations of these M traits for all N environmental scenarios. P_2 is then chosen as the initial vector of the estimation algorithm and the iterative process for parametric estimation from $(Obs_1, \cdots, Obs_N)$ is run to get a vector of parameters, P_3. If P_3 is equal to P_1 it means that the experimental data set composed of N scenarios for the M traits $(Trait_1, \cdots, Trait_M)$ available for each scenario is theoretically sufficient (discriminative enough) to estimate parameters for the SUNFLO model. This test was used to explore the effects of the

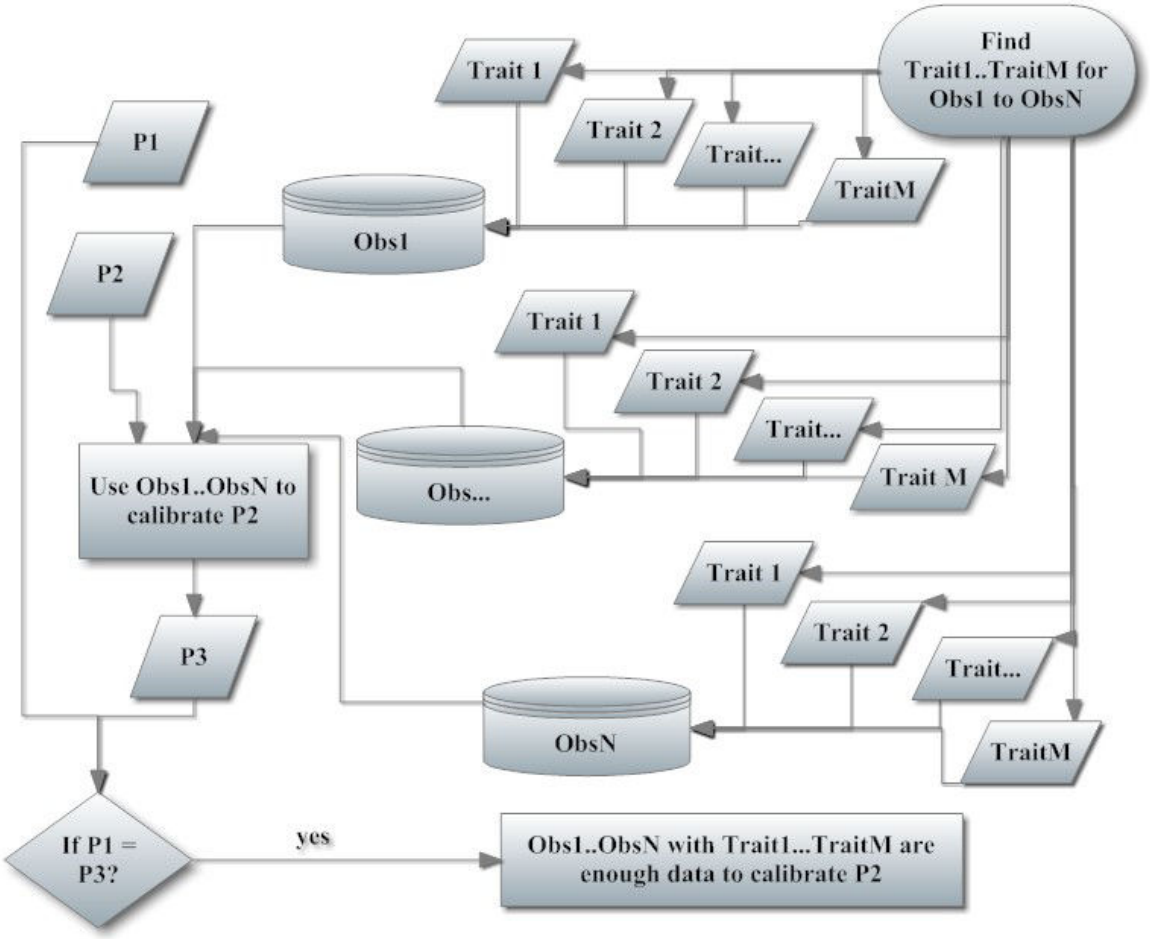

Fig. 4.2: Flowchart of the general strategy for the "proof of concept" test. Produce M trait data with P_1: named $(Trait_1, \cdots, Trait_M)$. From N climatic scenarios, we get the observations set $(Obs_1, \cdots, Obs_N)$, which are used to calibrate model parameters, starting the algorithm from an initial value P_2. The resulting parameter vector estimate is P_3. If P_3 is equal to P_1, N scenarios are sufficient to estimate model parameters.

number of scenarios and amount of observed traits on model inversion. The theoretical data set is generated from 27 years of real environmental information obtained in a meteorological station near Toulouse (South of France) from 1971-2007, including day-by-day information of global radiation, temperatures and rainfall (data described in section 2.3.1).

A virtual genotype parameter vector is chosen (denoted P_1). With the environmental information for the 27 years, 27 virtual observation data are simulated with SUNFLO with different kinds of details. Several experiments are designed, choosing different

amounts of observation scenarios N, diverse observation traits M, to test parameter estimation functions under a variety of conditions. Likewise, the initial value P_2 for the numerical algorithm of parameter estimation is chosen at different distances from P_1, to test the robustness of the numerical estimator regarding the initial condition.

Statistical Test

Bootstrapping is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample, most often with the purpose of deriving robust estimates of standard errors and confidence intervals on parameters. Bootstrapping is adopted for our statistical test to analyze the robustness and accuracy of the estimation methodology. The general strategy of such statistical property test is described in Fig. 4.3.

A parameter vector P_0 is used to generate virtual experimental data for M observed traits $(Trait_1, \cdots, Trait_M)$ and in N environmental scenarios $(Obs_1, \cdots, Obs_N)$. Then random perturbations of the observation vectors $(Obs_1, \cdots, Obs_N)$ are generated to produce a new groups of observation data: $(SampleObs_1, \cdots, SampleObs_N)$, which represents noisy observation data. The perturbation process is repeated K times to generate K samples of N groups of observation data: $(Sample_1Obs_1 \cdots Sample_1Obs_N)$ $, \cdots, (Sample_KObs_1 \cdots Sample_KObs_N)$. They represent K samples of virtual experimental data in N environments. For each $(Sample_iObs_1, \cdots, Sample_iObs_N)$, a parameter estimation is performed and an estimate P_i is deduced, for all $1 \leq i \leq K$. We can then compute the mean value, variance and confidence intervals for $P_1, \cdots, P_K$. The difference between the mean value and P_0 represents the bias of the method, and the standard deviation is a good indicator of its accuracy. The same virtual genotype is chosen as in the "proof of concept" test with its SUNFLO model parameters (denoted P_0 this time) and the same 27 climatic scenarios are chosen. 27 virtual observation data $Obs_1 \cdots Obs_{27}$ are simulated with SUNFLO, with several levels of details for the

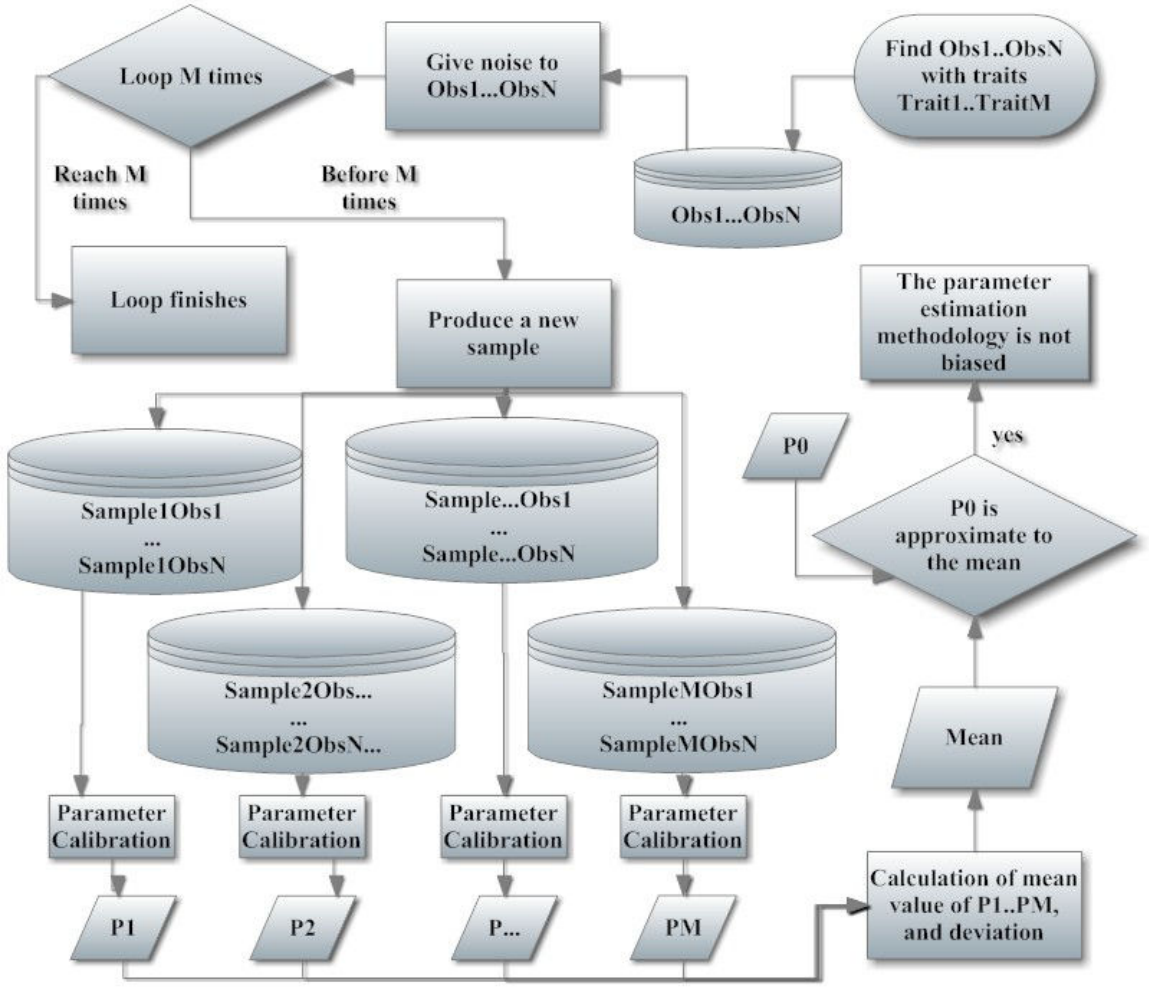

Fig. 4.3: Flowchart of the general strategy for the statistical test. 1) Generate virtual experimental data $(Obs_1, \cdots, Obs_N)$ of plant growth traits $(Trait_1, \cdots, Trait_M)$ based on parameter vector P_0. 2) Give random perturbations on $(Obs_1, \cdots, Obs_N)$ to produce $(SampleObs_1, \cdots, SampleObs_N)$. Repeat to generate K samples of perturbed observations. Each sample is used to produce an estimate, resulting in $P_1 \ldots P_K$ and their mean value and variance.

observed traits. Several experiments are designed to evaluate the bias and accuracy of the estimation method by choosing different amounts of traits and scenarios, and adjusting the level of noise and the number of samples. The perturbation is a multiplicative noise with a normal distribution, of mean 0 and standard deviation sigma. The level of noise is given by sigma ($\sigma = 0.05$, $\sigma = 0.1$, etc.)

4.2.2 Results

Proof of Concept Result

Through the "proof of concept" results, it is verified that the model SUNFLO does not need heavy trait data in every scenario for model inversion. If only one trait data $MS\_graine\_lastDay$ (seed biomass at harvest) is available in two scenarios, two parameters can be estimated with large initial perturbation. As illustrated in table 4.1, parameters $SFimax$ and Eb_0 are changed respectively by ratio 1.1, ratio 1.5 and ratio 2.0 to initialize the estimation algorithm. With two years trait data, $MS\_graine\_lastDay$ in 1971 and 1972, the estimated parameter values are the same as the original values. The successful calibration works for all genotypes and any two years in the 27 environmental year input. The choice of $MS\_graine\_lastDay$, corresponding to the sunflower final yield, is of course in keeping with the most important data of interest in breeding programs. Likewise, if we try to estimate 10 parameters

Tab. 4.1: "Proof of concept" test on 2 parameters

Trait	MS_graine
year 1971	271
year 1972	331

Parameters	Original Value	Ratio 1.1	Ratio 1.5	Ratio 2.0	Calibrated Value
SFimax	200	220	300	400	200
Eb_0	1.0	1.1	1.5	2.0	1.0

(some important genotype-dependent parameters and some genotype-independent parameters in contrast as shown in table 4.2), the tests show that 12 year scenarios are sufficient to retrieve the theoretical values, whatever the level of perturbation, or the choice of the 12 years among the 27 years.

Tab. 4.2: 10 significant parameters for SUNFLO model and the values for the virtual genotype used in statistical test

Parameters	LAI_a	SFimax	Eb_0	Eb_c	LAI_Kei	LAI_b	Position_SFimax	PHS	Phy2	Coeff_ext
Values	400	200	1.0	1.999	0.01379	1	15.4	0.05	16.34	0.96

Statistical Test Result

Based on the positive results of the "proof of concept" test, in the statistical property test, MS_graine is chosen as the main trait for parameter estimation. Different tests are performed as described in the Material and Methods section, and all show encouraging results regarding both bias and accuracy of the estimation methodology. An example of these tests is shown in table 4.3. For the bootstrap analysis, we use

Tab. 4.3: An example of the statistical test results ($sigma = 0.05$, 100 samples)

Parameter Set 1		Parameter Set 2				Parameter Set 3					
SFimax: 200		SFimax: 200		Eb_0: 1		SFimax: 200		Eb_0:1		PHS:0.05	
Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
201.17	7.28441	200.299	2.48166	1.01879	0.109574	200	e-8	1	e-8	0.503798	0.022322

a multiplicative normal perturbation ($sigma = 0.05$) to the trait MS_graine for 20 scenarios, and generate 100 perturbed samples. The statistical test is performed for three parameter sets, $SFimax$, $SFimax$ and Eb_0, and $SFimax$, Eb_0 and Eb_c. For the three parameter sets, the bias is quite small as well as the standard error. Test cases performed with higher levels of noise ($sigma = 0.1$) showed the same results. These results show the good robustness property of the methodology.

4.2.3 Conclusion

The proposed parameter estimation methodology MSPE relies on the idea that the level of information necessary for model inversion in the estimation process of plant models can be obtained with few trait data in a large number of environmental scenarios, and not necessarily with a lot of trait data, thus making this methodology adapted to breeding programs. This idea already suggested in Jeuffroy et al. [2006b] is here implemented and tested on the SUNFLO model.

When the data from which the parameter are estimated are virtual, that is to say when they are generated by the simulation of the model, the estimation methodology is extremely efficient (12 environmental scenarios are sufficient for the estimation of 10 parameters). Moreover, the test case was performed with environmental scenarios recorded in the same location, which means with weaker variations than if they were obtained in very different locations. It is necessary to note that such virtual experimental data should correspond to a 'perfect' model, a model describing perfectly the real plant growth. The robustness of the method is also very encouraging; little bias and good accuracy is observed in case of data perturbations. As complex breeding relationships exist among those genotypes, accurate parameter estimation should help develop the research to link model parameters and genes.

4.3 Implementation and Optimization Issues in MSPE.

In this section, we consider real data situations, with issues to consider for efficient and robust implementation of the method. Contrary to the virtual data case, corresponding to a perfect model, real parameter estimation problems prove more difficult to solve. It appears difficult to estimate a large number of parameters: system identifiability appears non-trivial. For this reason, we use global sensitivity analysis (Saltelli et al. [2008]) to select the most important parameters to estimate and thus reduce the problem of non-identifiability. Moreover, the yield is not a convex function of the parameters, which makes the use of descent methods (of Newton type) dangerous to use, since they may converge to local minima. Other non-convex optimization algorithms (simulated annealing, particle swarm optimization) are tested to circumvent this issue.

4.3.1 Parameter Selection

Ideally, all the parameters should be estimated. However, in real situations, it appears impossible, for different reasons. Therefore, the choice of the parameters to estimate is driven by several considerations. First, models like SUNFLO or CORNFLO have already been studied in other contexts, so that it is possible to provide reasonable parameter values. Some of the parameters are also considered as genotype-independent, so that their values can be fixed *a priori* from previous studies (for example for SUNFLO from the heavy experiments conducted in the first place for model calibration Lecoeur et al. [2011]). Therefore, the first argument for the choice of the parameters to estimate is generally given by the list of genotype-dependent parameters.

A few preliminary studies are then conducted to anticipate some estimation difficulties, particularly continuity and convexity studies. An application of sensitivity analysis method can finally be used to rank model parameters according to their importance on the output.

To illustrate the study, we use the CORNFLO model and dataset in section 2.2 as an example of preliminary model analysis for parameter selection and implementation of the estimation algorithm.

Continuity and Convexity

The parameter continuity and convexity analysis can be used to detect the estimation difficulty of model parameters. Each parameter of CORNFLO model was analyzed. The parameter value range is around its value estimated by direct measurement. Final grain yield $MSgraine$ in one scenario is simulated as a function of the parameter value and shown in Fig. 4.4. Since the environmental input is a 100 % irrigated scenario, water stress parameters RT, RE and RO have no effect on yield, and therefore are not considered in this analysis.

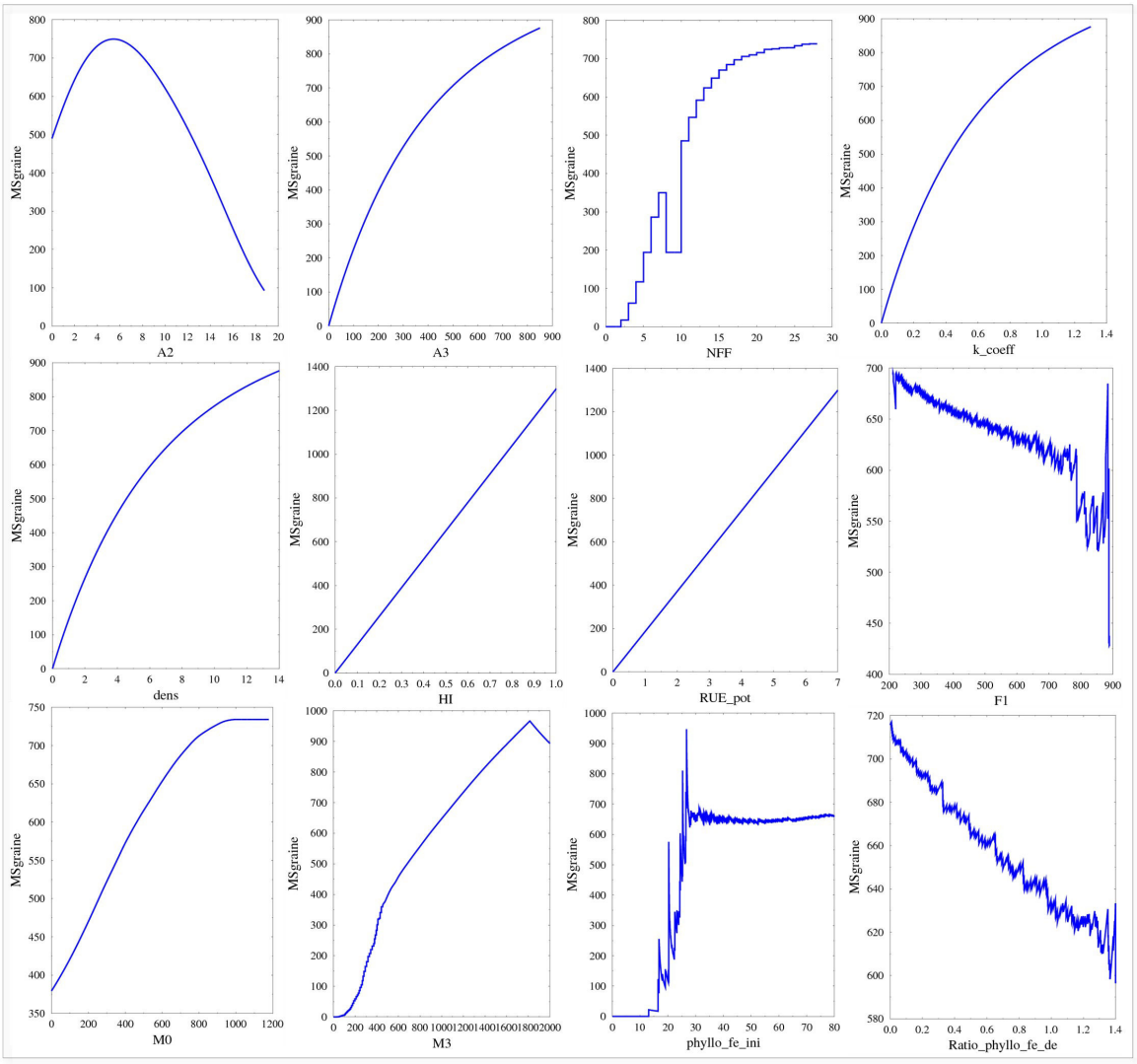

Fig. 4.4: Continuity of $MSgraine$ with respect to each parameter in CORNFLO model

Yields for multiple scenarios were also simulated as functions of the model parameters in their supposed ranges of variation. The example graphs of parameter $phyllo\_de\_ini$ in 2, 4, and 50 scenarios are shown in Fig. 4.5. The x-axis is still the variation of parameter value. The y-axis is the mean of yield values for all the scenarios. The increase of the number of scenarios damped the function irregularity.

