Crop model parameter estimation and sensitivity analysis for large scale data using supercomputers

by

Abhishes Lamsal

B.S., Tribhuvan University, 2007
M.S., Tribhuvan University, 2009

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2017

Abstract

Global crop production must be doubled by 2050 to feed 9 billion people. Novel crop improvement methods and management strategies are the *sine qua non* for achieving this goal. This requires reliable quantitative methods for predicting the behavior of crop cultivars in novel, time-varying environments. In the last century, two different mathematical prediction approaches emerged (1) quantitative genetics (QG) and (2) ecophysiological crop modeling (ECM). These methods are completely disjoint in terms of both their mathematics and their strengths and weaknesses. However, in the period from 1996 to 2006 a method for melding them emerged to support breeding programs.

The method involves two steps: (1) exploiting ECM's to describe the intricate, dynamic and environmentally responsive biological mechanisms determining crop growth and development on daily/hourly time scales; (2) using QG to link genetic markers to the values of ECM constants (called *genotype-specific parameters*, GSP's) that encode the responses of different varieties to the environment. This can require huge amounts of computation because ECM's have many GSP's as well as site-specific properties (SSP's, e.g. soil water holding capacity). Moreover, one cannot employ QG methods, unless the GSP's from hundreds to thousands of lines are known. Thus, the overall objective of this study is to identify better ways to reduce the computational burden without minimizing ECM predictability.

The study has three parts: (1) using the extended Fourier Amplitude Sensitivity Test (eFAST) to globally identify parameters of the CERES-Sorghum model that require accurate estimation under wet and dry environments; (2) developing a novel estimation method (*Holographic Genetic Algorithm,* HGA) applicable to both GSP and SSP estimation and testing it with the CROPGRO-Soybean model using 182 soybean lines planted in 352 site-years (7,426 yield

observations); and (3) examining the behavior under estimation of the anthesis data prediction component of the CERES-Maize model. The latter study used 5,266 maize Nested Associated Mapping lines and a total 49,491 anthesis date observations from 11 plantings.

Three major problems were discovered that challenge the ability to link QG and ECM's: 1) model expressibility, 2) parameter equifinality, and 3) parameter instability. Poor expressibility is the structural inability of a model to accurately predict an observation. It can only be solved by model changes. Parameter equifinality occurs when multiple parameter values produce equivalent model predictions. This can be solved by using eFAST as a guide to reduce the numbers of interacting parameters and by collecting additional data types. When parameters are unstable, it is impossible to know what values to use in environments other than those used in calibration. All of the methods that will have to be applied to solve these problems will expand the amount of data used with ECM's. This will require better optimization methods to estimate model parameters efficiently. The HGA developed in this study will be a good foundation to build on. Thus, future research should be directed towards solving these issues to enable ECM's to be used as tools to support breeders, farmers, and researchers addressing global food security issues.

Crop model parameter estimation and sensitivity analysis for large scale data using supercomputers

by

Abhishes Lamsal

B.S., Tribhuvan University, 2007
M.S., Tribhuvan University, 2009

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2017

Approved by:

Major Professor
Stephen M. Welch

Copyright

Abstract

Global crop production must be doubled by 2050 to feed 9 billion people. Novel crop improvement methods and management strategies are the *sine qua non* for achieving this goal. This requires reliable quantitative methods for predicting the behavior of crop cultivars in novel, time-varying environments. In the last century, two different mathematical prediction approaches emerged (1) quantitative genetics (QG) and (2) ecophysiological crop modeling (ECM). These methods are completely disjoint in terms of both their mathematics and their strengths and weaknesses. However, in the period from 1996 to 2006 a method for melding them emerged to support breeding programs.

The method involves two steps: (1) exploiting ECM's to describe the intricate, dynamic and environmentally responsive biological mechanisms determining crop growth and development on daily/hourly time scales; (2) using QG to link genetic markers to the values of ECM constants (called *genotype-specific parameters*, GSP's) that encode the responses of different varieties to the environment. This can require huge amounts of computation because ECM's have many GSP's as well as site-specific properties (SSP's, e.g. soil water holding capacity). Moreover, one cannot employ QG methods, unless the GSP's from hundreds to thousands of lines are known. Thus, the overall objective of this study is to identify better ways to reduce the computational burden without minimizing ECM predictability.

The study has three parts: (1) using the extended Fourier Amplitude Sensitivity Test (eFAST) to globally identify parameters of the CERES-Sorghum model that require accurate estimation under wet and dry environments; (2) developing a novel estimation method (*Holographic Genetic Algorithm,* HGA) applicable to both GSP and SSP estimation and testing it with the CROPGRO-Soybean model using 182 soybean lines planted in 352 site-years (7,426 yield

observations); and (3) examining the behavior under estimation of the anthesis data prediction component of the CERES-Maize model. The latter study used 5,266 maize Nested Associated Mapping lines and a total 49,491 anthesis date observations from 11 plantings.

Three major problems were discovered that challenge the ability to link QG and ECM's: 1) model expressibility, 2) parameter equifinality, and 3) parameter instability. Poor expressibility is the structural inability of a model to accurately predict an observation. It can only be solved by model changes. Parameter equifinality occurs when multiple parameter values produce equivalent model predictions. This can be solved by using eFAST as a guide to reduce the numbers of interacting parameters and by collecting additional data types. When parameters are unstable, it is impossible to know what values to use in environments other than those used in calibration. All of the methods that will have to be applied to solve these problems will expand the amount of data used with ECM's. This will require better optimization methods to estimate model parameters efficiently. The HGA developed in this study will be a good foundation to build on. Thus, future research should be directed towards solving these issues to enable ECM's to be used as tools to support breeders, farmers, and researchers addressing global food security issues.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I wish to take this opportunity to thank all the people who continuously helped to successfully complete my dissertation from Department of Agronomy, Kansas State University.

My great moment started when I first met with my advisor, Professor Dr. Stephen M. Welch. I would like to express my deepest gratitude to him for his faith and believe in me. It wouldn't be possible for me to be in this stage without his support, and guidance. His sheer brilliance, sound technical insights and constant support were highly inspiring. I feel fortunate to be a part of his research group.

I would also like to give my special thanks to my committee members: Drs. Kelly Thorp, Sanjoy Das, Jesse Poland, and Vara Prasad. Their insightful comments, feedback, suggestion, and encouragement had always broadened my knowledge and skill throughout my study period. I would also like to thank Dr. Jeffery White at USDA, Arizona, Drs. Jim Jones and Ken Boote at University of Florida and my lab's post doc. Dr. Wen Fung Leong for their valuable comment and suggestion throughout the research period.

I would also like to acknowledge Dr. David Turner, Department of Computer Science at KSU and Dr. John Fonner, Texas Advanced Computing Center (TACC) for their invaluable assistance while using Beocat and Stampede computing center.

Most importantly, my special thanks goes to my parents for believing in me and giving me the autonomy to pursue my dreams. Without their numerous sacrifices and support, I would not have been able to complete my Ph.D. and enjoy my personal and professional growth. My supportive wife, life partner, and soul mate, Ms. Anju Giri, my deepest love and thanks to you for always being on my side throughout these years.

Dedication

I would like to dedicate this

to my parents

(Father: Babu Ram Lamsal and Mother: Shanti Devi Lamsal)

My beautiful wife, Anju Giri

And my lovely son, Avyn Lamsal

CHAPTER 1 - INTRODUCTION

1.1 Crop Models

It has been predicted that world population will continue to grow and will likely reach 9 billion by middle of this century (Godfray et al., 2010). It is widely believed that global crop production must be doubled by this time to meet the worlds need for food, fiber and fuel resources (Ray et al., 2013). Meanwhile, farmers are also experiencing competition for land, water, and energy for crop production, and factors like climate change along with declining water resources are hindering crop production. To overcome this problem, novel crop improvement and management methods are essential to increase breeding rates of gain as well as on-farm yields through enhanced management strategies. A central requirement for these tasks is the reliable quantitative methods for predicting the behavior of different crop cultivars in novel, time varying environments.

In the 20th century two very different mathematical approaches emerged to address this need. Crop simulation models are ecophysiological process-based and have the ability to predict phenotypes of different cultivars in response to the environment and management (White, 2009). These models use differential equations to represent physiological (photosynthesis, respiration, growth and carbohydrate partitioning, development of reproductive structure), chemical and physical (soil chemical transformations, energy flows, diffusion of gases in to leaves) and other processes (White and Hoogenboom, 2010). They are complex and non-linear (Román-Paoli et al., 2000) but play an essential role in understanding and predicting the likely behavior of crop systems (Xu and Gertner, 2011).