Based on these graphs, we classify parameters into three classes: 1) parameters with smooth curves of $MSgraine$ variations, including $dens$, $A2$, $A3$, k_coeff, HI, RUE_pot, $M0$ and $M3$; 2) parameters with irregular curves of $MSgraine$ variations

including $F1$, $phyllo\_fe\_ini$, $phyllo\_de\_ini$ and $Ratio\_phyllo\_fe\_de$; 3) parameters to which the dependence of $MSgraine$ is discontinuous (discrete functions for example) including NFF. Parameters in the first class are easier to estimate.

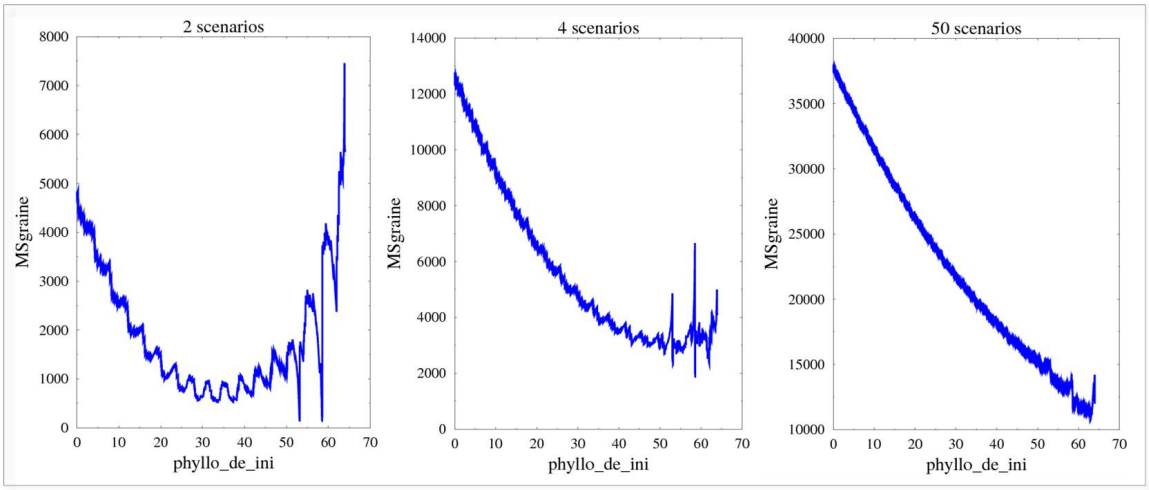

Fig. 4.5: Continuity of parameter $phyllo\_de\_ini$ in CORNFLO model based on multiple scenarios

Another interesting test is presented in Fig. 4.6, which shows the effect of an increase of the number of scenarios on residual sum of squares $RSS(n)$, which calculates the discrepancy between the data and an estimation model with $RSS(n,p) = \sum_{i=1}^{n}(Y_i(p) - Y_i(p_0))^2$, where n is the number of scenarios (here it is set as 2, 4 and

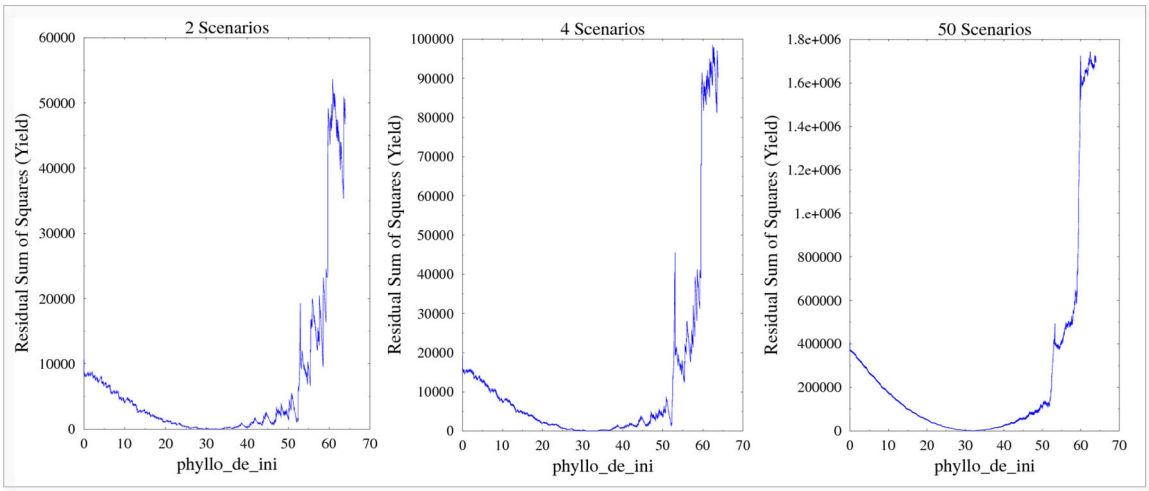

Fig. 4.6: Convexity test for 2, 4, 50 scenarios of the parameter $phyllo\_de\_ini$

50); p is the varying parameter values for the convexity test (0 to two times of measured parameter value in this test), whose simulation of yield for i^{th} scenario is $Y_i(p)$; p_0 is the original set of measured parameter values, whose yield simulation $(Y_i(p_0))_{1\leq i\leq n}$ are regarded as target observations, *i.e.* parameters are calibrated to reach the values which can reproduce such observation. We can see that the regularity of the cost function curve is improved, with less local minima and a better convexity, thus reducing the optimization difficulty.

Parameter Sensitivity

Sensitivity analysis is the study of how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input [Saltelli et al., 2008]. Sensitivity analysis can be useful in the modeling process for a range of purposes such as understanding relationships between input and output variables and searching for errors in the model [Pannell, 1997]. Here we use it to figure out yield sensitivity to each parameter in order to identify the most important parameters in the model, regarding yield elaboration.

There are a large number of approaches to performing a sensitivity analysis, including local methods(e.g. adjoint modeling[Cacuci, 2003; Cacuci et al., 2005] and automated differentiation [Grievank, 2000]), a sampling-based sensitivity [Helton et al., 2006] (e.g. input-output scatter plots), methods based on emulators (e.g. Bayesian[Oakley and O'Hagan, 2004]), screening methods (e.g. elementary effect method [Campolongo et al., 2007]), variance based methods [Homma and Saltelli, 1996; Saltelli et al., 2000; Wu et al., 2012], high dimensional model representations [Li et al., 2006, 2002], and methods based on Monte Carlo filtering [Hornberger and Spear, 1981; Saltelli et al., 2004]. In general most procedures adhere to the following outline: 1, specify the target function of interest; 2, quantify the uncertainty in each input (e.g. ranges, probability distributions); 3, run the model a number of times using some design

of experiments; 4, select a method for assessing the influence or relative importance of each input factor on the target function. Here we use a sampling-based method SRC (standardized regression coefficient,Helton et al. [2006]) as a first step to assess model nonlinearity. Since nonlinearity can not be neglected, we turn to variance based method Sobol [Sobol, 1993] for the sensitivity analysis of the models under investigation. Sobol method uses a unique decomposition of the model into summands of increasing dimensionality. All terms within the decomposition can then be calculated using multiple integrals. It has advantages of testing parameters sensitivity from individual effect and interactions, even though its computational cost is heavy and Morris method [Morris, 1991] is sometimes preferred for large dimension problems. Sobol method is used in this thesis for its good performance and the exhaustiveness of the information it can provide regarding the interactions and different types of effects in the model. We recall below the basic elements on Sobol's method.

Variance-based methods are a class of probabilistic approaches which quantify the input and output uncertainties as probability distributions, and decompose the output variance into parts attributable to input variables and combinations of variables. The sensitivity of the output to an input variable is therefore measured by the amount of variance in the output caused by that input. These can be expressed as conditional expectations [Homma and Saltelli, 1996; Saltelli et al., 2000]. For example in Sobol method, considering a model $Y = f(X)$ for $X = X_1, X_2, ...X_N$, there exist functions such that the output can be written as in equation 4.1.

$$f(X_1, \ldots, X_N) = f_0 + \sum_{i=1}^{N} f_i(X_i) + \sum_{1 \leq i \prec j \leq N} f_{ij}(X_i, X_j) + \ldots + f_{1\ldots N}(X_1, \ldots, X_N) \quad (4.1)$$

The solution to this problem can be written in terms of conditional expectation of Y:

$$\begin{aligned} f_0 &= E[Y], \\ f_i(X_i) &= E[Y \mid X_i] - f_0, \\ f_{ij}(X_i, X_j) &= E[Y \mid X_i, X_j] - f_i(X_i) - f_j(X_j) - f_0 \\ &\vdots \end{aligned} \tag{4.2}$$

Using formula 4.1, the variance of Y, $D = Var(Y)$ can be written as:

$$D = \sum_{i=1}^{N} D_i + \sum_{1 \le i \prec j \le N} D_{ij} + \ldots + D_{1\ldots N} \tag{4.3}$$

where

$$\begin{aligned} D_i &= Var(E[Y \mid X_i]), \\ D_{ij} &= Var(E[Y \mid X_i, X_j] - E[Y \mid X_i] - E[Y \mid X_j]), \\ &\vdots \end{aligned} \tag{4.4}$$

The Sobol sensitivity indices are defined by

$$S_i = \frac{D_i}{D}, \qquad S_{ij} = \frac{D_{ij}}{D}, \qquad \ldots \tag{4.5}$$

where

$$\begin{aligned} S_i \ge 0, \qquad S_{ij} \ge 0, \qquad \ldots \\ \textstyle\sum_{i=1}^{N} S_i + \sum_{1 \le i \prec j \le N} S_{ij} + \ldots + S_{1\ldots N} = 1 \end{aligned} \tag{4.6}$$

S_i is the first order index. It explains the part of the variance of Y that can be explained by the fluctuations of X_i. S_{ij} is the second order index. It explains the part of the variance of Y explained by the interaction of the fluctuations of the variable X_i and X_j. The total index ST_i is the sum of all indices relative to X_i, which expresses the sensitivity of Y with respect to X_i by itself or through its interactions with other variables [Ammari et al., 2012]. The computation method is based on Monte Carlo simulation and the estimation algorithm proposed by Wu et al. [2012] which is implemented in the Pygmalion platform. The results of CORNFLO parameters

sensitivity to model input "Yield", analyzed respectively by SRC and Sobol methods, are shown in table 4.4, and table 4.5. The configuration of our sensitivity analysis is: 1024 samples are taken into account in a Monte Carlo simulation, and two repetitions of sensitivity analysis are adopted to confirm the result's convergence. In order to

Tab. 4.4: SRC sensitivity index of the most influential parameters (with index > 1%).

NFF	$A2$	$M3$	HI	RUE_pot	$A3$	$F1$	$M0$	k_coeff
0.25	0.21	0.14	0.06	0.06	0.03	0.03	0.03	0.02

Tab. 4.5: Sobol first order and total order index values for the most influential parameters (with index > 1%).

	NFF	$A2$	$M3$	HI	RUE_pot	$A3$	$F1$	$M0$	k_coeff
First Order Index	0.26	0.21	0.13	0.06	0.06	0.05	0.03	0.03	0.02
Total Index	0.39	0.28	0.13	0.1	0.1	0.12	0.08	0.09	0.1

screen parameters, (select the parameters that can be fixed to some *a priori* values), the total order indexes SY_i are used: parameter j is screened if $ST_j < \epsilon$, with ϵ a threshold, for example 0.01. All these methods show that parameter $phyllo\_fe\_ini$, $phyllo\_de\_ini$ and $Ratio\_phyllo\_fe\_de$ have negligible effect. Therefore they can be screened and will not be estimated, at least in a first run.

4.3.2 Optimization Issues

In order to find the best optimization method for MSPE application on Cornflo model, Gauss-newton method [Walter and Pronzato, 2006], Simulated Annealing method [Laarhoven and Aarts, 1987], and Particle Swarm Optimization algorithm [Shi and Eberhart, 1998] are compared on their optimization capacity and computation efficiency.

Optimization Algorithms

Gauss-newton algorithm is able to solve non-linear least squares problems. It iteratively searches for the parameter values θ of a vector of m estimated parameters $(\theta_1, \ldots, \theta_m)$, which get the minimum of the sum square error $f(\theta)$:

$$f(\theta) = \sum_{i=1}^{n} r_i^2(\theta) \tag{4.7}$$

where r is a vector of the residual squared error between n observations and n estimations. The iteration starts from the initial guess of parameters $\theta^{(0)}$, with the step gradient as:

$$\theta^{(s+1)} = \theta^{(s)} - (J_r^T J_r)^{-1} J_r^T r(\theta^{(s)}) \tag{4.8}$$

where J_r is a $n \times m$ Jacobian matrix equating:

$$J_r = \begin{vmatrix} \frac{\partial r_1}{\partial \theta_1} & \cdots & \frac{\partial r_1}{\partial \theta_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_n}{\partial \theta_1} & \cdots & \frac{\partial r_n}{\partial \theta_m} \end{vmatrix} \tag{4.9}$$

Simulated Annealing algorithm (SA) and Particle Swamp Optimization algorithm (PSO) are all computational intelligence methods with iterative optimization techniques. Their parameter searching space is all m-dimensional space representing respective values of parameters $\theta_1 \ldots \theta_m$. Their objective is to find the best position (with minimum cost function value) in the m-dimensional space.

SA mimics the metallurgical process of annealing. It compares the current solution $\theta^{(s)}$ with a randomly generated potential solution $\theta^{(s+1)}$. If the system energy has decreased (the cost function is more minimized by new solution), $\theta^{(s+1)}$ is accepted and set as the current system position. If not, an acceptance probability P_A of the new solution is calculated:

$$P_A = exp(-\partial f * (\frac{T_0}{T})) \tag{4.10}$$

where ∂f is the change of cost function between existing solution and new solution. T_0 is the initial temperature value. T is the current temperature, which is high at the beginning and is gradually cooling down with the cooling speed ratio α. A random value between $[0-1]$ is compared with P_A. If P_A is greater than this value, the new solution $\theta^{(s+1)}$ is accepted. This is repeated until the system freezes into a steady state. T_0 and α are meta-parameters to configure the algorithm's efficiency.

PSO mimics the biological behavior of a swarm of bees. The swarm of bees (particles) cooperates to find the target (global minima in optimization) in a partially random way. Each particle receives information from other members about their swarm's best position and records its best position of current minimum. Thereby, a particle's movement in search space for each iteration is calculated:

$$\begin{aligned} v^{(s+1)} &= c_1 v^{(s)} + c_2(p^{(s)} - \theta^{(s)}) + c_3(g^{(s)} - \theta^{(s)}) \\ \theta^{(s+1)} &= \theta^{(s)} + v^{(s+1)} \end{aligned} \tag{4.11}$$

where $v^{(s)}$ is the particle's self velocity at s^{th} iteration, $p^{(s)}$ is the minimum position for this particle, and $g^{(s)}$ is the minimum position for all particles. c_1, c_2 and c_3 are given parameters, which together with the amount of particles t used for optimization, are named meta-parameters to configure PSO algorithm. Iterations are executed until a stable state is achieved.

Optimization Capacity

The limit of the cost function found by Gauss-newton algorithm is a stationary point if the algorithm converges. However, the convergence is not guaranteed. When the initial guess is far from the minimum or the Hessian matrix $J_r^T J_r$ is ill-conditioned, the

algorithm converges slowly or cannot converge. In the MSPE tests in this thesis, the initial guess of SUNFLO and CORNFLO parameters are from another estimation by direct measurements. To ensure the scenarios provide enough information for reaching convergence, different combinations of parameter sets are tested and the amount of scenarios is increased, for example in the feasibility test in section 4.2, 12 scenarios are used for estimating 10 SUNFLO parameters. Other disadvantages of Gauss-newton algorithm include that it fails when the derivative of the cost function cannot be computed (which is the case for example for the NFF parameter, the number of leaves); it performs well for local minimum optimization problem, but for multi local minima problems, it is easily stuck at local minima and fails to reach the global minimum.

SA algorithm is able to deal with highly non-linear models and find global minimum. In SA processes, T is high at the beginning, which allows the algorithm to search in a wide range of solutions, including many that are worse than the current solution to avoid sticking to local minimum. Meta-parameters α and T_0 are critical factors. A low T_0 may cause the failure of reaching global minimum; a high T_0 may bring in unnecessary cost of time for the algorithm execution. Similarly, α affects whether the parameter space are sufficiently searched for finding the global minimum. They are adapted to concrete optimization problems for ensuring the optimization capacity. While SA has good functionality in global minimum searching, it has the problem of costly computing time. For example, in this thesis tests, the configuration of $T_0 = 1000$ and $\alpha = 0.0995$ succeeds to estimate six CORNFLO parameters, but it costs around 20 hours to achieve one round of parameter estimation. When an experiment involves many rounds of parameter estimation, such as the one in section 4.4 for the bootstrap algorithm in order to evaluate the estimation uncertainty, the computation becomes very heavy.

As another meta-heuristic artificial intelligence based algorithm, PSO has also good

capacity in finding global minimum. Its meta-parameters are also critical on ensuring the optimization capacity. In the tests of this thesis, its meta-parameters are $c_1 = 0.01$, $c_2 = 0.05$, $c_3 = 0.02$, and particle amount $t = 3000$. The quantitative comparison between SA and PSO for assessing their capacity in finding global minimum is not investigated in this thesis. It is only demonstrated here that both algorithms are able to be well used for MSPE optimization in CORNFLO model. PSO has also the problem of heavy computational cost. These algorithms' computational efficiencies are discussed in the next section.

Computational Characteristics and Efficiency

Two types of computers are used in the thesis. The first computer "Dell P8600" has an Intel dual core processor with the feature of 2.4 GHz. For the FlOPS (the floating-point operations per second) benchmark, which is a principle measurement of computer performance, this computer has 16370 MFLOPS. Another computer is named "mesocenter", which is a large computing machine located in Ecole Centrale Paris, France. It has 10 TFLOPS and comprises nearly 1000 calculation units. The mesocenter has obviously more computing power than the first computer. For example, the 50 scenarios convexity analysis in Fig. 4.6 has 20000 samples to construct the x-axis and for each sample, 50 simulations need to be produced. In total, it needs 1 million simulations, which takes around 48 hours computed by "Dell P8600", but only 2 hours 35 minutes computed with the "mesocenter", *i.e.* the calculation of a simulation in "Dell P8600" spends around 0.2 seconds, while its calculation spends around 0.01 seconds in "mesocenter". In our tests, "mesocenter" calculation is around 20 times faster than the "Dell P8600". The results for parameter estimation are comparable.

Gauss-newton algorithm has much faster computation than SA and PSO algorithms. For estimating 6 CORNFLO parameters in MSPE, Gauss-newton spends few minutes in "Dell P8600" while SA and PSO take tens of hours in "mesocenter". But as

the optimization capacity is the main concern, SA and PSO are preferable choices despite their slow computations. Between the two algorithms, PSO seems to have better computational efficiency than SA as already underlined by Qi et al. [2010] in a plant growth modeling context. With the algorithm configurations detailed in the previous section, 6 parameters optimization in SA needs around 20 hours, while PSO needs around 10 hours, showing less computational cost. However, as the algorithms' optimization capacities are not fully tested, changing meta-parameters for algorithms' configurations may lead to different performances. Considering tests carried out in this thesis, PSO is recommended for the excellent balance of its optimization capacity and computational efficiency.

4.3.3 Parameter Non-Estimability

Identifiability is a property which a model must satisfy in order for inference to be possible. We say that the model is identifiable if it is theoretically possible to learn the true value of this model underlying parameter after obtaining an infinite number of observations from it. Mathematically, this is equivalent to saying that different values of the parameter must generate different probability distributions of the observable variables. For example, when a model has been written in such as way that two or more parameters are nonseparable, the non-separable parameters are not estimable with any data set. In this case, the parameters are often referred to as non-identifiable [Ponciano et al., 2012] or structurally non-identifiable. Rannala [2002] showed a simple exponential modeling case with this type of non-identifiability. Beside the identifiability problem, another situation can lead to parameters' non-estimability (also called practical non-identifiability). It takes place when the sampled data contains absolutely no information about the parameter of interest, yet other data sets might. We term such cases as identifiable but non-estimable [Ponciano et al., 2012] or practically non-identifiable.

The identifiable but non-estimable problem makes the determination of data sufficiency important in MSPE. When the environment scenarios for parameter estimation are few, only limited parameters can be estimated. In our parameter estimation practice, we observed that adopting more environment scenarios can help estimating more parameters. The sufficiency of observation data limits the amount of parameters which we can estimate in MSPE. Moreover, when not enough scenarios are available, MSPE computes parameter estimates outside their validity range (or at the boundary since some constraints are imposed for the parameters to remain in the validity range). When it occurs, we consider that we are in a non-estimability situation. The range of validity for five CORNFLO parameters, $M0$, $M3$, $A2$, $A3$, k_coeff is given in (table 4.6). For example, the parameter $A2$, representing the leaf rank of biggest

Tab. 4.6: Parameter optimization searching value range

	Lower	Upper		Lower	Upper
M0	442	1326	M3	739	2216
A2	7	21	A3	323	968
k_coeff	0.27	0.8	RUE_pot	1.75	5.25

surface leaf has the range $[7; 21]$. This is a reasonable setting because of $A2$'s physical feature. When using 50 randomly chosen scenarios for the estimation, $A2$ has big probability to reach its value boundary, as illustrated in one specific estimation provided in table 4.7. But some combinations of 50 scenarios do not exhibit such

Tab. 4.7: Estimated parameters values from 50 scenarios: parameters reaching boundary

M0	M3	A2	A3	k_coeff
1022	1283	7	385	0.45

problem, as illustrated for another choice of 50 scenarios in (table 4.8). We can see

Tab. 4.8: Estimated parameters values from 50 scenarios: parameters not reaching boundary

M0	M3	A2	A3	k_coeff
1157	1200	16	668	0.68

that estimated parameter values are not close to their value range. The reason that some combinations of scenarios make it possible to estimate parameters while some others cannot is the scenarios correlation which determines the information amount of the scenarios set. It means that scenarios for the table 4.7 have bigger correlations than those for table 4.8. The environmental scenarios similarity and difference and their effect on multi-scenario parameter estimation will be explained in Chapter 5.

To estimate more parameters with MSPE, a larger number of scenarios is required. To estimate four parameters $M0, M3, A3, k$, as least as 30 scenarios chosen randomly in the database (real data, as presented in section 2.2 for one sepcific genotype) are enough for parameter estimation. We do not have the data insufficiency problem. 50 scenarios are generally enough to estimate five parameters $M0, M3, A2, A3, k$, but not always. Finally, adopting 500 scenarios make most of cases estimable and 720 scenarios makes it work (table 4.9). More scenarios make it possible to estimate more parameters.

Tab. 4.9: Estimated parameters values from 500 and 720 scenarios.

	M0	M3	A2	A3	k_coeff
500 Scenarios:	1134	1339	9	441	0.4
720 Scenarios:	879	1452	9	441	0.4

However, some parameters cannot be estimated no matter how many scenarios are taken. Because MSPE has the special characteristic that it uses only small amount of data for each input dataset, some parameters have practical identifiability problem arising with the MSPE strategy. Since the non-identifiability is practical and not structural, such problems would not occur if more traits were available. In our case for example, the calculation of yield $MSgraine$ in CORNFLO, the harvest index parameter HI cannot be estimated, even by increasing the number of scenarios. It can be solved by using two traits (data $MSgraine$ and $MStot$) in one scenario of MSPE, or by improving harvest index mechanisms, for example as in the the SUNLAB model

in Chapter 3.3.

4.3.4 Conclusion

From the above analysis, we can see that among all CORNFLO parameters, $F1$ has strong oscillating curves in continuity test, which make it more difficult to estimate, $phyllo\_de\_ini$, $phyllo\_fe\_ini$ and $Ratio\_phyllo\_fe\_de$ have low sensitivity, and HI is non-identifiable when the available data are restricted to crop final yield. Besides, $dens$ and NFF, representing field density and crop average total amount of leaves, can be measured from field experiments with low uncertainty. Therefore parameters in the first priority to estimate are: $A2$, $A3$, k_coeff,RUE_pot, $M0$ and $M3$. These parameters are used in our following tests to prove MSPE prediction capacity.

From the previous section 4.3.2, we know that MSPE parameter optimization can be a difficult issue from a numerical point of view. An important reason is the existence of the multiple local minima in cost function. Increasing scenario amount can help damping estimation and optimization difficulty in MSPE. Fig. 4.5 has shown that the increase of scenario amount make an oscillating function curve smoother.

4.4 Evaluation of MSPE Estimation Accuracy

Deriving from the level of information in the experimental data used in model inversion for parameter estimation, there is a degree of uncertainty for model parameter values. The distribution of parameter and its statistics are of interest because they represent the accuracy and confidence of relative parameters' estimation. In this thesis, we study the distribution of four CORNFLO model parameters $M0$, $M3$, $A3$, and k_coeff based on 720 real experimental scenarios (dataset in section 2.2), with the following steps:

1 A single test TM is to choose randomly a set from the $C(n, m)$ complete combinatorial group of m scenario samples among n total scenarios. n is 720 of 100% irrigated scenarios in this study, for one specific genotype. The m scenario samples will produce an estimated value for a selected parameter $ParamM$.

2 We repeat TM 100 times to produce samples of estimation values $ParamM\_0$, $ParamM\_1 \ldots ParamM\_100$. A discrete approximation of the distribution is thus obtained. The mean and standard deviation of the estimate distribution is denoted as $paramMean\_M$ and $paramSd\_M$.

3 Different values of m are chosen: 40, 50, 70, 100 and 200 in this study. Then $paramMean\_40 \ldots paramMean\_200$ and $paramSd\_40 \ldots paramSd\_200$ can be obtained to research on the change of parameter distribution due to the increase in the number of scenarios for the estimation.

Fig. 4.7, Fig. 4.8 and Fig. 4.9 are examples of parameter distributions with m respectively as 40, 70 and 200. Our hypothesis of the m effect on standard deviation values of parameters distribution is that the bigger m is, the smaller standard deviation will be. When m reaches big enough value, the variance will converge as in Fig. 4.10. The standard deviation in our tests with m respectively equal to 40, 50, 70, 100, 200 are shown in Fig. 4.11. Current results show that the variance is not converged yet. But for all the four parameters, the standard deviation has an obvious tendency to decrease when m increases. The bigger the number of scenarios is, the more accurate parameter estimates are.

This study is crucial for the use of the MSPE method to characterize genotypes and to discriminate between them. For the statistical test to consider whether the estimated parameters for two different genotypes can be considered as different or not, the power of the test will be bigger for better estimate accuracy, and thus with a bigger number of scenarios.

Likewise, regarding the predictive capacity of the model, a reduction of input uncertainty (here parameter estimates) should also lead to better performances. The impact of the number of scenarios used for the estimation is thus considered in the next section.

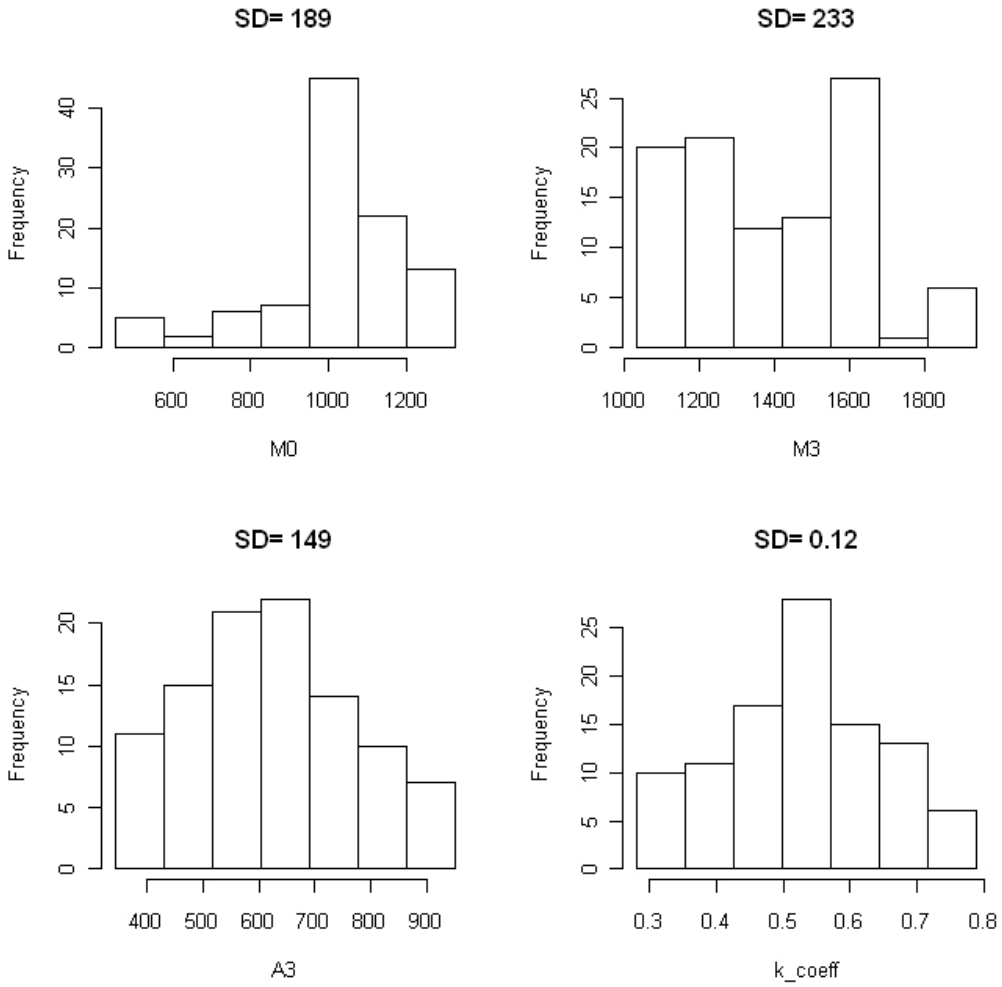

Fig. 4.7: Distribution of 100 samples for four CORNFLO parameters based on 40 scenarios

4.5 Evaluation of MSPE Prediction Error

Model evaluation aims at determining how well a model fulfills its initial objectives. For crop models, whose main purpose is yield prediction, the evaluation consists in

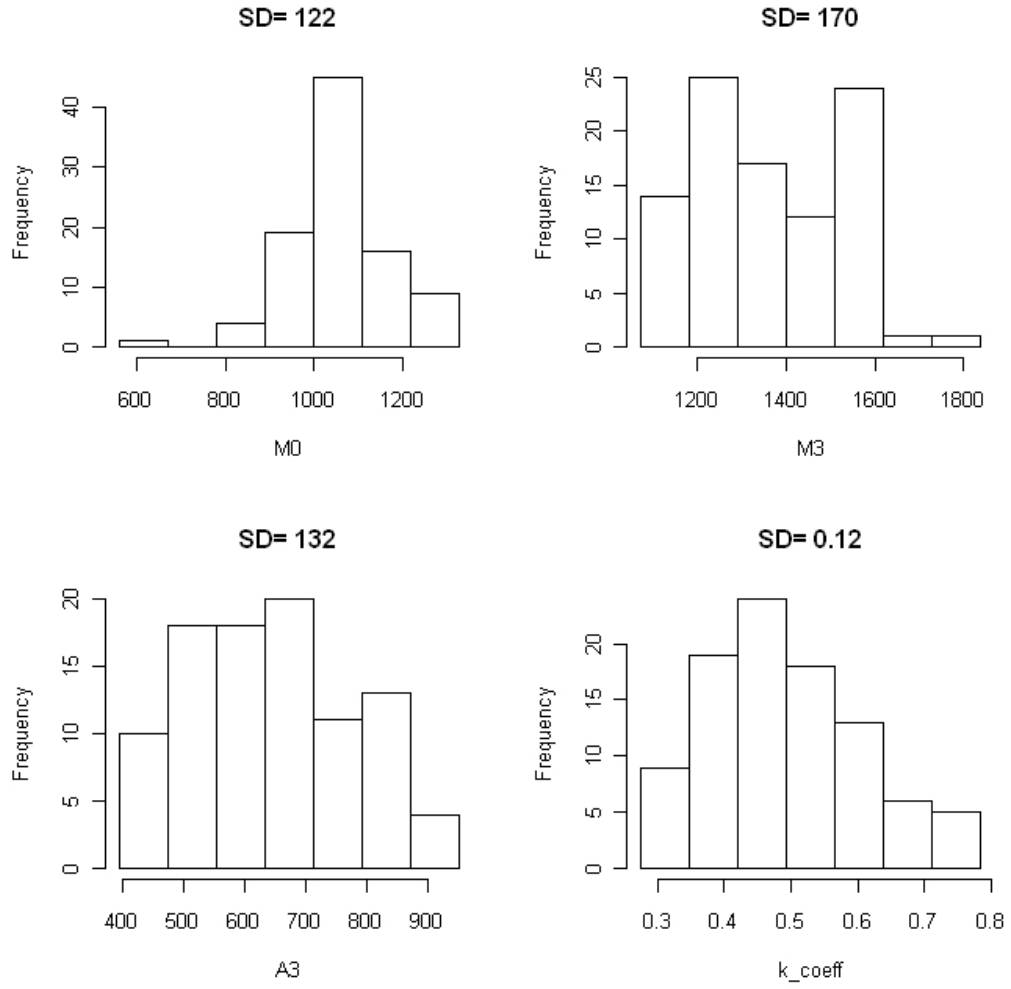

Fig. 4.8: Distribution of 100 samples for four CORNFLO parameters based on 70 scenarios

comparing between observed and predicted values, graphically and with numerical and statistical measures of model quality [Wallach, 2006]. Model evaluation is important because it tells the users and developers the quality of their model and may also give hints to improve the model quality. For MSPE, a hypothesis is that the scenario amount can influence the prediction error / predictive ability of the model associated with its estimated parameters. Two tests based on CORNFLO model were carried out to research on this feature.

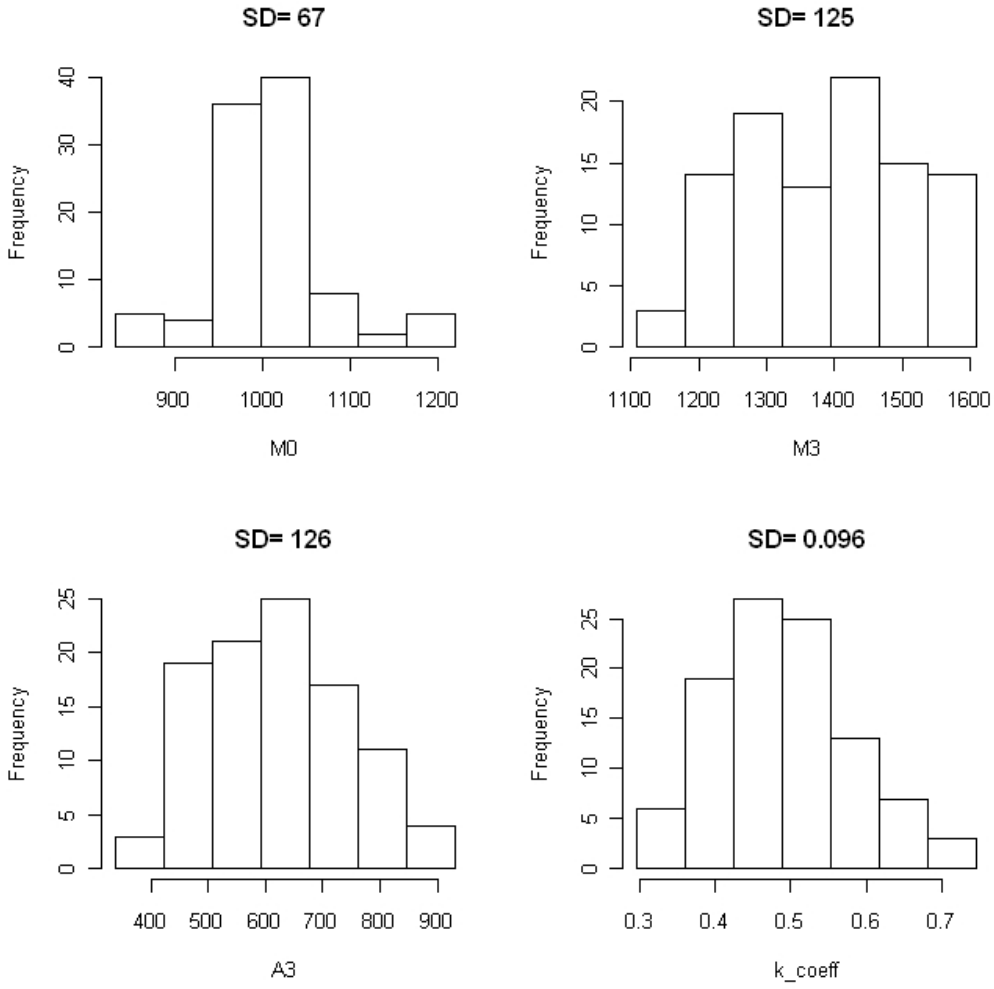

Fig. 4.9: Distribution of 100 samples for four CORNFLO parameters based on 200 scenarios

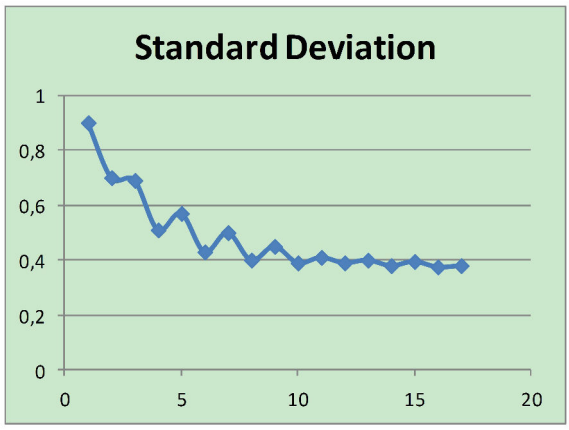

Fig. 4.10: The illustration of our hypothesis on the evolution of the standard deviation value

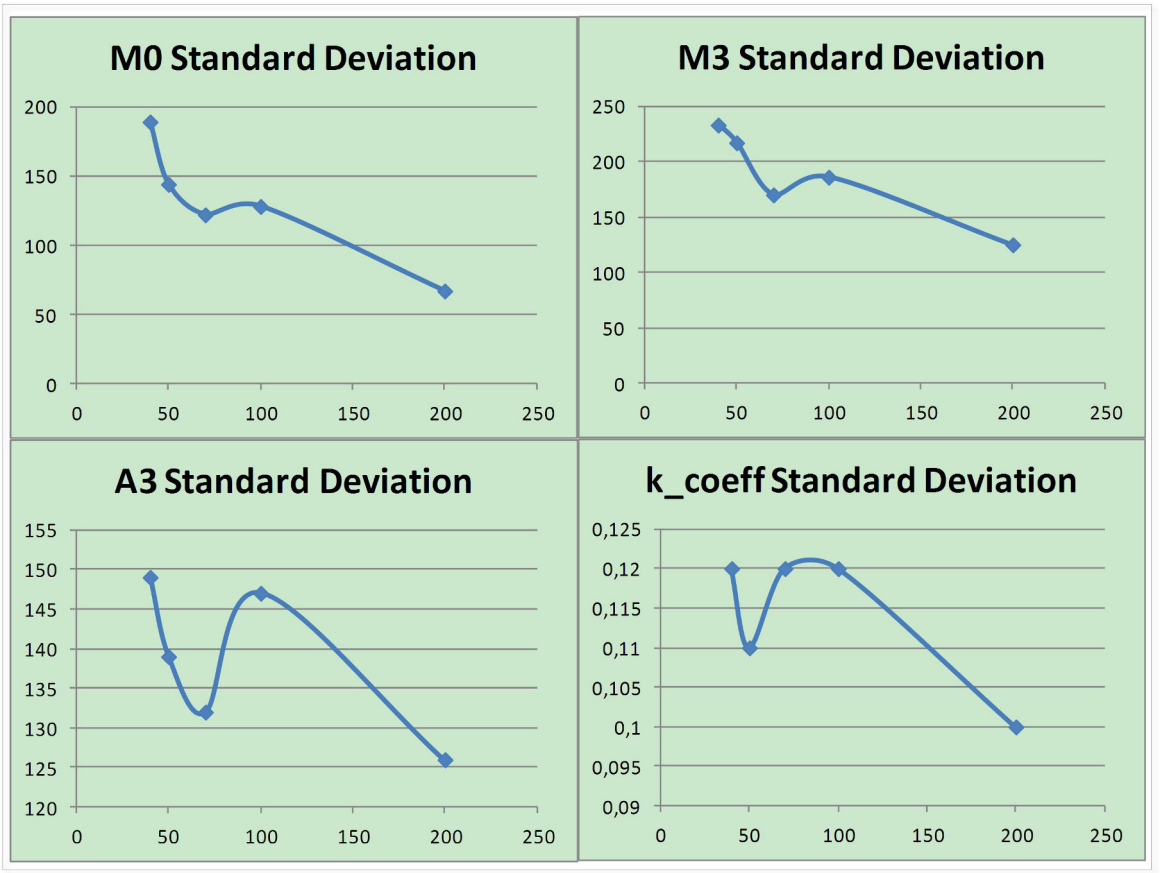

Fig. 4.11: The illustration of standard deviation values of parameter distribution in MSPEJ

4.5.1 Number of Scenarios Increase Test

In this test, among the 720 available scenarios (for one genotype, and no water stress), 240 are isolated to serve as the validation set. The remaining 480 scenarios serve as the learning set. Among these, we choose samples, with an increasing number of scenarios from 10 to 480. These samples of scenarios are used for MSEP, with Gauss-Newton method as optimization algorithm. They produce different parameter estimation result $P10 \ldots P480$. Each parameter set result P is used to simulate the trait of interest, the yield, on the 240 scenarios of the validation test. For each estimate, the root mean square error of prediction $RMSEP = \sqrt{(1/N)\sum_{i=1}^{N}(Yi - \hat{Y}i)^2}$ is used to assess the difference between observation and simulation, *i.e.* model pre-

diction ability, where N is the total amount of considered scenarios in the validation test (240 in this test), Yi is the measured experimental value of i^{th} scenario (experimental yield), and $\hat{Y}i$ is the simulated value with the parameters estimated from a subsample of the learning set. $RMSEP$ has the same unit as yield. Therefore, the statistics $RMSEP_{10}, \ldots, RMSEP_{480}$ represent model parameters' prediction ability along with the number of scenarios.