These models are widely utilized in agriculture as research tools to predict cropping system outcomes under different climate, soil and management conditions (Jones et al., 2003). They

typically contain large numbers of parameters (i.e., numerical constants) which represent the physical and biological characteristics of the crop and its environment. Models simulate crop growth and development as a function of these many parameters that describe the climatic, genetic, soil, and crop management features. The accuracy of model predictions strongly depend on the accuracy of many (but not necessarily all) of these parameters (DeJonge et al., 2012).

The second approach is quantitative genetics. Geneticists are continuously trying to identify the genes and genetic loci involved in plant adaption to different environmental conditions to support the efforts of breeders in developing new genotypes that are optimized for challenging environmental conditions. Quantitative genetic models are typically based on linear algebra and often seek to predict phenotypic endpoints (e.g., yields, flowering dates) rather than the progress of crop development through time. Their strength is that, unlike crop simulation models, their independent variables are explicitly genetic; i.e., marker states, alleles, etc.

Because the two approaches are so different, the communities that use and advance them are essentially disjoint. However, there has been an increasing amount of research on how to meld them (Hoogenboom et al., 2004; Messina et al., 2006; Reymond et al., 2003; Yin et al., 2003) for use in breeding programs. Hammer et al. (2002) stated that crop simulation models can help in interpreting breeding results, enhance breeding strategies, develop superior traits in combinations with genetic markers via improved phenotype prediction of prospective genotypes in novel, time–varying environments, and improve understanding of specific gene alterations affecting plant behavior. All of these advances should increase the rate of genetic gain, thus reducing the number of breeding cycles needed to reach a given target. Other improvements would accrue from enhanced, model-based production management.

The basic notion has two parts. The first is to exploit ecophysiological crop models to describe the intricate, dynamic, and environmentally responsive biological mechanisms that determine crop growth and development on daily/hourly time scale to predict the phenotypes of interest within different possible environments and in-field management options. The second part is to use quantitative genetic methods (Cooper et al., 2016; Technow et al., 2015) to relate the values of the model's biological parameters (herein after called *genotypic specific parameters*, GSP's) that encode the responses of different genetic lines to genotypic markers. In this way one might predict the behavior of an offspring whose parents have not yet been crossed in an environment that might not yet exist. To do so, one would (1) use the offspring's markers to predict the values of its ecophysiological model constants and then (2) use the crop simulation model to predict the phenotypes in the environments of interest (Reymond et al., 2003).

1.2 Overall Objective

The methodology just described imposes an enormous computational burden. First of all the ecophysiological models have many GSP's. Secondly, it may not be feasible to directly measure all of the environmental properties (e.g. depth profiles of soil hydraulic parameters), so they may need to be estimated concurrently with the GSP's. These environmental properties can be referred to as *site-specific parameters* (SSP's). Thirdly, to employ quantitative genetic methods, the GSP's of hundreds to thousands of lines must be estimated. *Therefore, the overall objective of this dissertation is to examine ways to reduce this burden.*

Stated briefly, the three chapters that follow describe (1) the use of global sensitivity analysis to identify the subset of parameters most requiring estimation, (2) a novel estimation procedure applicable when both plant and soil parameters must be estimated, and (3) an estimation effort involving an exceptionally large number of lines. The latter reveals and diagnoses some

particular issues requiring resolution if the scheme for melding ecophysiological and quantitative genetic models is to be truly viable. The remaining sections of this chapter provide (1) background on sensitivity analysis and (2) parameter estimation by way of justifying (3) the specific objectives and undertakings the statements of which follow.

1.3 Sensitivity Analysis

Sensitivity analysis (SA) is a fundamental tool for supporting mathematical model development and use (Taranatola and Salteli, 2003). SA studies "how the variation of the output of a model can be apportioned to different input variations" (Saltelli et al., 2000). SA eventually helps to reduce and simplify models by avoiding redundancies in their structure along with over parameterization (Taranatola and Salteli, 2003). Crop models in general are over parameterized for reasons that reflect biological reality. (Stated in one way, plant evolution is under no mandate to produce mathematically concise systems; the only imperative is successful offspring.) However, parameter estimation in high-dimensional parameter spaces is both algorithmically and computationally challenging. Dimensionality is reduced by restricting estimation to only those parameters found to be influential by SA (Van Griensven et al., 2006). Additionally, because errors in unimportant parameters will not greatly affect model outputs, they can be set to nominal values, thus reducing over parameterization.