Fig. 4.12 illustrates the $RMSEP$ results of two sub tests following the above methodology scheme. In the prediction error test 1, the optimization method Gauss-newton is carried out with initial parameter values as 0.9 times of measured CORNFLO parameters and the prediction error test 2 uses 1.1 times of those values. It is demonstrated

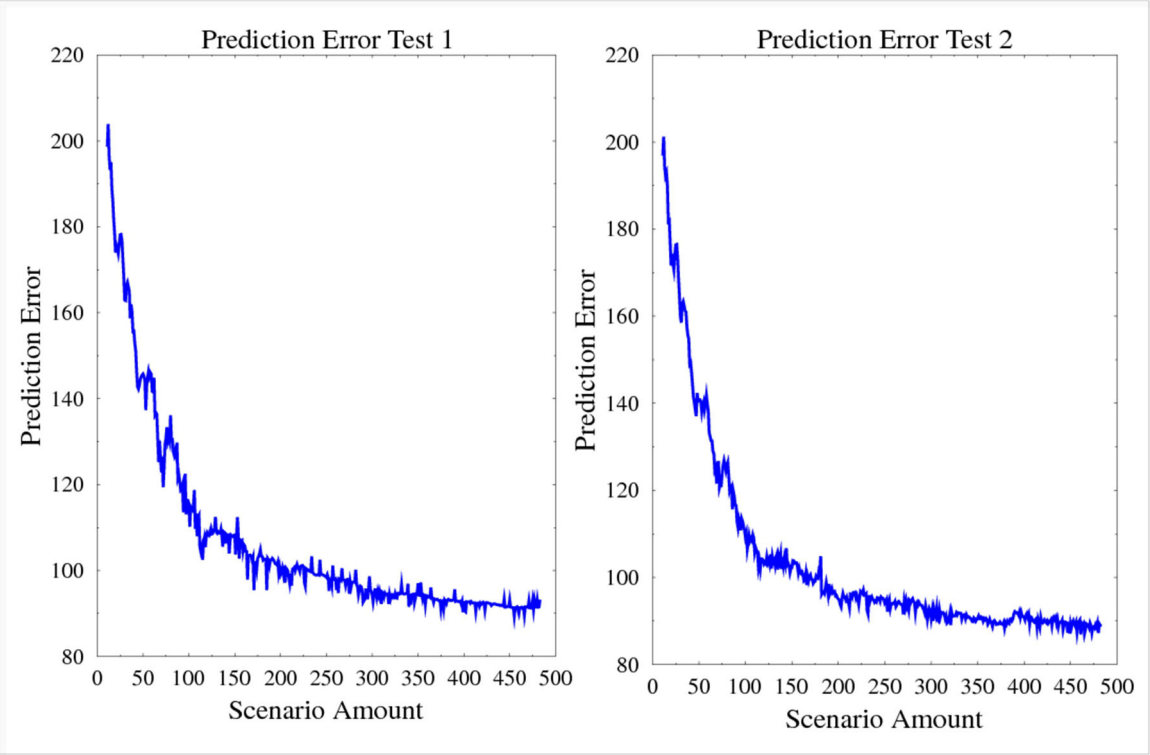

Fig. 4.12: Prediction error (Root mean square error) changes along with the increase in the number of scenarios. In prediction error test 1, the optimization algorithm Gauss-newton method takes 0.9 times of *a priori* CORNFLO parameters as initial parameter values for optimization; Prediction error test 2 takes 1.1 times of those values

that the increase in the number of scenarios used by MSPE can improve the estimated

parameters' prediction ability, as the prediction errors are shown to reduce in both sub tests. Since the Corn yield is around 1000, the prediction errors with the same unit reduce from around 200 to 90, showing that the prediction error decreases from around 20 % of yield to 9 %, which corresponds to what was expected.

However, there are two points in this test that can be improved. Firstly, it was performed with Gauss-newton optimization method. As explained in section 4.3.2, the specific cost function of CORNFLO model to optimize decides that the Gauss-newton method is easily stuck in local minima. The estimation values in this figure may not be global minima. Although the tests with different initial values used in Gauss-newton still show that the prediction error is reduced with an increase of the number of scenarios, even for different local minima combinations, the prediction ability should be improved if based on global minima. The second problem is that, since the 480 scenarios for parameter estimation and 240 scenarios for model evaluation are random scenarios in dataset 2.2, the specific choice of these two sets may affect the result. For these reasons, we consider a more complex test in the next section.

4.5.2 Cross Validation Test

In this test, PSO algorithm with proper meta-parameters configuration instead of Gauss-newton is used to ensure reaching global minima in optimization. The root mean square error of prediction $RMSEP$ (Eq. 4.12) is used to measure the difference between observation and simulation, which also has the same unit as yield.

$$RMSEP = \sqrt{E\left\{[Yi - \hat{Y}i(\hat{\theta})]^2|\hat{\theta}\right\}} \tag{4.12}$$

where θ is the estimated parameter set for a specific situation. The $RMSEP$ is different from the one used in the previous test because it takes into account all possible interested situations while in the previous section, only a specific configuration for the

choice of scenarios was considered. In detail, this test takes advantage of the test in section 4.4, whose steps are adapted as follows: for a single test TM, while m scenario samples are taken from n for parameter estimation, 200 random scenarios from the rest of $n-m$ scenario produce prediction simulations; 100 different choices of TM are realized, and thus 20000 (100×200) scenarios' simulation points are produced and used to compute the expectation of prediction error $RMSEP(m)$; setting values for m as 40, 50, 70, 100, and 200, we can get the prediction error's evolution with the increase of the number of scenarios (obtaining $RMSEP(40), \ldots, RMSEP(200)$). The test is designed with cross validation strategy. Cross validation is based on the principle of data splitting, and aims to remove the bias coming from arbitrarily assigning certain selected scenarios for estimation and prediction. It has been tested on crop models by Jones and Carberry [1987] and Colson et al. [1995]. It is used in this thesis to eradicate the choice of scenarios influence on our prediction error measurement. Under this strategy, the calculation of prediction error is illustrated in Fig. 4.13. The left graph shows the reduction of prediction error from scenario amount 40 to 200. The prediction error from initially 8% is reduced to 7.5%. It proves that more scenarios used in MSPE cause better prediction ability for its estimated parameters. However, the gain in prediction is far more reduced with an increase in the number of scenarios, showing that with this strategy (global minimization + jackknife averaging of the estimation), 40 scenarios seem sufficient to ensure a good level of prediction.

The mean squared error of prediction $MSEP(\hat{\theta})$, which is the square of $RMSEP(\hat{\theta})$, can be decomposed into three terms $MSEP(\hat{\theta}) = \lambda + \delta + \alpha$. $\lambda = E\left[var(Y|X)\right]$, is the population variance term representing the observation trait (Y, yield)'s variance for fixed values of models' environmental inputs X, *i.e.* it measures the missing consideration of important environmental variable's effect. δ is the square of model bias, and depends on the form of the model. $\alpha = E\left\{var[\hat{Y}i(\hat{\theta})|X]\right\}$ is the the model variance due to the variability of model parameters. The variance of model parameters

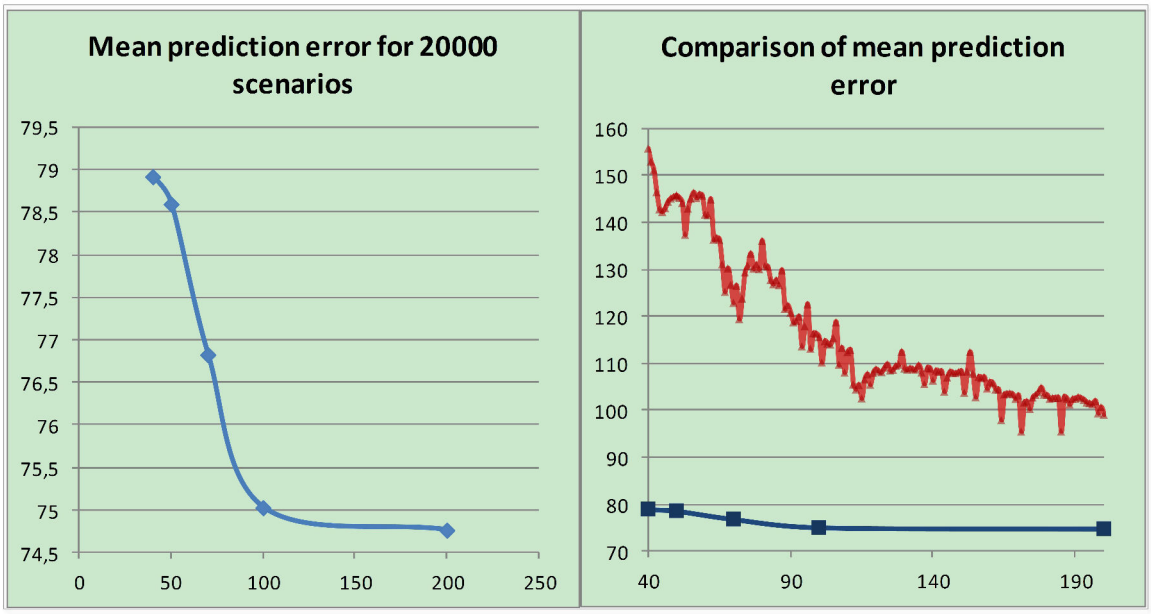

Fig. 4.13: Prediction error is reduced with the number of scenarios, from 40 to 200. Left graph: the prediction error reduction in cross-validation test; Right graph: the comparison between prediction error in cross validation test(blue line) and in number of scenarios increase test (in section 4.5.1, red line). Y-axis is the root mean square error of prediction with the same unit as yield; X-axis is the the number of scenarios used for MSPE parameter estimation.

$var[\hat{Y}i(\hat{\theta})|X]$, which has been proved to reduce for the increase of scenario amount, explains the prediction error's reduction.

The right graph is the comparison of prediction error in this cross validation test and in the previous simple test increasing the number of scenarios (in section 4.5.1). For smaller numbers of scenarios, the improvement made by the cross validation test quite big (the percentage of error reduction is about 7.5% of yield for 40 scenarios), while the improvement is lower when the number of scenarios is bigger (for 200 scenarios, the prediction error in cross validation test is reduced of 2.5% of yield compared to the simple test with Gauss-Newton and without cross-validation). It demonstrates the superiority of global optimization and averaging. However, since both convexity and estimation accuracy are improved when the number of scenarios is increased (as shown in figures 4.6 and 4.11 respectively), the improvement of global optimization and averaging is lower. Besides, both methods show that the prediction error converges

(to $\lambda+\delta$) when the number of scenarios increase. The quantitative analysis of scenario amount's convergence critical values need to be studied in the future: can we compute an error bound allowing us to decide the proper number of scenarios for an acceptable error level *a priori* ?

4.6 Jackknife Based MSPE: an Extended Version

In this section, we propose an alternative version of the MSPE method, which is based on the Jackknife strategy. A distribution of the estimate is thus obtained, with a standard deviation of estimated parameter value and whose mean appears to be a more stable estimate than the direct MSPE, specifically in the case of low numbers of scenarios.

4.6.1 Delete-m Jackknife Estimator and MSPEJ Methodology

The jackknife estimator introduced by Quenouille (1949) has become an important tool in simulation and data analysis. It creates a series of statistics, usually a parameter estimate, from a single data set by generating that statistic repeatedly on the data set leaving each time some data values out (not used for the calibration). It is used mainly for bias reduction and interval estimation. A generalized definition of Jackknife estimator is as below: n samples are splitted into g groups of size h where $n = g * h$. Let $Y_1, \ldots, Y_n$ be a sample of independent and identically distributed random variables. Let $\hat{\theta}$ be an estimator of the parameter θ based on the sample of size n. Let $\hat{\theta}_{-i}$ be the corresponding estimator based on the sample of size $(g-1) * h$, where the i-th group of size h has been deleted.

A popular form for many researches is when $g = n$ and $h = 1$, which is also the case discussed here. For the bias reduction aspect, define

$$\tilde{\theta}_i = \ n * \hat{\theta} - (n-1) * \hat{\theta}_{-i} \quad i = 1, \ldots, n. \tag{4.13}$$

The estimator

$$\tilde{\theta} = \frac{1}{n} * \sum_{i=1}^{n} \tilde{\theta}_i = n * \hat{\theta} - (n-1) * \frac{1}{n} * \sum_{i=1}^{n} \hat{\theta}_{-i} \tag{4.14}$$

has the property that it eliminates the order $1/n$ term from a bias of the form Miller

[1974]

$$E(\hat{\theta}) = \theta + a1/n + O(1/n^2) \tag{4.15}$$

For its interval estimate aspect, the jackknife confidence interval CI is calculated as

$$CI(95\%) = \tilde{\theta} \pm 1.96\sqrt{\frac{var}{n}} \tag{4.16}$$

where

$$var = \frac{1}{n-1} * \sum_{i=1}^{n} (\tilde{\theta_i} - \tilde{\theta})^2 \tag{4.17}$$

The above jackknife method is called delete-1 jackknife because for the estimator $\hat{\theta}_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$, the $\hat{\theta}_{-i}$ is constructed by leaving out one observation Y_i. Instead of removing one single observation from the samples, the delete-m jackknife subsamples are computed by leaving out m observations from $Y_1, \ldots, Y_n$ at a time. The subsamples of the number $\binom{n}{m} = \frac{n!}{m!(n-m)!}$ produce the sample distribution for calculating the corresponding bias reduction or confidence interval. Delete-m jackknife is adopted normally for non-smooth statistics, such as median of samples. It can fix the inconsistency problem for jackknife subsamples in delete-1 jackknife method. MSPEJ adopts delete-m jackknife method, researching on the bias reduction compared with MSPE and MSPEJ's interval estimate. It also research the effect of different number of m on bias and interval estimate features in delete-m jackknife based MSPEJ.

MSPEJ is defined as follows: there are totally n sample scenarios $Y_1, \ldots, Y_n$. m scenarios ($m <= n$) are used to produce a point estimator $\hat{\theta}_{m(i)}$, where $n - m$ sample scenarios are taken out from n. Since the complete combinatorial group number $c = \binom{n}{m}$, there produced $\hat{\theta}_{m(1)}, \ldots, \hat{\theta}_{m(c)}$, totally c estimated point values which construct a distribution of parameter values relating to m. It is denoted $MSPEJ_m$. An example on SUNFLO model parameters $SFiMax$ and $position\_SFiMax$ is obtained with this estimation, whose parameter distribution are shown in Fig. 4.14. Note that the

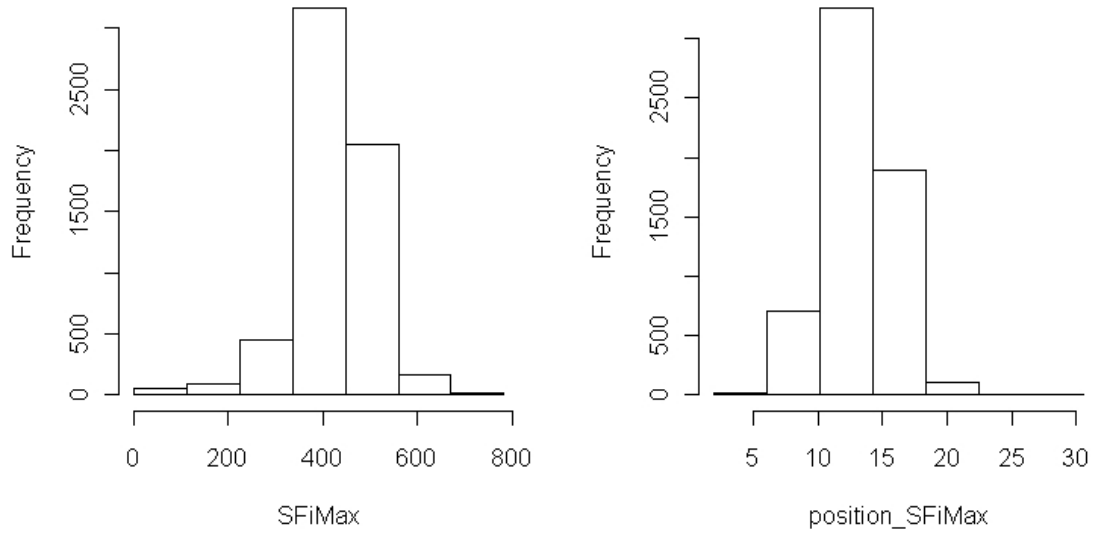

Fig. 4.14: The distribution of all combinatorial samples for using 17 scenarios out of 21 scenarios in MSPEJ for SUNFLO Model

method used in section 4.2.2 to compute parameter distributions and uncertainty is pretty similar to the MSPEJ here proposed, except for the sampling strategy among the whole set of scenarios.

4.6.2 A Real Data Test

Compared with MSPE, besides the fact that MSPEJ offers mean and standard deviation of parameter distribution, it is derived to overcome the problem raised when only small numbers of environmental scenarios are available in MSPE. Normally in breeding programs the number of available environmental scenarios for a given genotype is large enough, resulting in estimated parameters with acceptable estimation and prediction errors. Even though some well designed models such aim at predicting crop yields in a large range of environmental conditions, its predictive ability still face some limitations resulting from the impossibility to handle all local environmental conditions precisely. Uncertainty also comes from experimental data. Huge

scenarios used in estimation will diminish these influence of uncertainty. When the number of scenarios is small, the uncertainty influence is amplified. For example, a series of tests are carried out on the Sunflower genotype "Melody" with the SUNFLO model. Detailed experiments were conducted in SupAgro Montpellier to determine via direct measurements the SUNFLO parameters of the genotype 'Melody' [Lecoeur et al., 2011], denoted $P0$. For the same genotype, 21 experimental scenarios recorded in different places and years (Chapter 2.1.2) are available, with environmental data, crop management information, and final seed yield. In each scenario, only the trait $MS\_graine\_lastDay$) is used for parameter estimation. Gauss-newton algorithm is utilized for generalized least square error optimization. The estimation of two parameters $SFimax$ and Eb_0 based on 21 scenarios produce an unreasonable results that $SFimax$ obtains the value 4904, which is unrealistic from a biological point of view. Two strategies to manage the limited scenario data based on MSPEJ are tested.

The first strategy screens scenarios to keep those whose yield simulations are good compared with experimental observations. Six scenarios are selected. A test takes two out of the six scenarios producing 15 combinatorial groups to estimate two parameters $SFimax$ and Eb_0. The seed yields in the six selected scenarios are supposed to have errors coming from field measurements. To correct the effect of measurements errors on resulted parameters, the six scenarios are grouped two by two, and it thus obtains 15 combinatorial groups (corresponding to the number of different choices of two elements in a group of six). Every group has two scenarios to calibrate 2 parameters. Calibrated parameters values in every group are recorded. At last, the average of parameter values in all groups is admitted as the final estimated parameter for this genotype, and is compared to the original value of P0. The standard error is a measure of parameter uncertainty. Table 4.10 shows the estimation result. Likewise, the 6 scenarios can be grouped three by three, 20 combinatorial groups are thus obtained, which provide three scenarios to estimate three parameters $SFimax$, Eb_0, and Eb_c in every group

Tab. 4.10: Measured value of two parameters, their estimated values in 20 samples and the calculation of mean and standard error by MSPEJ.

	Measured Value	Calibrated parameters values in every sample															Calibrated values	
		Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10	Group 11	Group 12	Group 13	Group 14	Group 15	Mean	SE
Sfimax	674.3	337.2	674.3	84.29	168.6	337.2	1348	1348	674.3	674.3	674.3	337.2	674.3	674.3	674.3	674.3	623.7	396.8
Eb_0	1.1	1.122	0.998	1.470	1.211	1.128	0.992	1.001	1.068	1.080	1.087	1.181	1.104	1.111	1.120	1.126	1.120	0.111

(20 is the number of 3-combinations from the set of 6 scenarios), the results are shown in table 4.11; Grouped by four, 15 combinatorial groups are obtained to calibrate four

Tab. 4.11: Measured value of three parameters, their estimated values in 20 samples and the calculation of mean and standard error by MSPEJ.

	Measured Value	Calibrated parameters values in every sample																		Calibrated values	
		G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16	G17	G18	G19	G20	Mean	SE
Sfimax	613	594	613	613	548.9	613	613	464	306.5	306.5	56.4	1226	307	613	613	613	613	613	613	613	1872	621.2	359
Eb_0	1	0.74	1.04	0.97	0.92	1.05	1.04	0.70	1.15	1.14	1.29	1.16	1.46	1.09	1.35	1.1	1.1	1.11	1.11	1.12	0.86	1.08	0.18
Eb_c	1.999	0.45	1.991	1.563	1.165	1.99	1.956	0.07	1.986	1.788	e-4	3.1	3.91	1.984	4.16	1.986	1.98	1.99	1.98	1.98	0.958	1.849	1.03

parameters $SFimax$, Eb_0, Eb_c, and LAI_Kei (see table 4.12). There is a good agreement between the estimated values and the measured values of the parameters. But the standard errors of parameters are quite high, thus showing that the number of scenarios from which the parameters are estimated is not sufficient. For a practical usage case, this strategy is not possible to be carried out because while our aim is to estimate parameters, we don't have produced simulation for screening scenarios depending on their prediction ability. Here it is used to verify the possibility to use MSPEJ for small amount of scenarios under a realistic situation.

Tab. 4.12: Measured value of four parameters, their estimated values in 15 samples and the calculation of mean and standard error by MSPEJ.

	Measured Value	Calibrated parameters values in every sample															Calibrated values	
		Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10	Group 11	Group 12	Group 13	Group 14	Group 15	Mean	SE
Sfimax	613	628.4	931.4	1031	613	673.7	980.8	591.8	613	1215	306.5	973	697.9	1127	613	495.7	766	249.6
Eb_0	1	0.824	0.828	0.811	1.054	0.951	0.671	1.21	1.071	0.827	1.199	1.255	1.208	0.857	1.1	1.03	0.992	0.176
Eb_c	1.999	0.811	0.991	0.984	1.991	1.654	0.247	3.104	1.987	0.987	2.298	3.75	2.95	0.99	1.98	1.38	1.74	0.94
LAI_Kei	0.01379	0.0076	0.0259	0.0096	0.0125	0.0081	0.025	0.0106	0.0104	0.021	0.006	0.025	0.022	0.046	0.0328	0.0083	0.018	0.011

The second strategy screens the estimated result samples. A test takes 12 scenarios from 21 scenarios to estimate 10 parameters shown in table 4.13. The knowledge of parameters obtained from direct measurement gives us reasonable ranges for parameter values. The interval area of parameter values (around 0.5 times and 1.5 times of the measured value) screens out the estimated samples with parameter values outside of the range. Estimated parameters values are reasonable compared with direct measurement values. Their standard errors are also reduced compared with the first strategy. Those estimated result samples are screened out possibly for two reasons. One is that those scenarios used for estimation may include too many scenarios with bad prediction. Another is that those scenarios used for estimation do not provide sufficient information level and we face non-estimability / practical non-identifiability problems, as explained in section 4.3.3. The environment selection methodology introduced in Chapter 5 can assess the cause of data insufficiency in scenario sets. This methodology is more robust than using directly MSPE by discarding estimated result samples suffering from this effect.

Tab. 4.13: Measured value of ten parameters, the calculation of their estimation values' mean and standard error by MSPEJ.

Parameter Name	Measured Value	Calibrated values	
		Mean	SE
SFimax	613	591.084	89.1862
Eb_0	1	1.02512	0.0744608
Eb_c	1.999	1.97377	0.0957802
LAI_a	400	404.12	22.9021
LAI_Kei	0.01379	0.018096	0.00788875
Phy2	16.34	16.191	1.79002
LAI_b	1	0.939928	0.22347
pos_SFiMax	15.4	15.2662	0.788192
PHS	1	1.01076	0.0377396
coeff_extinct	0.96	0.992638	0.0751752

In conclusion, we have carried out tests on estimating SUNFLO parameters confronting to data in Chapter 2.1.2 based on MSPEJ methodology. Up to ten parameters can be calibrated with good agreement to the measured values. Parameters $SFimax$,Eb_0, Eb_c, LAI_Kei. $SFimax$ are chosen since they are key parameters in the model, with important differences between genotypes [Lecoeur et al., 2011] and with a very direct biological meaning and potential accurate direct measurements to check the validity of our approach. The others are important parameters that are not easy to compute from field measurements, and for which the method has a strong interest. MSPEJ helps us to test the capacity to calibrate complex parameters. These tests indicate the potential to calibrate unknown parameters for new genotypes with this type of dataset.

4.7 Conclusion

In this chapter, an original parameter estimation methodology MSPE (Multi-scenario Parameter Estimation Methodology) is designed to overcome the difficulty of crop model parameter estimation by model inversion when the experimental cost is heavy. The principle of the methodology is to use large numbers of environmental scenarios with simple traits to estimate parameters rather than detailed experimental data. Its feasibility has been demonstrated by the theoretical test on SUNFLO parameters. Practical solutions for carrying out MSPE are discussed, including optimization issues and the analysis of parameters sensitivity, continuity, convexity, and identifiability. The effect of the number of scenarios used in MSPE on the accuracy of estimation and the prediction error are investigated, indicating that more scenario amount leads to more accurate parameter estimation and better prediction capacity. Finally, an extended version Jackknife based MSPE, entitled as MSPEJ, is developed for a special case of parameter estimation with only small numbers of scenarios, with simple traits.

5. ENVIRONMENT CLUSTERING AND INTERACTION WITH CROP MODELING

The MSPE methodology relies on the level of information available from the different scenarios used for model inversion. This level of information can be enhanced by well choosing the environmental scenarios on which is based the estimation. In this chapter, based on data clustering methods, environment scenarios are studied for two purposes:

- the visualization of environment clusters provides suggestions to multi-environmental trials (section 5.2). We consider a classification of the experimental locations based on climatic / soil data or crop yield information;
- multi-scenarios parameter estimation methodology is improved by selecting scenarios from environment clusters (section 5.3). In this section, we consider clusters of scenarios, based on one year climatic data.

5.1 Data Clustering Methods

The objective of data clustering is to group data objects with more similarity into clusters. It can be achieved by many clustering methods. Typical classes of clustering methods include connectivity based clustering [Hastie et al., 2009] and centroid based clustering [Kanungo et al., 2002].

Connectivity based clustering connects objects based on their distances. Objects are

more related to nearby objects in the clustering result than those farther away (An example result is in Fig. 5.1). Hierarchical clustering method used in section 5.2 is a connectivity based clustering method. The user can visualize the similarity between objects and get a clue of the number of clusters with this method. But it does not provide a clear partitioning of objects dataset. The clusters are chosen by the user according to the similarity hierarchy.

Centroid based clustering is based on an iterative search of the centers of each cluster and each object's distance to the center until it converges, and thus provides the final clustering. K-means clustering method used in section 5.3 is a centroid based clustering method. It produces a clear partitioning of objects, but it needs the input of the number of clusters before the method begins. There are algorithms to guess reasonable number of clusters according to the dataset.

We recall the principle of the algorithms of hierarchical method and k-means methods as used in this thesis.

Hierarchical Clustering. Two strategies can be used to build a hierarchy of objects clusters. Agglomerative strategy is a "bottom up" strategy. It considers each object as a cluster at first, and merge gradually other clusters to build up hierarchy. Divisive strategy is a "bottom down" strategy, which considers all objects as a cluster at first and split them step by step. The merge or split of sets of objects depends on their dissimilarity, i.e. their distances. Two factors are taken into account for calculating the dissimilarity: metric and linkage criteria. The metric approach considers the distance between two objects a and b. Common metrics are euclidean distance $\sqrt{\sum_i (a_i - b_i)^2}$, squared euclidean distance $\sum_i (a_i - b_i)^2$ or Manhattan distance $\sum_i |a_i - b_i|$, where a_i and b_i are coordinates of a and b. Linkage criterion determines the distance between two sets of objects. Common criteria include complete linkage $max\{d(a,b) : a \in A, b \in B\}$, single linkage $min\{d(a,b) : a \in A, b \in B\}$ or average linkage $\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d(a,b)$, where d is the chosen metrics, a and b are

objects in sets A and B.

K-means Clustering. K-means clustering considers every object as a n-dimensional vector. To cluster m objects $X_1, \ldots, X_m$ into k groups $G_1, \ldots, G_k$, we solve the following optimization problem:

$$\underset{G}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{X_j \in G_i} \|X_j - \mu_j\|^2 \tag{5.1}$$

where μ_j is the mean n-dimensional position of group j. The optimization is an iterative process. Before the iteration begins, each group is given a random mean position to produce groups' mean positions $\mu_1, \ldots, \mu_k$. Each iteration includes two steps:

Step 1, Assign each object to the group whose mean position is closest to this object's position. Each group G_i contains a set of objects:

$$G_i = \{X_p \text{ such that } \|X_p - \mu_i\| \leq \|X_p - \mu_j\| \forall\ j \neq i,\ 1 \leq j \leq k\} \tag{5.2}$$

Step 2, Calculate new mean positions for each group:

$$\mu_i = \frac{1}{|G_i|} \sum_{X_j \in G_i} X_j \tag{5.3}$$

The two steps are iterated until it converges to get stable k means and groups.

5.2 Environment Scenarios Clustering Based on Different Information Strategies

Multi-environmental trials are traditionally used to assess cultivar adaptation within a target population of environments [Messina et al., 2006]. The adequate selection of en-

vironment scenarios to represent the target environmental space is therefore becoming a significant process in the development of better-adapted genotypes. The decisions about which environments should be selected to conduct field trials are based on the understanding of environments' characterizations and representative environment scenarios in each category. The need to characterize the environments used for multi-environmental trials has been widely documented (e.g. Comstock [1977]; Cooper et al. [1993]; Loffler et al. [2005]). Attempts to characterize crop environments largely fall into three categories [Messina et al., 2006]: 1, Classification based on climatic and soils data. It is useful for describing environmental variables, but it does not identify the environments' influences on crops ecophyiological functions. 2, Classification based on the statistical analysis of variety performance data. This approach has been widely used, but it does not provide a measure of the environment independent of crop performance. 3, Classification using crop models to integrate weather, soil and management information. Model outputs can be used to produce categorical variables that describe environments [Messina et al., 2006]. In this section, the first and second approaches are adopted to classify environments for Corn, based on the US database presented in 2.2. Another involved approach combining the first and second approaches, considers both climatic, soil data and crop performance data to classify the environments. We entitle the three kinds of approaches as A, B and C Classification.

The clustering results for the environment scenarios is not unique, it depends on the method and the information data used for the classification. Based on hierarchical clustering methods, we adopt three approaches to cluster locations used for the crop field experiments in USA (data described in section 2.2). The variables considered for A-type classification include daily minimal temperature $Tmin$, daily maximal temperature $Tmax$, daily radiation RG, daily precipitation P, and daily evapo-transpiration reference $ET0$. Each location contains 10 years' daily data of the five variables, therefore a 'data' has $365 * 10 * 5 = 18250$ dimensions. The data clustering method classifies

in total 720 locations, each of which has 18250 dimensions. B-type classification considers only crop yield data because this trait is our main research concern. For others test, more traits could be considered for the classification. Each 'data' has thus 10 dimensions. C-type classification considers all the information involved in A and B classification. Each location has 18260 dimensions. Fig. 5.1 is an example of the visualization of environmental clusters under the third approach.

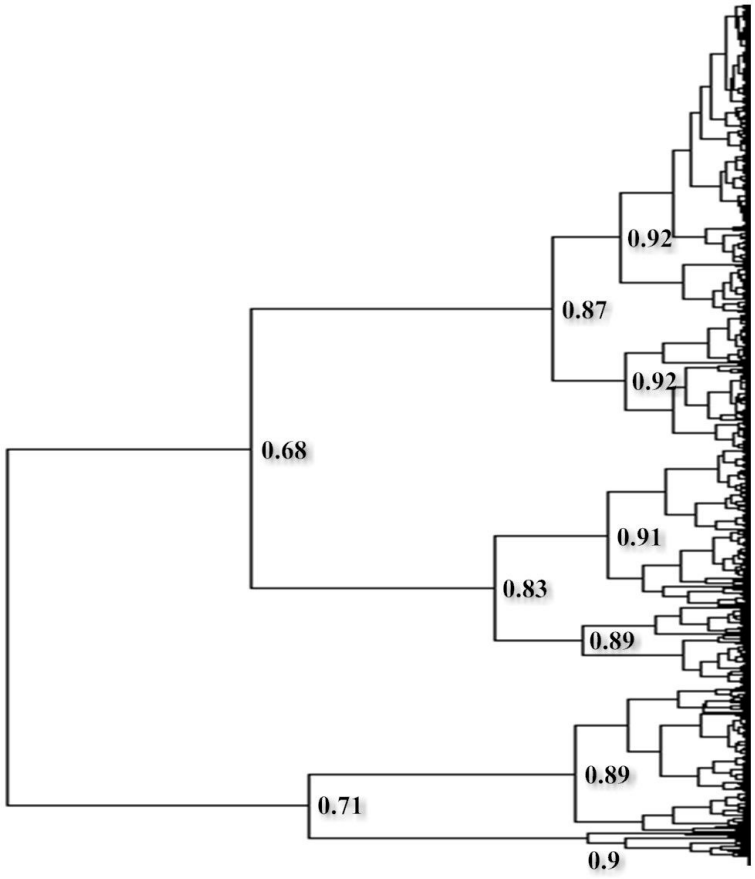

Fig. 5.1: County scenario's clustering based on the combination of environment information and yield

We consider that a level of correlation above 0.9 is signficative enough to make each scenario representative of the cluster. With this 0.9 level, A, B, and C classifications all indicate that six clusters exist for our scenarios. The obtained clusters however are not similar.

For example, scenario county No. “17095”, “17063”, “17105” are all from Illinois state. They belong to the same cluster in both A, B and C classification methods (this group is named “Cluster1” here). County “17131” is also from Illinois state. But it does not belong to the same cluster “Cluster1” as the other three Illinois counties in A classification approach. In B classification approach considering yield, “17131” belongs to “Cluster1”. In C classification approach considering both information, the balance between the two sides finally decides that “17131” should be in “Cluster1”.

Likewise, for example, scenario counties “18113” and “18087” of Illinois state belongs to “Cluster1” in A classification method, but they both belong to a different cluster in C classification method. Among the three classification approaches, clustering results of C classification are recommended because it combines environment and crop features information. Each scenario in the analysis adopts ten years data, which is to avoid the categorization variation from year to year, or at least get a classification result based on long-term historical samples. In further study, a quantitative comparison among the three approaches based on a more advanced clustering method can be carried out to confirm the advantage of C classification.

5.3 Multi-Scenario Parameter Estimation based on Scenario Selection

In practical applications of the multi-scenarios estimation method, facing large amount of crop yield data for estimation, there come two significant questions: “how many scenarios should be used” and “which scenarios should be chosen”. The first question about the influence of the number of scenarios was investigated in Chapter 4 and the second one is discussed in this section. An adjusted methodology “environment scenario selection based multi scenarios parameter estimation methodology (MSPEE)” is proposed here. The strategy of MSPEE is to use data clustering methodology to

cluster environmental scenarios. Then we choose representative scenarios from each cluster in order to make the selected scenarios as different as possible from each other. The hypothesis is that using data clustering selected scenarios in MSPEE instead of random scenario in MSPE can improve the numerical efficiency of the method. Three aspects are supposed to be improved:

1. improvement on practical identifiability. Data insufficiency problem coming from lack of scenarios is supposed to be weakened;
2. improvement on estimation accuracy. The estimation uncertainty (represented by the variance of the estimated parameter) should be reduced in MSPEE compared to MSPE when using the same number of scenarios;
3. improvement on prediction error. Fewer selected scenarios based on data clustering can represent the population of environments. Therefore using the same number of scenarios, the prediction error of estimated parameters in MSPEE should be less than in MSPE.

All tests are based on CORNFLO model and its parameters.

5.3.1 Practical Identifiability

In the MSPE tests of the previous chapter, only four CORNFLO parameters were estimated because for more parameters, the data insufficiency problem arised, specifically for small numbers of scenarios, like 40: six parameters cannot be estimated in most combinations of 40 scenarios.

In this section, K-means Clustering method is used to make scenario categories. Note that contrary to the previous section, we really consider scenarios: two different years for the same county count for two scenarios. Taking one representative scenario from each category provides for the group of selected scenarios better properties in terms of identifiability. For example, 6 parameters can be estimated with the multi-scenario

parameter estimation methodology with 40 scenarios, selected from 40 clusters obtained by environment clustering of the US dataset (described in 2.2), while it was not possible when the scenarios were chosen randomly, see results in Table 5.1.

Tab. 5.1: Parameters estimated from scenarios with high diversity, selected from 40 different environment clusters.

M0	M3	A2	A3	k_coeff	RUE_pot
960	1073	14	541	0.5	4

To illustrate the idea, we reversely choose 40 scenarios from the same cluster whose correlation is 0.91. Table 5.2 shows that it has nonestimibility difficulty, as explained in section 4.3.3: parameters reach the boundary values of the optimization interval, meaning a wrong estimation. Other similar tests have shown the same identifiability problem.

Tab. 5.2: Parameters estimated from scenarios with high similarity, selected from one single environment cluster (of correlation level >0.91).

M0	M3	A2	A3	k_coeff	RUE_pot
664	2027	14	323	0.35	5.25

Since scenarios clustered for MSPEE correspond to one year environmental data (two years in one county count for two scenarios), we found that the clustering tends to classify among years rather than counties. So scenarios from the same year tend to have more similarity than those from the same county or state. It indicates that when choosing scenarios database for MSPEE estimation, scenarios from multiply years are recommended for obtaining widely-used model parameters. However, from a practical point of view in a breeding context, this solution is not optimal in terms of cost. A clustering based on experimental location should be more interesting.

5.3.2 Improvement on MSPE Estimation Accuracy

A similar test as described in section 4.4 to investigate MSPE parameter estimates' distributions and statistics is carried out. It also takes m scenarios from n total scenarios to estimate a set of parameter values, entitled as a single test TM. Repeating the single test TM for 100 times can produce 100 sets of parameter values, thus approximating the parameter estimates' distributions. The difference in the MSPEE test is that a single test TM is not taking m random scenarios from the total n scenarios as in MSPE. It requires the following steps to complete a single test TM:

1 Categorize n scenarios into m groups based on k-means clustering.

2 Take a random scenario S_0 from n total scenarios. Find its group ranking G in m groups.

3 Every scenario's distance is defined by parameters number n building the n-dimensional space. Add S_0 in a set $\mathcal{S}$ which will contain all the m selected scenarios. The center position of $\mathcal{S}$ is the position of S_0 currently.

4 From one of the $m-1$ groups excluding the G^{th} group, take a scenario which is the longest distance (Euclidean distance) from the center position of S; add this selected scenario in $\mathcal{S}$; recalculate the center position of $\mathcal{S}$.

5 Repeat the same strategy in step 4 to take one scenario from all the other groups, resulting in the S with m scenarios. S is used to produce a parameter estimation.

MSPEE manages to estimate 6 parameters. The parameter estimates' distributions based on $m = 40$ scenarios are illustrated in Fig. 5.2. The standard deviations of parameters are different, as shown in the figure, from around 1% ($M0$) to 10% (A3) of parameter value. It indicates that most parameters have already good confidence intervals, while for some of them, the uncertainty may still be reduced by testing with higher number of scenarios.

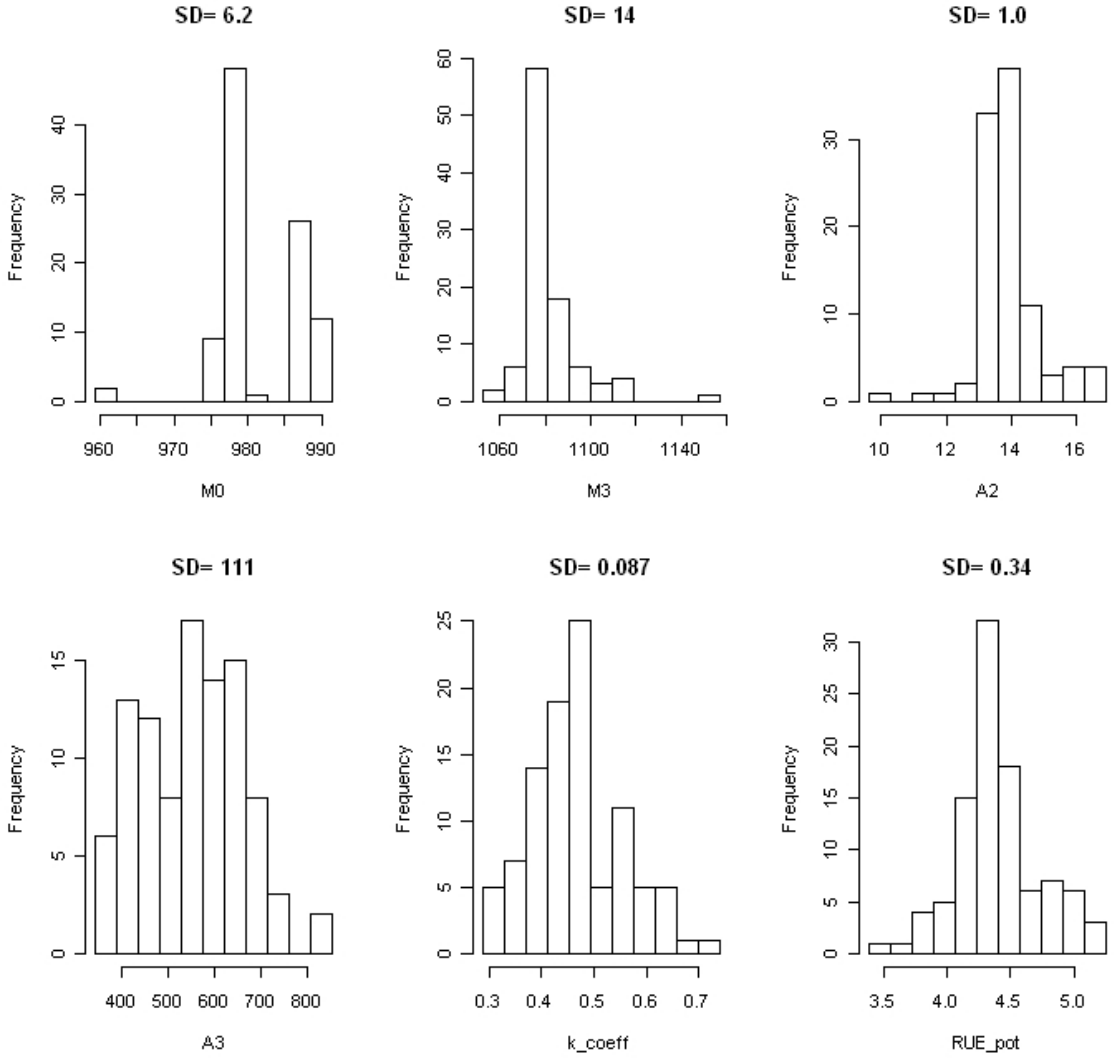

Fig. 5.2: Parameter estimates distributions and their standard deviations for six SUNFLO parameters based on 40 scenarios

To compare with parameter estimation in MSPE, four parameters are estimated with 40 and 100 scenarios. Fig. 5.3 is the example of four parameters' distributions based on $m = 40$ scenarios.

The comparison of standard deviations with MSPE based on m equal to 40 and 200 scenarios, and MSPEE based on m equal to 40 scenarios is illustrated in Fig. 5.4. MSPEE improves significantly the accuracy of parameters by reducing estimates' variance and uncertainty, with less scenarios used for multi-scenario parameter estimation.