According to Saltelli et al. (2000), there two major categories of SA methods, local and global. Local SA methods calculate sensitivity by varying a single parameter at a time in the neighborhood of some nominal set of values while holding all others constant. This method is most valid when applied with a linear model but can be quite misleading when applied for nonlinear models because model responses can depend strongly on interactions between variables that are

separately influenced by individual parameters. Thus, although they are constants, the parameters interact (Christopher Frey and Patil, 2002).

Global SA methods overcome these drawbacks. They calculate the sensitivity by examining the model output responses across the parameter space by simultaneous parameter perturbations (McRae et al., 1982). This allows examination of model output responsiveness to both single and multiple interacting of parameters. This can lead to substantial reductions in estimation dimensionality by limiting attention to the select few parameters that most influence model outputs.

However, while global sensitivity analysis can evaluate interactions between parameters in particular environments, there has been little work done on applying this type of analysis across greatly different environments (DeJonge et al., 2012; Jones et al., 2012). Such comparative study can lead to further computational savings when particular types of environments are of interest. For example, understanding environment-dependent patterns of sensitivity is particularly important in a state like Kansas where the availability of water is changing in the face of aquifer dewatering and a shifting climate. Therefore, Chapter 2 compares the results of a global sensitivity analysis algorithm applied to simulations conducted at multiple sites under comparatively wet and dry scenarios.

1.4 Parameter Estimation

Throughout most of the history of crop simulation, it has been impractical to directly measure all the GSP's and SSP's that models require for more than a few varieties and sites. Therefore, researchers have had no choice but to estimate these parameters indirectly from field data on phenology, yield and seed size (Alderman et al., 2015; Mavromatis et al., 2002, 2001; Pathak et al., 2012; Welch et al., 2002) or reproducing gene effects on crop development and yield

(Messina et al., 2006). Parameter estimation is a process of iteratively adjusting model parameter values to so that model predictions align closely with observed data sets, where "closeness" is evaluated by a goodness-of-fit function. The goal is to maximize the model's predictive power when applied to conditions outside of the training datasets. Successful completion of this step is a prerequisite for any application of a model (He et al., 2009).

Because crop models are often complex and non-linear (Román-Paoli et al., 2000), there is little likelihood that the algorithms traditionally used in statistics to optimize model parameter will be effective (Klepper and Hendrix, 1994; Klepper and Rouse, 1991). Such methods are much more likely to converge to some parameter values that are "optimal" only in relation to those within their local neighborhood in the parameter space (Wallach et al., 2011, 2011). Therefore, studies have compared different algorithms for parameter estimation in complex models. Surprisingly, a very frequent approach is trial and error (Wallach et al., 2001), wherein different parameters values are tested manually until an acceptable appearing match between simulated and observed data is found. This approach, of course, becomes highly inefficient as the amount of parameter space increases. Thus, numerous off-the-shelf automated optimization techniques have also been utilized. Examples include use of global optimization techniques such as the genetic algorithm (J. P. Pabico et al., 1999; Thorp et al., 2012), simulated annealing (Thorp et al., 2008), sequential search software (GENCALC; Hunt et al., 2001), Uniform covering by Probabilistic Region (UCPR; Klepper and Hendrix, 1994; Román-Paoli et al., 2000), the simplex method (Royce et al., 2001; Xinli et al., 1995), iterative grid searches (Mavromatis et al., 2001; Welch et al., 2002), particle swarm optimization (PSO; Koduru et al., 2007), and generalized likelihood uncertainty estimation (GLUE;He et al., 2010, 2009).

Most recently, however, with advancements such as high-throughput phenotyping and plantings containing the huge numbers of lines necessitated by quantitative genetics studies, data sets are becoming available whose size greatly exceeds what previously passed as "large" (e.g., Mavromatis et al., 2002; Welch et al., 2002). This provides a motivation to examine parameter estimation in the particular case of "big data" – a context likely to characterize most, if not all crop model applications in the future. Chapters 3 and 4 undertake this task using two different data sets (one originating in the private sector and one from the public sector) and estimation approaches.