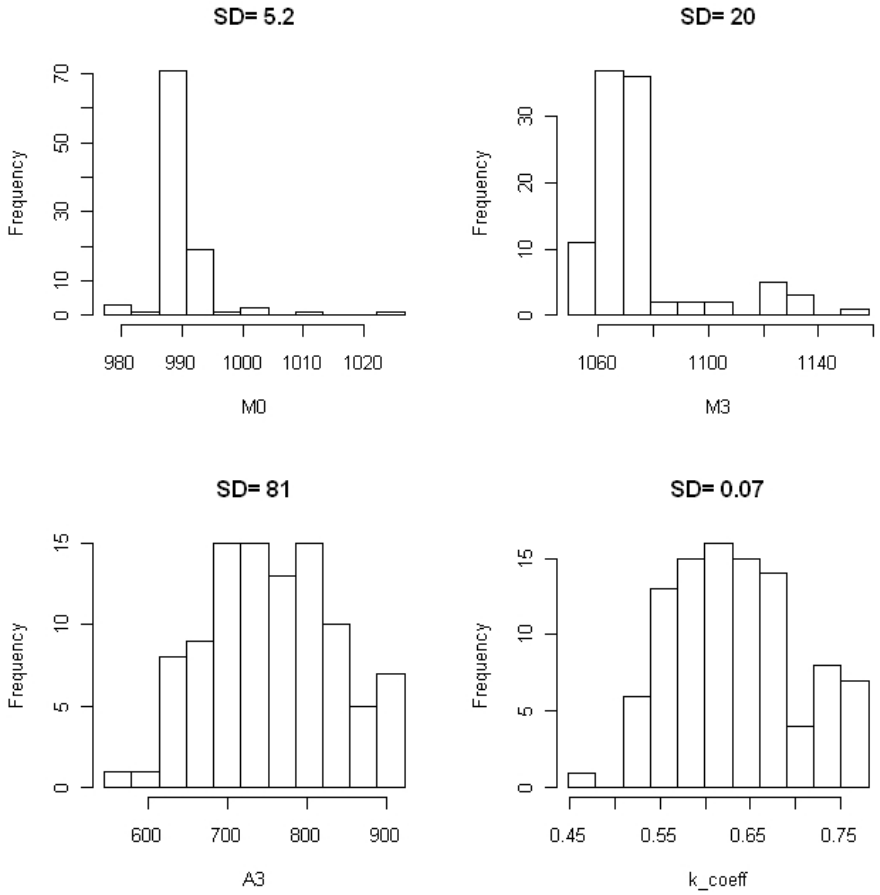

Fig. 5.3: MSPEE parameters' distributions and their standard deviations for four SUNFLO parameters based on 40 scenarios

5.3.3 Improvement on MSPE Prediction Ability

A similar cross-validation test as described in section 4.5.2 is carried out. The method's difference with the test in MSPE is still how to choose scenarios for estimation. In MSPE, scenarios are chosen randomly, while here in MSPEE it is based on data clustering and it picks scenarios with longest distance to each other. The set of scenarios for validation and prediction test is selected with the same cross-validation strategy as in MSPE.

Fig. 5.5 shows the model prediction error based on the estimation of four parameters from 40 scenarios and 100 scenarios with the MSPEE strategy compared with the

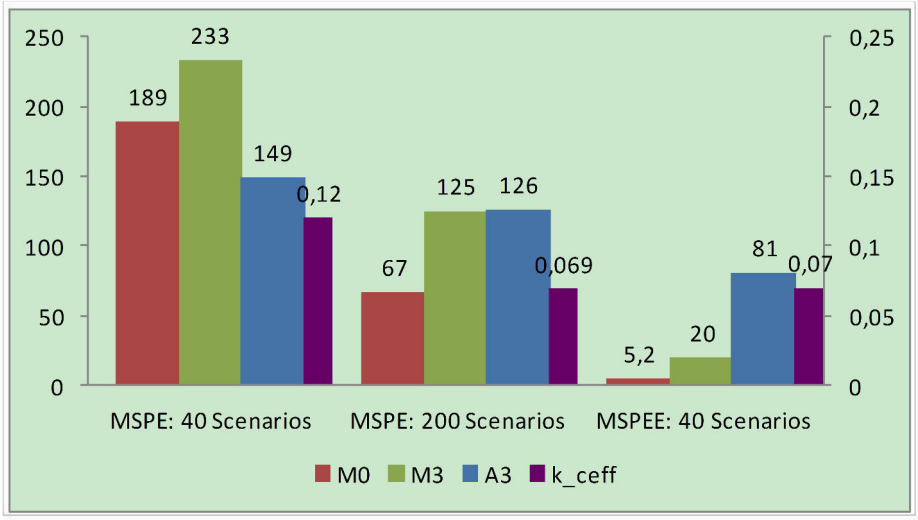

Fig. 5.4: The comparison between MSPEE parameters' standard deviations (40 scenarios) and MSPE parameters standard deviation (40 scenarios and 200 scenarios)

prediction errors in MSPE. It is demonstrated that both in MSPE and MSPEE, larger number of scenarios improves the model predictive capacity. In MSPEE, the prediction error decrease from around 8% of yield from the test of 40 scenarios to around 7% of yield from the test of 100 scenarios. However, the results are not striking since the he prediction ability converges to a limit, which is nearly reached with 40 scenarios. The MSPEE improves the parameter estimation efficiency in terms of using less scenarios for estimation while achieving similar prediction ability.

5.4 Conclusion

Environmental scenarios are clustered based on data clustering methods. The clustering results provide suggestions for multi-environmental trials in order to select the optimal experimental locations. Three clustering strategies, respectively based on weather and soil information, crop performance information, and the combination of both information, are tested and their clustering results are discussed. Another important usage of environmental clustering is to improve the efficiency of multi-scenario parameter estimation methodology. By categorizing scenarios in order to increase the

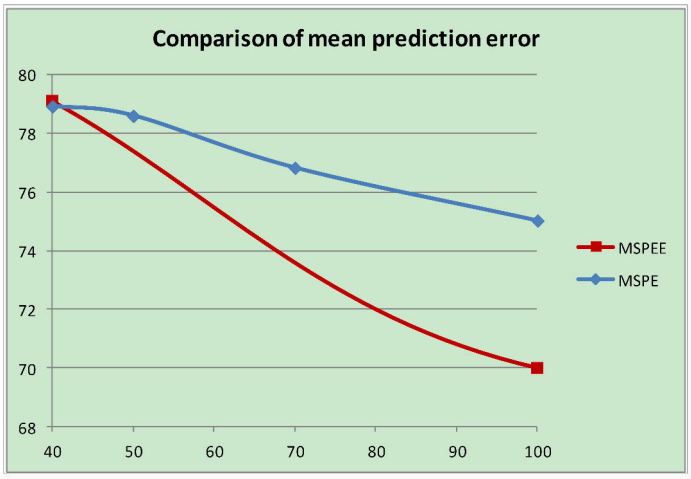

Fig. 5.5: Model prediction error based on the estimation of 4 parameters from 40 and 100 scenarios with MSPEE methods, compared with prediction error in MSPE methods. Y-axis is the root mean square error of prediction with the same unit as yield; X-axis is the the number of scenarios used for multi-scenario parameter estimation methods.

level of discriminative information for model inversion, we derived the adapted version of MSPE, named scenario selection based multi-scenario parameter estimation (MSPEE). It is demonstrated to improve the identifiability, estimation accuracy and prediction ability of MSPE.

Part III

APPLICATIONS AND DISCUSSION

6. MODEL SIMULATIONS OVER LARGE GEOGRAPHICAL AREAS AND TIME SCALES

Jones et al. [2006] state that the four most important applications of crop model are *(i)* prediction, *(ii)* determination of optimal management, *(iii)* large spatial-scale applications, and *(iv)* characterization of plant varieties and plant breeding. The application presented in this chapter deals with irrigation problem, therefore involving all the four aspects.

A non-negligible and increasing amount of water is unsustainable. As the majority of water use is dedicated to agriculture, optimizing irrigation strategies plays a key role in water sustainability. This chapter presents a tool to simulate irrigation demand of sunflower crop for large-scale geographic areas and time scales. Simulations were carried out on 20 genotypes of sunflower, 25 European farming regions and a time span from year 1951 to 2100. These results provide insights into the impact of the choices of farming regions and crop genotypes on irrigation. They also bring to a deeper understanding of irrigation demand evolution for researching future water management scenarios. We propose a way of optimizing irrigation water use by selecting crop genotypes and farming regions, whilst taking harvest yield constraints into consideration.

6.1 Context

Water is essential for growing food; for household water uses, including drinking, cooking, and sanitation; as a critical input into industry; for tourism and cultural purposes; and for its role in sustaining the earth's ecosystems. The supply of water, an essential resource for agriculture and industry, is under threat [Rosegrant et al., 2002]. Towards 35% of human water use is unsustainable, drawing on diminishing aquifers and reducing the flows of major rivers: this percentage is likely to increase if climate change impacts become more severe, populations increase, aquifers become progressively depleted and supplies become polluted and unsanitary [Clarke and King, 2004]. Water security and food security are inextricably linked. In the 1990s, it was estimated that humans were using 40-50% of the globally available freshwater in the approximate proportion of 70% for agriculture, 22% for industry, and 8% for domestic purposes with total use progressively increasing [Shiklamov and Baser, 1998]. Agricultural behaviors have significant impacts on the global water cycle, especially irrigation, accounting, for example, for about 80 percent of global and 86 percent of developing country water consumption in 1995 [Rosegrant et al., 2002]. Population and income growth will boost demand for irrigation water to meet food production requirements, household and industrial water demand. From 1961 to 2001 water demand doubled - agricultural use increased by 75% [Millennium Ecosystem Assessment, 2005]. By 2025, global population will likely increase to 7.9 billion. In response to population growth and rising incomes, calorie requirements and dietary trends will translate to even greater water demand if the food produced is to supply adequate nutrition [Rosegrant et al., 2002]. Therefore, the study of irrigation optimization significantly influences water sustainability and agricultural sustainability.

This chapter is concerned with two questions related to irrigation optimization. The first is to produce knowledge of irrigation demand under various scenarios. Water

management relies on reasonable information on water availability as well as on water demands by different sectors. Estimation of irrigation demand at large scale is therefore a key need for more precise water management. The second problem is to optimize irrigation demand under the requirement of a satisfactory harvest. Besides irrigation techniques, two determinants significantly affect irrigation demand: these are drought tolerance capacity of genotypes and environmental conditions of farming regions. The usage of irrigation could be significantly reduced by appropriate selections of crop genotypes and farming locations. To address the two questions, our investigation relies on a spatially distributed modeling of crop growth and water balance. We propose a tool to simulate irrigation demand and carry out a preliminary study on large scenarios. These simulations thus provide affluent knowledge of geographical, genotypic and time influences on simulated irrigation demand, which provides materials for water sustainability and sustainable agricultural study. In detail, the research was carried out on large data set of scenarios on 20 genotypes of a crop Sunflower, 25 farming regions in Europe, and weather information across 150 years from 1951 to 2100. These simulation results are useful for the study of irrigation optimization and enable us to analyze advantages of specific genotype and farming region, in order to give recommendations in term of irrigation saving and strong harvest yields for irrigation water management and agricultural strategy decisions. An interesting contribution of this research is the effect of crop genotype diversity on irrigation demand, whereas previous studies tend to focus on irrigation intensity of various farming regions. Due to the large variation of irrigation demand over different genotypes, research on irrigation optimization needs to consider both genotype and farming region diversity.

6.2 Irrigation Demand Simulation

The sunflower model SUNFLO (Chapter 3.2) is the core to simulate sunflowers' growth, with considering environmental impacts and genotype diversity. It models plant photosynthesis, morphogenesis, biomass production, and biomass distribution. Thanks to its water budget module, SUNFLO is able to simulate the sensitive influences of water deficit scenarios on plant growth. Its water cycle is mainly co-functioned by root water absorption and transpiration from the plant side, and rain, irrigation, and evaporation from the environment side (Fig. 3.5). Fractional soil water index ($FTSW$) represents the crop water stress. Its value ranges from 0 to 1. The bigger the value is, the more water deficiency the crop has. Depending on genotypes and plant functions, critical values RT and RO determine the plant drought tolerance. For example, for a sunflower genotype "Albena", the critical value of radiation usage efficiency is 0.32. When $FTSW$ is below 0.32, the radiation usage is badly affected. When FTSW indicates that crops will be under water deficiency, irrigation is supplied to give the plant minimum water with keeping normal plant development. Fig. 6.1 illustrates simulated $FTSW$ differences between the case with irrigation and without irrigation. Clearly the irrigation case improved $FTSW$ and relieved water stress.

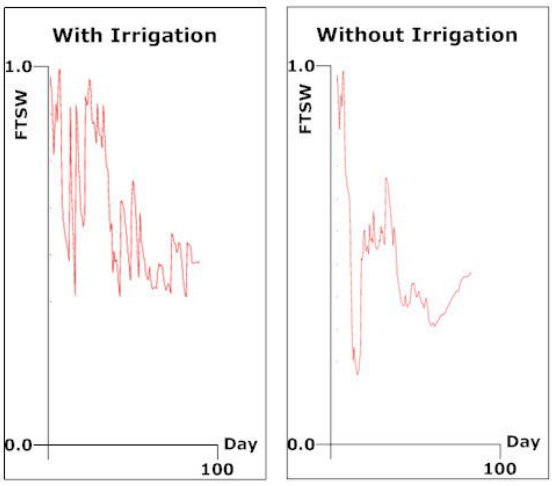

Fig. 6.1: (a) Left picture: FTSW value in plant growth period in the scenario with irrigation; (b) Right picture: without irrigation

The irrigation information is integrated to analyze agricultural water use and agricultural impacts on water resources. The simulations of irrigation amount and crop harvest are paid the most attention. Executing plenty of simulations produces data for researching irrigation evolution, and genotype and region selection in term of water saving optimization, under conditions of multiple genotypes, farming locations, and environment scenarios.

6.3 Simulation Experiments

Three series of data are utilized, including a European irrigation map, an environmental information database, and genotypic parameters. The European irrigation map supplies knowledge for location selections in our experiments. The majority of irrigated areas are concentrated in the Mediterranean region. France, Greece, Italy, Portugal, and Spain account for 12 million ha corresponding to 75% of the total area equipped for irrigation in EU [Gunter et al., 2008]. From a global irrigation map, named GMIA [Siebert et al., 2007], 25 regions with different irrigation density in above 5 countries are chosen (see Fig. 6.2).

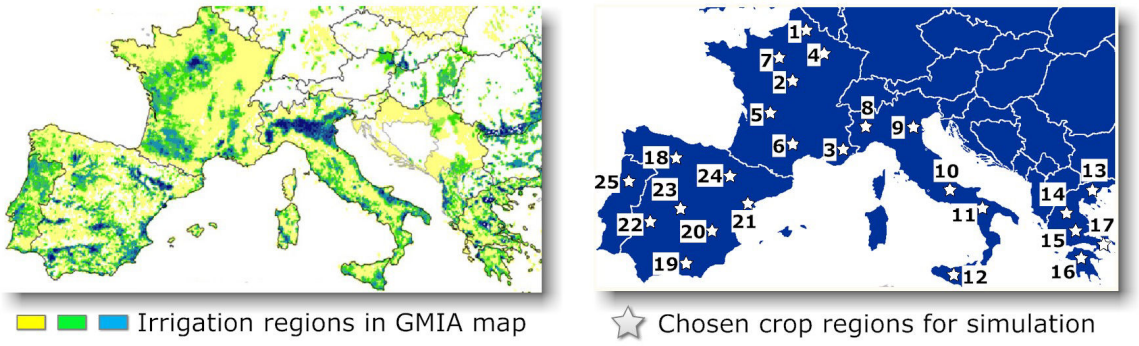

Fig. 6.2: (a) Left: irrigation regions in GMIA map (yellow to blue: more irrigation); (b) Right: 25 farming regions (stars) in simulation

The environmental information used as a simulation input comes from ENSEMBLES dataset (detailed in section 2.3.2). There are environmental data for the above Eu-

ropean locations, including temperature, radiation, precipitation, evaporation. Data from 1951 to 2011 are used for past irrigation examination, and those from 2011 until 2100 are used for irrigation evolution prediction, based on climatic data simulation. Genotypic parameters, representing sunflower genotypic diversity and diverse interactions with environment, are measured parameters of botanical experiments by SupAgro Montpellier [Lecoeur et al., 2011]. 20 sunflower genotypes are concerned (see table 6.1). The tool Pygmalion is used to simulate irrigation demand and plant growth features for multiple scenarios. It can compare irrigation and harvest for chosen scenarios and recommend a corresponding genotype and region. It also produces harvest and irrigation evolution graphs.

Tab. 6.1: Names of 20 sunflower genotypes concerned for irrigation demand simulations.

Peredovik	INRA6501	Remil	Airelle	Relax
Mirasol	Primasol	Cargisol	Viki	Frankasol
Albena	Vidoc	Euroflor	Santiago	DK3790
Prodisol	Melody	LG5660	Allstar	Heliasol

6.4 Results

Using the above data and our simulator, we produced a collection of irrigation demand simulations. Our irrigation demand simulations qualitatively agree with real observations. In Fig. 6.3, two irrigation maps are contrasted. The left is an irrigation map GMIA. The right is our irrigation demand simulation. They have a similar pattern of irrigation demand differences between regions. This result provides support for the reliability of our irrigation simulation tool.

The evolutions of irrigation demand and harvest amounts have been produced for 1951 to 2100. An example illustrated in Fig. 6.4 is the harvest and irrigation evolution of all genotypes from 1990 to 2100. The genotype “Melody” offers the biggest harvest in the future, but its irrigation amount is also high. There are two levels of irrigation

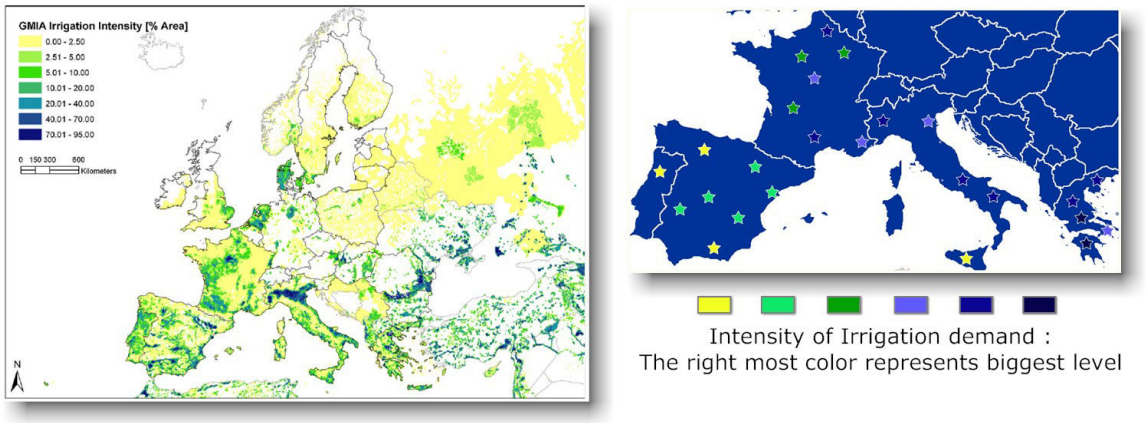

Fig. 6.3: Contrast of irrigation demand between GMIA map and simulations; (a) Left: GMIA map; (b) Right: simulation map

demand. Most genotypes are in the smaller one. So it's possible to find a genotype with the small level irrigation, but with a comparably big harvest. The genotype "Euroflor" fits the standard. It has the second largest harvest. Although this amount is less than the one from "Melody", its large irrigation saving could make up for the loss. Especially in drought regions, it could be a good selection for water sustainability reasons. Generally speaking, the harvest evolution is decreasing, and the irrigation evolution has a slight tendency to increase in simulations. This is not in agreement with reality as the sunflower harvest has increased in the last few decades, because the simulation ignores genotype evolution and other positive factors. This statement of harvest evolution is made by only considering climate change in long time range. It predicts a negative influence of future climate in both harvest and irrigation. While harvest has been widely recognized to face potential declines because of water shortage and potential increase of farming land, this conclusion puts more pressure. Our irrigation optimisation strategy on appropriate genotype and farming region selection is one way to mitigate it.

Comparisons are carried out among genotypes for particular regions. For example, for two farming regions location 3 in France and location 16 in Greece, the total irrigation demand and harvest of 20 genotypes is shown in Fig. 6.5.

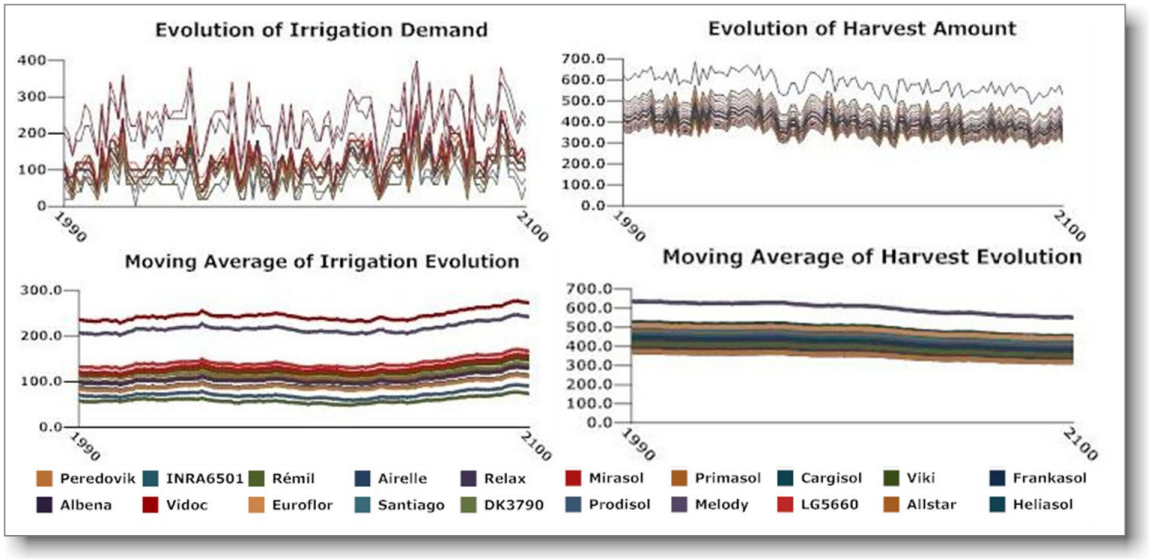

Fig. 6.4: Evolution graphs for 20 sunflower genotypes from 1990 to 2100 and their moving average; (a) Top left: irrigation demand evolution; (b) Top right: harvest evolution; (c) Bottom left: irrigation moving average; (d) Bottom right: harvest moving average

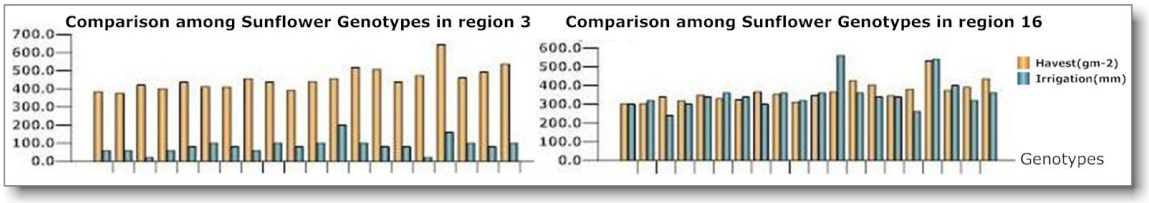

Fig. 6.5: Harvest and irrigation comparisons; Horizontal axis is ordered genotypes in table 6.1; (a) Left: region 3; (b) Right: region 16

For location 3, the genotype "Melody" has a distinctly higher harvest and a slight higher irrigation demand than the others. Therefore it is recommended for this region. For location 16, the biggest harvest genotype is still "Melody", but it requires substantial irrigation. Searching for a genotype that has smaller irrigation demand will result in a reduced harvest. The genotype "Heliasol" has the second largest harvest with a clear decrease of irrigation. This genotype is preferred considering the drought condition in this area. Comparisons are also made among regions for particular genotypes. For example, Fig. 6.6 illustrates the irrigation demand and harvest for genotype "Remil" and "Melody" in 25 farming region.

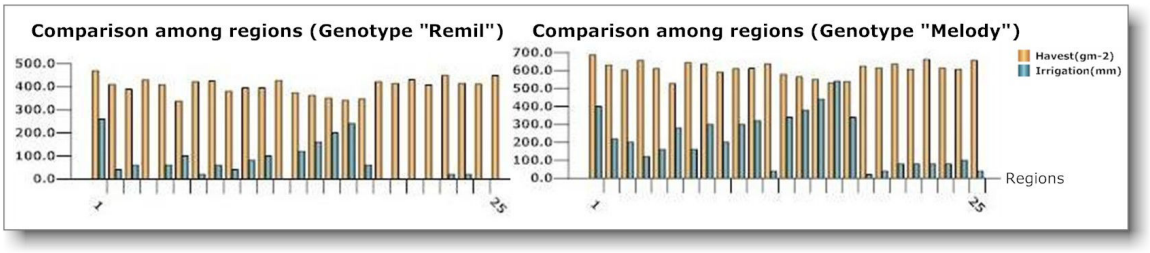

Fig. 6.6: Harvest and Irrigation for sunflower genotypes in 25 regions; (a) Left: genotype "Remil"; (b) Right: genotypes "Melody"

Irrigation amounts vary widely among regions, whilst harvest levels are sorted in little various levels. The best farming region for "Melody" is location 19 in Portugal, because it has a low irrigation demand and a comparably good harvest. Compared with location 16 in Greece, its irrigation amount is 10 times smaller, and its harvest is bigger. Compared with location 1 in France, its harvest is a slightly smaller, but its irrigation demand costs 10 times less. Selections require a compromise between irrigation and harvest, and depend on the project objectives and actual situation, such as the irrigation capacity.

6.5 Discussion

In this chapter it is proposed to use crop genotype selection and farming region control for irrigation optimization and improving water sustainability. A tool is developed for offering irrigation demand and harvest amount information for three dimensional scenarios: diverse genotypes, multiple farming regions, and a long timeframe. A preliminary study and analysis are carried out on a crop sunflower with 20 genotypes, 25 European farming regions, and over 150 years, in order to demonstrate the proposition and to test the tool. This study results in a methodology usable in future research, to study interactions between irrigation, crop genotypes and the environment.

These results are very encouraging. Firstly, this kind of simulation offers a large number of characteristic features for various contrasting scenarios. Moreover, for

diverse genotypes and farming regions, the simulation results predict a variety of distinct irrigation demands and harvest amounts. This indicates that appropriately selecting a combination of these parameters can result in improved results. Secondly, these simulations qualitatively produce a good fit to real data. This suggests that it is reliable to use simulated information and that we may have confidence in our analysis. Lastly, rules to achieve optimal results are explored. Particular analyses proved that proper selection of the genotype and farming region may save a considerable amount of water for irrigation. For concrete policy decisions, selection rules integrate the project objective, the yield requirements, the irrigation budget, and the technological level available in the target location. Using the dynamic system formulation of the plant growth model, the tool develops numerical optimisation techniques to determine multi-constraint optimal farming strategies [Qi et al., 2010; Wu et al., 2005].

This simulation tool can be applied easily to irrigation demand problems for other crops in other scenarios. An effective water sustainability management requires global crop and farming region control. The water cycle modeling used in the tool simulates the interactions among environment, plants, and human operations. Although the tool produces reasonable simulation results, it can still be improved. The quantity of irrigation and harvest amount forecast by our simulations should not be considered as scientifically proven at this stage. More real data for sunflower harvest and irrigations need to be gathered for model calibration. Moreover, the modeling could consider more factors such as irrigation techniques and management.

7. DISCUSSION

One of the challenges of modern plant breeding is to provide genetic solutions to increase plant productivity. A breeding program can be considered as the process of developing improved cultivars by manipulating available genetic variability to create new allelic combinations best adapted to target environments and applications [Messina et al., 2006]. Traditionally, breeding is based only on phenotypic observations, which makes the work costly, long, and highly based on breeder's experience. This is particularly true when breeding populations and cultivars are characterized by high genetic diversity and substantial genotype × environment interactions: breeding programs in that case require precise and efficient phenotyping [Walter et al., 2012]. Most current efforts therefore focus on developping sophisticated high-throughput phenotyping equipements. But genotypic information is also a key point. As stated in [Messina et al., 2006], the breeding of higher-yielding crop plants would be greatly accelerated if the phenotypic consequences of changes at some genetic markers of an organism could be reliably predicted. Plant growth models, which aim to simulate the genotype × environment interactions in order to predict the corresponding phenotypes, naturally appear as relevant tools to advance the analysis of phenotypes and breeding strategy. They can be used to assist genetic improvement in four main ways: environmental characterization for testing genotypes, assessment of specific putative traits for designing improved plant types, analysis of responses of probe genotypes for improved interpretation of multi-environment trials, and optimizing combinations of genotype and management for target environment [Messina et al., 2006].

In this context, this thesis addresses the development of plant growth models and

model analysis methodologies in order to facilitate phenotype analysis and breeding strategies. Four objectives were mentionned in the introduction 1.4: each of them will be discussed on the basis of our results and perspectives will be evoked.

Modeling. We used two ecophysiological crop models, SUNFLO for sunflower (*Helianthus annuus* L.) and CORNFLO for corn (*Zea mays* L.). They simulate plant phenology, morphogenesis, photosynthesis, biomass production and biomass distribution under temperature and drought stress. Model parameters have biological meaning and are designed to be grounded potentially in gene-level understanding. Since they had been previously validated on several genotypes and different conditions, we mainly used them in this thesis for our tests and applications. One of our contributions consists also in the analysis of these models (section 4.3.1).

In an attempt to build a more mechanistic model on the basis of SUNFLO, an original sunflower model, SUNLAB, has been developed: it mainly improves the biomass allocation process, by adopting a classical source and sink approach. SUNLAB is a joint concept of SUNFLO and GREENLAB. While it is hard to find a balance for a model design, SUNLAB model is an interesting trial. It produces more details on organs structure and mass than SUNFLO. It models ecophysiological functions of photosynthesis and morphogenesis to ensure a more accurate and a better representation of crop physiology for biomass production than GREENLAB. SUNLAB proves that this combination of concepts is effective. The parameters of four sunflower genotypes were estimated in SUNLAB based on two years experimental data, including one with drought stress. It helped to explain the internal competition for biomass by simulating organ biomass distribution. The parameter estimation procedure, benefiting from both direct measurements and model inversion strategies, preserves a good capacity for genotypic differentiation and is provides robust results on multiple phenotypic traits.

For testing of predictive capacity, SUNLAB outputs were confronted to an additional

experimental data set. However, this remains a weak point of our study since the validation dataset is clearly not independent of the training set: more diverse datasets are expected for further exploration of this predictive aspect.

SUNLAB is able to simulate specific leaf area (SLA). SLA is an important variable in plant growth modeling since it usually determines blade surface area values based on blade biomass for further simulation loops, such as for instance in GREENLAB [Christophe et al., 2008]. SLA is usually considered constant in those models. In reality, SLA varies according to genotypes, leaf ranks and leaf growing periods, as it has been observed for instance for the SLA variations of wheat [Rawson et al., 1987]. For sunflowers, the variations of SLA and the factors influencing them are still poorly known. Accurate estimation of SLA is mentioned as a major source of error in models and implies difficulties in obtaining a reliable computation of leaf area index, which is the main component of biomass production modules [Heuvelink, 1999; Marcelis et al., 1998]. As SUNLAB simulation outputs include, independentely, individual blade mass and blade areas, variations of SLA at every simulation steps can be produced. However, these preliminary results have to be considered with caution since the current SUNLAB parameters were estimated using reconstructed data of individual blade mass (that were not measured individually). Besides, this feature would need to be reconsidered when one aims at taking into account a feedback effect of biomass allocation on production: in the current model, blade mass do not play a role in the determination of blade area and therefore do not have effect on the produced biomass.

This raises interesting questions in terms of how mechanistic a model should be, in our context of phenotype analysis to assist breeding programs. With more mechanistic models generally come hidden parameters that cannot be experimentally measured, because feedback effects can produce emergent properties that can be difficult to disantangle *a posteriori* from the resulting phenotype. Therefore, modeling crop growth

through empirical experimental analysis and direct parameter measurements has the advantage that parameters have a biological meaning but hampers the consideration of complex mechanisms or internal regulations. For instance, biomass partitioning was not modeled in SUNFLO because of the heavy experiments and the difficulty to understand the organs interaction, while it could be done when introducing hidden parameters, sinks, that cannot be measured but had to be estimated relying on optimization algorithms, as done in SUNLAB. Hidden parameters of mechanistic models, that can simulate the internal processes regulating plant growth, are more likely to be genetically determined (or, at least, stable under varying environmental conditions) than directly measured parameters that can be strongly influenced by the environment. They could therefore offer more potentials on the development of genotype-to-phenotype predictive models. However, practical considerations should also be examined in our context of model application, *i.e.* transferring model-based informations to breeders. This kind of information could be for instance recommendations on optimal environmental conditions or management practices for a given genotype; or identification of particular features (a subset of the model parameters, for instance) to focus on in the breeding process in order to create variants with some targeted traits. A highly mechanistic and complex model whose parameters cannot be observed might appear as a 'black box' whose results would not be easily trusted by end users that have not participated to its development. It also implies that, because of their interactions, parameters cannot be obtained independentely from each other: the whole estimation process needs to be performed on all the data at the same time (it is not possible to optimize sequentially on data for different types of organs, for instance). Therefore, there is a balance to find between an empirical model that would be easily applied but with limited ability to represent the plant internal regulations and a fully mechanistic model that would reveal too complex to be of practical use.

Parameter estimation. Multi-scenario Parameter Estimation Methodology (MSPE) is designed to overcome the problem of difficult model inversion coming from insufficient complex experimental data. It is able to deal with limited or aggregated kinds of data as soon as they are collected under a large amount of diverse scenarios. Practical issues for carrying out MSPE are discussed, including setting priorities on parameters to be estimated, optimization and computation solutions, and parameter identifiability. The hypothesis that "the increase of scenarios amount makes estimated parameters possessing better estimation accuracy and better prediction ability" is explored by cross-validation tests. However, the convergence of these estimation accuracy and prediction ability could not be fully tested, because of its heavy computational requirements.

The MSPE methodology relies on the level of information available from the different scenarios used for model inversion: the more different and diverse the environmental scenarios are, the more robust model inversion will be. This issue on information level was central in several points of the thesis: the issue of non-estimability / practical non-identifiability (see section 4.3.3), the convergence of parameter estimation with the number of scenarios available (see section 4.4) and the optimal choice of environmental scenarios based on clustering techniques (see section 5.3). However, the approach developed in this PhD regarding the concept of 'information' was purely empirical: numerical tests were used to illustrate the behavior of the methodology in virtual or real test cases. An important perspective of this research work would thus be to make stronger links with the statistical information theory, in order to determine *a priori* results on convergence, error bounds or uncertainty estimation based on a theoretical analysis of the system and data.

An interval estimator MSPEJ was developed, based on the delete-m Jackknife method: m scenarios are taken from n scenarios to produce ${}_nC_m$ combinatorial parameter distribution samples. The methodology was tested on SUNFLO parameters. A perspective

of this work would be to study the estimation and prediction features of MSPEJ, particularly the effect of n and m on estimation accuracy and prediction ability. Jackknife methods are generally used to test the bias and variance of some statistics, as done in the estimation and prediction tests of MSPE. Compared with single estimation, Jackknife methods have been shown to reduce bias for some statistics: it would be interesting to investigate that point in our application. Identifiability issues could also be compared with MSPEJ and with MSPE.

Values, not converge

Choice of environmental protocols. Environmental scenarios were clustered by hierarchical and centroid-based clustering analysis based on weather and soil information including temperature, radiation, precipitation, and evapo-transpirational reference, and its influence on plant growth features, such as crop yield. For experimental design, selecting one representative scenario from each cluster can help deciding necessary trials in field experiments. MSPEE, a multi-scenario parameter estimation methodology based on environment clustering and scenario selection, improved practical identifiability of parameters in comparison with the basic MSPE. It also improved the efficiency of MSPE, in term of utilizing fewer scenarios for getting the same prediction ability and variance of parameter distribution. However, for reason of the heavy computation cost of the cross-validation method used in MSPEE tests, only few data points have been calculated to support our conclusion. Calculations on more data points should be carried out.

Applications of phenotype analysis based on crop models. SUNFLO was used to simulate phenotypic traits of 20 sunflower genotypes over large geographical areas (25 locations) and time scales (150 years). In particular, it predicts irrigation demand of different genotypes and potential yields under varied scenarios. This illustrates several aspects of future model use.

A final perspective to discuss concerns the study of linking crop model parameters to

genetic information: the application was planned but was regretfully not completed in this thesis. Models in agriculture systems are characterized by having many organizational levels. From the individual components within a single plant or animal cell, through constituent plants or animals to farms or a whole agricultural region or nation, and finally to the world agricultural economy, lies a whole range of agricultural systems [Cheeroo-Nayamuth, 1999]. Organ-level plant growth models can help 'navigate a path through this complexity'. They provide means to link phenotypic consequence to changes in genomic regions via stable associations with model coefficients. If they can capture the system dynamics and much of the fine detail is not directly required, robust coarse-grained models might be the tool needed to integrate phenotypic and molecular approaches to plant breeding [Hammer et al., 2006]. Recently, quantitative trait loci (QTL) information has been incorporated into some organ-level crop models. For example, to connect model coefficients to genomic regions (or genes), Reymond et al. [2003] dissected the parameters of a model of maize (*Zea mays*) leaf elongation rate into effects of quantitative trait locus. Yin et al. [2006] has identified a few quantitative trait locus to model-input traits in the model of predicting spring barley (*Hordeum vulgare L.*) flowering time. To address the link between model parameters and QTL, well designed models and suitable experimental data are required. Appropriate model structures allow sufficient physiological feedback features to be incorporated. Model input parameters should be designed to be grounded potentially in gene-level understanding [Yin et al., 2004]. It always requires the plant growth model parameters having biological meaning to represent genetic coefficients [Tardieu, 2003; Yin and Struik, 2010]. The organ-level model SUNFLO and its parameters meet the requirements. The experimental database of 90 sunflower genotypes (section 2.1.2) is a good dataset to study the subject. These 90 genotypes are F1 hybrid of the first filial generation resulting from a cross mating of 9×10 distinctly different parental types. A main obstacle to this study was coming from

the insufficiency of the collected experimental traits in this dataset for a proper parametric estimation: this problem is now solved by our MSPEJ parameter estimation methodology. It was tested on one genotype, “Melody”, among the 90 genotypes: the results were compared to those obtained by direct measurements, in order to validate the parameter estimation methodology. Since it was proved to be a successful attempt, this thesis paves the way to this promising study. The next steps would be to estimate SUNFLO parameters for the 90 genotypes, then perform statistical analyses to study the correlations between parameters of different genotypes. Similarities or differences in parameter values could reveal genetic links.

To sum up, this thesis produced promising results on crop modeling and crop model's multi-scenario parameter estimation. Further studies on these perspectives will boost the development of phenotype analysis' tools to move towards cheaper, faster, and more efficient breeding processes.

LIST OF TABLES

3.1 SUNFLO model: non-genotypic parameters 59
3.2 SUNFLO model: genotypic parameters 60
3.3 Main SUNFLO inherited parameters values. 67
3.4 Sensitivity analysis of SUNLAB parameters: first-order indices of the most influential parameters (with index > 1%). 70
3.5 Estimated parameter values of SUNLAB for four genotypes. 71

4.1 "Proof of concept" test on 2 parameters 91
4.2 10 significant parameters for SUNFLO model and the values for the virtual genotype used in statistical test 91
4.3 An example of the statistical test results ($sigma = 0.05$, 100 samples) . 92
4.4 SRC sensitivity index of the most influential parameters (with index > 1%). 100
4.5 Sobol first order and total order index values for the most influential parameters (with index > 1%). 100
4.6 Parameter optimization searching value range 106
4.7 Estimated parameters values from 50 scenarios: parameters reaching boundary 106
4.8 Estimated parameters values from 50 scenarios: parameters not reaching boundary 106
4.9 Estimated parameters values from 500 and 720 scenarios. 107
4.10 Measured value of two parameters, their estimated values in 20 samples and the calculation of mean and standard error by MSPEJ. 123
4.11 Measured value of three parameters, their estimated values in 20 samples and the calculation of mean and standard error by MSPEJ. 123
4.12 Measured value of four parameters, their estimated values in 15 samples and the calculation of mean and standard error by MSPEJ. 123
4.13 Measured value of ten parameters, the calculation of their estimation values' mean and standard error by MSPEJ. 124

5.1 Parameters estimated from scenarios with high diversity, selected from 40 different environment clusters. 134
5.2 Parameters estimated from scenarios with high similarity, selected from one single environment cluster (of correlation level >0.91). 134

6.1 Names of 20 sunflower genotypes concerned for irrigation demand simulations. 148

LIST OF FIGURES

2.1 $FTSW$ for three datasets "2001", "2002a", "2002b" 31

3.1 Flowchart of plant growth modeling. 40

3.2 Leaf area expansion curve (GRe) for leaf ranks 1, 10, 20 of genotypes "Melody" and "Albena". 48

3.3 Potential radiation usage efficiency(Ebp) for genotypes "Melody" and "Albena". 50

3.4 Thermal factor for sunflower genotypes. 51

3.5 Water budget module in SUNLAB: (a) Left: processes considered in the water cycle model; (b) Right: the three soil layers C1, C2 and C3. . 53

3.6 Experimental data (dots) and simulation (lines) comparisons of blade dry mass, internode dry mass, and capitulum dry mass for the four genotypes - "Albena", "Melody", "Heliasol", and "Prodisol" - and for dataset "2001"(blue) 72

3.7 Sink ability based on SUNLAB parameters and sink competition theory in biomass distribution module 73

3.8 Graphs A to D: Experimental data (dots) and simulation (lines) comparisons for the "2001" (blue) and "2002a" (red) conditions of the radiation interception efficiency $RIE(d)$, total blade area $AA(d)$, leaf number $N(d)$, accumulated above-ground dry biomass $CDM(d)$ and biomass compartments (capitulum, blades, petioles, internodes) for the "Melody" genotype. Graphs E and F: biomass compartments of "Prodisol" and individual leaf mass profile for "Heliasol". 74

3.9 Model validation for genotype "Albena" using an additional experimental dataset: "2002b" 75

3.10 Comparison of simulation and field data for individual blade area and biomass of genotype "Melody"; the right graph is the simulation of specific leaf area for the four genotypes 76

4.1 Illustration of the idea of MSPE methodology 85

4.2 Flowchart of the general strategy for the "proof of concept" test. Produce M trait data with P_1: named $(Trait_1, \cdots, Trait_M)$. From N climatic scenarios, we get the observations set $(Obs_1, \cdots, Obs_N)$, which are used to calibrate model parameters, starting the algorithm from an initial value P_2. The resulting parameter vector estimate is P_3. If P_3 is equal to P_1, N scenarios are sufficient to estimate model parameters. . 88

4.3 Flowchart of the general strategy for the statistical test. 1) Generate virtual experimental data $(Obs_1, \cdots, Obs_N)$ of plant growth traits $(Trait_1, \cdots, Trait_M)$ based on parameter vector P_0. 2) Give random perturbations on $(Obs_1, \cdots, Obs_N)$ to produce $(SampleObs_1, \cdots, SampleObs_N)$. Repeat to generate K samples of perturbed observations. Each sample is used to produce an estimate, resulting in $P_1 \ldots P_K$ and their mean value and variance. 90

4.4 Continuity of $MSgraine$ with respect to each parameter in CORNFLO model 95

4.5 Continuity of parameter *phyllo_de_ini* in CORNFLO model based on multiple scenarios 96

4.6 Convexity test for 2, 4, 50 scenarios of the parameter *phyllo_de_ini* . . 96

4.7 Distribution of 100 samples for four CORNFLO parameters based on 40 scenarios 110

4.8 Distribution of 100 samples for four CORNFLO parameters based on 70 scenarios 111

4.9 Distribution of 100 samples for four CORNFLO parameters based on 200 scenarios 112

4.10 The illustration of our hypothesis on the evolution of the standard deviation value 112

4.11 The illustration of standard deviation values of parameter distribution in MSPEJ 113

4.12 Prediction error (Root mean square error) changes along with the increase in the number of scenarios. In prediction error test 1, the optimization algorithm Gauss-newton method takes 0.9 times of *a priori* CORNFLO parameters as initial parameter values for optimization; Prediction error test 2 takes 1.1 times of those values 114

4.13 Prediction error is reduced with the number of scenarios, from 40 to 200. Left graph: the prediction error reduction in cross-validation test; Right graph: the comparison between prediction error in cross validation test(blue line) and in number of scenarios increase test (in section 4.5.1, red line). Y-axis is the root mean square error of prediction with the same unit as yield; X-axis is the the number of scenarios used for MSPE parameter estimation. 117

4.14 The distribution of all combinatorial samples for using 17 scenarios out of 21 scenarios in MSPEJ for SUNFLO Model 121

5.1 County scenario's clustering based on the combination of environment information and yield 131

5.2 Parameter estimates distributions and their standard deviations for six SUNFLO parameters based on 40 scenarios 136

5.3 MSPEE parameters' distributions and their standard deviations for four SUNFLO parameters based on 40 scenarios 137

5.4 The comparison between MSPEE parameters' standard deviations (40 scenarios) and MSPE parameters standard deviation (40 scenarios and 200 scenarios) 138

5.5 Model prediction error based on the estimation of 4 parameters from 40 and 100 scenarios with MSPEE methods, compared with prediction error in MSPE methods. Y-axis is the root mean square error of prediction with the same unit as yield; X-axis is the the number of scenarios used for multi-scenario parameter estimation methods. 139

6.1 (a) Left picture: FTSW value in plant growth period in the scenario with irrigation; (b) Right picture: without irrigation 146

6.2 (a) Left: irrigation regions in GMIA map (yellow to blue: more irrigation); (b) Right: 25 farming regions (stars) in simulation 147

6.3 Contrast of irrigation demand between GMIA map and simulations; (a) Left: GMIA map; (b) Right: simulation map 149

6.4 Evolution graphs for 20 sunflower genotypes from 1990 to 2100 and their moving average; (a) Top left: irrigation demand evolution; (b) Top right: harvest evolution; (c) Bottom left: irrigation moving average; (d) Bottom right: harvest moving average 150

6.5 Harvest and irrigation comparisons; Horizontal axis is ordered genotypes in table 6.1; (a) Left: region 3; (b) Right: region 16 150

6.6 Harvest and Irrigation for sunflower genotypes in 25 regions; (a) Left: genotype “Remil”; (b) Right: genotypes “Melody” 151

BIBLIOGRAPHY

Ahmad, S., Ahmad, R., Ashraf, M., Ashraf, M., and Waraich, E. (2009). Sunflower (helianthus annuus l.) response to drought stress at germination and seedling growth stages. *Pak J Bot*, 41(2):647–654.

Albuquerque, F. and Carvalho, N. (2003). Effect of type of environmental stress on the emergence of sunflower (helianthus annuus l.), soyabean (glycine max (l.) merril) and maize (zea mays l.) seeds with different levels of vigor. *Seed Sci Technol*, 31:465–467.

Allen, M., Prusinkiewicz, P., and Dejong, T. (2005). Using L-systems for modeling source-sink interactions, architecture and physiology of growing trees, the L-peach model. *New Phytologist*, 166:869–880.

Allinne, C., Maury, P., Srrafi, A., and Grieu, P. (2009). Genetic control of physiological traits associated to low temperature growth in sunflower under early sowing conditions. *Plant Science*, 177:349–359.

Ammari, H., Garnier, J., Jugnon, V., Kang, H., Lee, H., and M.Lim (2012). Enhancement of near-cloaking. part 3: Numerical simulations, statistical stability, and related questions. *Contemporary Mathematics*, 577:1–23.

Bertheloot, J., Cournède, P.-H., and Andrieu, B. (2011). Nema, a functional-structural model of n economy within wheat culms after flowering: I. model description. *Annals of Botany*, In press.

Brockington, N. (1979). Computer modelling in agriculture. *Oxford University press*.

Buck-Sorlin, G. H., Kniemeyer, O., and Kurth, W. (2005). Barley morphology, genetics and hormonal regulation of internode elongation modelled by a relational growth grammar. *New Phytologist*, 166(3):859–867.

Cacuci, D. G. (2003). Sensitivity and uncertainty analysis: Theory, volume i. *Chapman and Hall*.

Cacuci, D. G., Ionescu-Bujor, M., and Navon, M. (2005). Sensitivity and uncertainty analysis: Applications to large-scale systems (volume ii). *Chapman and Hall*.

Campolongo, F., Cariboni, J., and Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling and Software*, 22:1509–1518.

Casadebaig, P., Guilioni, L., Lecoeur, J., Christophe, A., Champolivier, L., and Debaeke, P. (2011). Sunflo, a model to simulate genotype-specific performance of the sunflower crop in contrasting environment. *Agricultral and Forest Meteorology*, 151:163–178.

Chapman, S., Cooper, M., Podlich, D., and Hammer, G. (2003). Evaluating plant breeding strategies by simulating gene action and dryland environment effects. *Agronomy Journal*, 95:99–113.

Cheeroo-Nayamuth, B. (1999). Crop modelling/simulation: an overview. *Annual Meeting of Agricultural Scientists*, pages 11–26.

Chen, Y. and Cournède, P.-H. (2012). Assessment of parameter uncertainty in plant growth model identification. In Kang, M., Dumont, Y., and Guo, Y., editors, *Plant growth Modeling, simulation, visualization and their Applications (PMA12)*. IEEE Computer Society (Los Alamitos, California).

Christophe, A., Letort, V., Hummel, I., Cournède, P.-H., de Reffye, P., and Lecoeur, J. (2008). A model-based analysis of the dynamics of carbon balance at the whole-plant level in arabidopsis thaliana. *Functional plant biology*, 35:1147–1162.

Christopoulos, A. and Michael, J. L. (2000). Beyond eyeballing: Fitting models to experimental data. *Critical Reviews in Biochemistry and Molecular Biology*, 35(5):359–391.

Clarke, R. and King, J. (2004). The atlas of water.mapping the world's most critical resource. *Earthscan*, 127:22–23.

Colson, J., Wallach, D., Bouniols, A., Denis, J.-B., and Jones, J. (1995). Mean squared error of yield prediction by soygro. *Agronomy Journal*, 87:397–402.

Comstock, R. (1977). Quantitative genetics and the design of breeding programs. *Proceedings of the international conference on quantitative genetics*, pages 705–518.

Cooper, M., Byth, D., and Delacy, I. (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.

Cournède, P.-H., Letort, V., Mathieu, A., Kang, M., Lemaire, S., Trevezas, S., Houllier, F., and de Reffye, P. (2011). Some parameter estimation issues in functional-structural plant modelling. *Mathematical Modeling of Natural Phenomena*, 6(2):133–159.

de Bruin, H. A. R., Trigo, I. F., Jitan, M. A., Enku, N. T., van der Tol, C., and Gieske, A. S. M. (2010). Reference crop evapotranspiration derived from geostationary satellite imagery –a case study for the fogera flood plain, nwethiopia and the jordan valley, jordan. *Hydrology and Earth System Sciences Discussions*, 7:4925–4956.

Ford, E. D. and Kennedy, M. C. (2011). Assessment of uncertainty in functional-structural plant models. *Annals of Botany*, 108(6):1043–1053.

Granier, C., Aguirrezabal, L., Chenu, K., Cookson, S. J., Dauzat, M., Hamard, P., Thioux, J.-J., Rolland, G., Bouchier-Combaud, S., Lebaudy, A., et al. (2005). Phenopsis, an automated platform for reproducible phenotyping of plant responses to soil water deficit in arabidopsis thaliana permitted the identification of an accession with low sensitivity to soil water deficit. *New Phytologist*, 169(3):623–635.

Grievank, A. (2000). Evaluating derivatives, principles and techniques of algorithmic differentiation. *SIAM*.

Gunter, W., Marijn, V., Aloe, A., and Bouraoui, F. (2008). A european irrigation map for spatially distributed agricultural modelling. *Agricultural Water Management 2008*, 96:771–789.

Guo, Y., Ma, Y., Zhan, Z., Li, B., Dingkuhn, M., Luquet, D., and de Reffye, P. (2006). Parameter optimization and field validation of the functional-structural model greenlab for maize. *Annals of Botany*, 97:217–230.

Hadi, H., Khazaei, F., Babaei, N., and Daneshian, J.and Hamidi, A. (2012). Evaluation of water deficit on seed size and seedling growth of sunflower cultivars. *International Journal of AgriScience*, 2(03):280–290.

Hammer, G., Cooper, M., Tardieu, F., Welch, S., Walsh, B., van Eeuwijk, F., Chapman, S., and Podlich, D. (2006). Models for navigating biological complexity in breeding improved crop plants. *Trends in Plant Science*, 11(12):1360–1385.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). 14.3.12 hierarchical clustering. *The Elements of Statistical Learning*, pages 520–528.

Helton, J., Johnson, J., Salaberry, C., and Storlie, C. (2006). Survey of sampling based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety*, 91:1175–1209.

Heuvelink, E. (1996). Re-interpretation of an experiment on the role of assimilate transport-resistance in partitioning in tomato. *Annals of Botany*, 78:467–470.

Heuvelink, E. (1999). Evaluation of a dynamic simulation model for tomato crop growth and development. *Annals of Botany*, 83:413–422.

Homma, T. and Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety*, 52:1–17.

Hornberger, G. and Spear, R. (1981). An approach to the preliminary analysis of environmental systems. *Journal of Environmental Management*, 7:7–18.

Jeuffroy, M.-H., Barbottin, A., Jones, J., and Lecoeur, J. (2006a). Chapter 10: Crop models with genotype parameters. *Working with Dynamic Crop Models*, pages 281–307.

Jeuffroy, M.-H., Valantin-Morison, M., Champolivier, L., and Reau, R. (2006b). Azote, rendement et qualite des graines : mise au point et utilisation du modele azodyn-colza pour ameliorer les performances du colza vis-a-vis de l'azote. *OCL*, 13(6):388–392.

Jones, J., Makowski, D., and Wallach, D. (2006). Chapter 8: Introduction to section ii. *Working with Dynamic Crop Models*, pages 251–256.

Jones, P. and Carberry, P. (1987). A technique to develop and validate simulation models. *Agricultural Systems*, 46:427–442.

Jullien, A., Mathieu, A., Allirand, J.-M., Pinet, A., de Reffye, P., Cournï£¡de, P.-H., and Ney, B. (2011). Characterisation of the interactions between architecture and source:sink relationships in winter oilseed rape (brassica napus l.) using the greenlab model. *Annals of Botany*, 107(5):765–779.

Kang, F., Letort, V., Magaldi, H., Cournede, P.-H., and Lecoeur, J. (2012a). Sunlab: a functional-structral model for genotypic and phenotypic characterization of the sunflower crop. *Plant Growth Modeling, Simulation, Visualization and Applications (PMA)*, pages 148–151.

Kang, M., Heuvelink, E., Carvalho, S., and de Reffye, P. (2012b). A virtual plant that responds to the environment like a real one: the case for chrysanthemum. *New Phytologist*, 195(2):384–395.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis*, 24(7):881–890.

Keating, B., Robertson, M., Muchow, R., and Huth, N. (1999). Modelling sugarcane production systems. i. description and validation of the apsim sugarcane module. *Field Crops Research*, 61:253–272.

Kellomaki, S., Oker-Blom, P., and Kuuluvainen, T. (1985). The effect of crow and canopy structure on light interception and distribution in a tree stand. *Crop Physiology of Forest Trees*, pages 107–115.

Laarhoven, P. and Aarts, E. (1987). *Simulated Annealing: Theory and Applications*. D. Reidel Publishing Company.

Lecoeur, J., Poire-Lassus, R., Christophe, A., Pallas, B., Casadebaig, P., Debaeke, P., Vear, F., and Guiloni, L. (2011). Quantifying physiological determinants of genetic variation for yield potential in sunflower. sunflo: a model-based analysis. *Functional plant biology*, 38(3):246–259.

Letort, V. (2008). *Multi-scale analysis of source-sink relationships in plant growth models for parameter identification. Case of the GreenLab model.* PhD thesis, Ecole Centrale Paris.

Letort, V., Mahe, P., Cournï£¡de, P.-H., de Reffye, P., and Courtois, B. (2008). Quantitative genetics and functional-structural plant growth models: Simulation of quantitative trait loci detection for model parameters and application to potential yield optimization. *Annals of Botany*, 101(8):951–963.

Li, G., Hu, J., Wang, S.-W., Georgopoulos, P., Schoendorf, J., and Rabitz, H. (2006). Random sampling-high dimensional model representation (rs-hdmr) and orthogonality of its different order component functions. *Journal of Physical Chemistry*, A(110):2474–2485.

Li, G., W., S. W., and R., H. (2002). Practical approaches to construct rs-hdmr component functions. *Journal of Physical Chemistry*, 106:8721–8733.

Li, Z., Le Chevalier, V., and Cournède, P.-H. (2009). Towards a continuous approach of functional-structural plant growth. In Li, B.-G., Jaeger, M., and Guo, Y., editors, *3rd International Symposium on Plant Growth and Applications(PMA09), Beijing, China*. IEEE.

Loffler, C., Wei, J., Fast, T., Gogerty, J., Langton, S., Bergman, M., Merril, R., and Cooper, M. (2005). Classification of maize environments using crop simulation and geographic information systems. *Crop Science*, 45:1708–1716.

Ma, Y., Wen, M., Guo, Y., Li, B., Cournède, P.-H., and de Reffye, P. (2008). Parameter optimization and field validation of the functional-structural model greenlab for maize at different population densities. *Annals of Botany*, 101(8).

Makowski, D., Hillier, J., Wallach, D., Andrieu, B., and Jeuffroy, M.-H. (2006). Parameter estimation for crop models. In Wallach, D., Makowski, D., and Jones, J., editors, *Working with Dynamic Crop Models*, pages 55–100. Elsevier.

Marcelis, L., Heuvelink, E., and Goudriaan, J. (1998). Modelling of biomass production and yield of horticultural crops: a review. *Scientia Horticulturae*, 74:83–111.

Messina, C., Boote, K., Loffler, C., Jones, J., and Vallejos, C. (2006). Chapter 11: Model-assisted genetic improvement of crops. *Working with Dynamic Crop Models*, pages 309–335.

Millennium Ecosystem Assessment (2005). Millennium ecosystem assessment: ecosystems and human well-being; a framework for assessment. *World Resources Institute, Washington, DC*, pages 51–53.

Miller, R. G. (1974). The jackknife - a review. *Biometrika Trust*, 61:1–15.

Minchin, P. and Lacointe, A. (2005). New understanding on phloem physiology and possible consequences for modelling long-distance carbon transport. *New Phytologist*, 166:771–779.

Monteith, J. (1977). Climate and the efficiency of crop production in britain. *Proceedings of the Royal Society of London B*, 281:277–294.

Morris, M. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33:161–174.

Oakley, J. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a bayesian approach. *J. Royal Stat. Soc.*, B(66):751–769.

Pallas, B., Loi, C., Christophe, A., Cournède, P. H., and Lecoeur, J. (2010). Comparison of three approaches to model grapevine organogenesis in conditions of fluctuating temperature, solar radiation and soil water content. *Annals of Botany*.

Pannell, D. (1997). Sensitivity analysis of normative economic models: Theoretical framework and practical strategies. *Agricultural Economics*, 16:139–152.

Ponciano, J. M., Burleigh, J. G., Braun, E. L., and Taper, M. L. (2012). Assessing parameter identifiability in phylogenetic models using data cloning. *Syst Biol 2012*, 61(6):955–972.

Qi, R., Ma, Y., Hu, B., De Reffye, P., and Cournède, P.-H. (2010). Optimization of source-sink dynamics in plant growth for ideotype breeding: a case study on maize. *Computers and Electronics in Agriculture*, 71(1):96–105.

Rannala, B. (2002). Identifiability of parameters in mcmc bayesian inference of phylogeny. *Syst. Biol.*, 51:754–760.

Rawson, H., Gardner, P., and Long, M. (1987). Sources of variation in specific leaf area in wheat grown at high temperature. *Australian Journal of Plant Physiology*, 14(3):287–298.

Reymond, M. (2001). Variabilite genetique des reponses de la croissance foliaire du mais a la temperature et au deficit hydrique. combinaison d'un modele ecophysiologique et d'une analyse qtl. *These de l'Ecole Nationale Superieure Agronomique de Montpellier, Montpellier, France*, page 70.

Reymond, M., Muller, B., Leonardi, A., Charcosset, A., and 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.

Rosegrant, M. W., Cai, X., and Sarah, A. (2002). World water and food to 2025: Dealing with scarcity. *IFPRI*.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global Sensitivity Analysis*. John Wiley&Sons, the primer edition.

Saltelli, A., Tarantola, S., and Campolongo, F. (2000). Sensitivity analysis as an ingredient of modeling. *Statistical Science*, 15(4):377–395.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). Sensitivity analysis in practice: A guide to assessing scientific models. *John Wiley and Sons*.

Shi, Y. and Eberhart, R. (1998). A modified particle swarm optimizer. In Belew, K. and L.B., B., editors, *Evolutionary Computation Proceedings (IEEE World Congress on Computational Intelligence)*, pages 69–73. Morgan Kaufmann.

Shiklamov, I. and Baser, I. (1998). World water resources. a new appraisal and assessment for the 21st century. *International Hydrological Programme (IHP) report.*

Siebert, S., Doll, P., Feick, S., Frenken, K., and Hoogeveen, J. (2007). Global map of irrigation areas version 4.0.1. *University of Frankfurt (Main), Germany, and FAO, Rome, Italy.*

Sievanen, R., Nikinmaa, E., Nygren, P., Ozier-Lafontaine, H., Perttunen, J., and Hakula, H. (2000). Components of functional-structural tree models. *Annals of Forest Science*, 57:399–412.

SINCLAIR, T. and SELIGMAN, N. (1996). Crop modelling: from infancy to maturity. *Agronomy Journal*, 88:698–704.

Sobol, I. (1993). Sensitivity analysis for non-linear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414.

Sotirios A, T. and Christos, N. (2009). Plant phenotyping with low cost digital cameras and image analytics. *Information Technologies in Environmental Engineering Environmental Science and Engineering*, pages 238–251.

Tardieu, F. (2003). Virtual plants: modelling as a tool for the genomics of tolerance to water deficit. *Trends in Plant Science*, 8(1):9–14.

Taylor, W. (1977). Small sample properties of a class of two-stage aitken estimator. *Econometrica*, 45(2):497–508.

Tsaftaris, S. A. and Noutsos, C. (2009). Plant phenotyping with low cost digital cameras and image analytics. *Information Technologies in Environmental Engineering Environmental Science and Engineering*, pages 238–251.

Turhan, H. and Baser, I. (2004). In vitro and in vivo water stress in sunflower (helianthus annuus l.). *Helia*, 27(40):227–236.

Wallach, D. (2006). Chapter 2: Evaluating crop models. *Working with Dynamic Crop Models*, pages 11–50.

Walter, A., Studer, B., and Kolliker, R. (2012). Advanced phenotyping offers opportunities for improved breeding of forage and turf species. *Annals of Botany*, 110(6):1271–1279.

Walter, E. and Pronzato, L. (2006). *Identification de Modèles Paramètrique*. Masson.

Wells, J. (1992). Analysis and interpretation of binding at equilibrium, in receptor-ligand interactions: A practical approach. *Oxford University Press*, 61:289–395.

Wu, L., De Reffye, P., Hu, B., Le Dimet, F.-X., and Cournï£¡de, P.-H. (2005). A water supply optimization problem for plant growth based on greenlab model. *ARIMA*, 3:194–207.

WU, Q. and Cournède, P.-H. (2010). The use of sensitivity analysis for the design of functional structural plant models. *Sixth International Conference on Sensitivity Analysis of Model Output*, 2(6):7768–7769.

Wu, Q., Cournède, P.-H., and Mathieu, A. (2012). An efficient computational method for global sensitivity analysis and its application to tree growth modelling. *Reliability Engineering and System Safety*, 107:35–43.

Yin, X., Struik, P., and Kropff, M. (2004). Role of crop physiology in predicting gene-to-phenotype relationships. *Trends in Plant Science*, 9(9):426–432.

Yin, X. and Struik, P. C. (2010). Modelling the crop: from system dynamics to systems biology. *Journal of Experimental Botany*, 61(8):2171–2183.

Yin, X., Struik, P. C., van Eeuwijk, F. A., Stam, P., and Tang, J. (2006). Qtl analysis and qtl-based prediction of flowering phenology in recombinant inbred lines of barley. *Journal of Experimental Botany*, pages 1–10.