1.5 Objectives/Outline

Currently, ecophysiological crop models exist for all major crops, many minor ones, and some weed species. The studies in the next three chapters will focus on three particular models, CERES-Sorghum, CROPGRO-Soybean, and CERES-Maize, in that order. These models are all part of a single software suite, DSSAT (Decision Support System for Agro-technology Transfer). DSSAT has been used for well over two decades by researchers all over the world. The full package supports simulation of 42 different species including cereal, legume, forage, and oilseed crops (Hoogenboom et al., 2015; Jones et al., 2003).

Chapter 2 presents a global SA of the CERES-Sorghum model using the Extended Fourier Amplitude Sensitivity Test (eFAST) algorithm (Saltelli et al., 1999). Kansas is the second largest sorghum producing state in the U.S. Both dryland and irrigated sorghum production is common in Kansas (Assefa et al., 2013). Although, as discussed above, one would ideally wish to estimate all parameters, this is not likely to be computationally feasible in a big-data world. Thus, knowing which parameters in which environments most influence outputs (and, therefore, can most inflate uncertainties in model predictions) can help greatly in focusing estimation efforts (Staggenborg and Vanderlip, 2005; Wang et al., 2013). Therefore, the main objective of this chapter is to

determine the most influential CERES-Sorghum input parameters under wet and dry conditions in Kansas.

The objective of Chapter 3 is to develop a novel algorithm that increases the computational efficiency of parameter estimation by exploiting the data structure embedded in large-scale field performance trials along with other relationships that may exist among the parameters too. The algorithm is specifically designed for situations wherein the parameters of a nonlinear model are subject to linear constraints. In our situation these were largely bound constraints expressed in linear forms. We call our method the "Holographic Genetic Algorithm" (HGA) because, just as all small areas in a hologram contain information on all parts of the 3D image, the software analyses all of the data to determine many additional constraints imposed by the exact pattern of site-year-line combinations provided. The data set comprises 7426 site-year-line soybean yield means. Applied to the CROPGRO-Soybean model, HGA estimates seven GSP's for each of 182 lines and three SSP's for each of 353 site-years for a total of 2333 parameters. The ratio of 7426 observations to 2333 parameters (i.e., 3.18) would be undesirably low but it was acceptable in this instance because of large number of constraints that HGA extracted from the data.

The test data were posted by Syngenta AG as part of a predictive soybean modeling contest conducted in 2015-16 in collaboration with the Institute for Operations Research and the Management Sciences (INFORMS). HGA and all required CROPGRO-Soybean model executions were performed on the Stampede and Beocat supercomputers located, respectively, at the Texas Advanced Computing Center of the University of Texas and Kansas State University.

With the advances in plant genomics and the falling costs of locating genetic markers, efforts are being made to link GSP's to actual genes and/or QTL's (quantitative trait loci) (Boote et al., 2003; Messina et al., 2006; Wilczek et al., 2009). The original goal of Chapter 4 was to do

this for four CERES-Maize GSP's that control anthesis date. Even though we observed very good estimation result using our new HGA algorithm in Chapter 2, we also saw some possible equifinality issues. Thus, in Chapter 4, we used different approach for estimation which allowed us to study the parameter response surface and characterize the equifinality. We used the 5266 lines of the Maize Nested Association Mapping population, subsets of which had been grown at 11 site-years (49491 total site-year-line combinations). The first novel part of this study used a quasi-random multidimensional search wherein the total number of time consuming model runs does not depend on the number of lines but only on the number of site-years and desired parameter precision. As in Chapter 2, all computer processing was done on Stampede and Beocat.

The second novel part was to be a comparison of the GSP QTL's with quantitative trait loci that had been previously found by directly mapping anthesis dates in these same lines and plantings. However, despite the high predictive quality of the data fits obtained, several very mysterious artifacts emerged during estimation which rendered mapping impossible. This unexpected and adverse finding has serious negative implications for the methodology that most workers currently envision as the route forward for combining ecophysiological and quantitative genetic models. Therefore, the second objective of the paper became to analyze these artifacts and determine their source. This was achieved and future work can focus on finding remedies.

1.6 Reference

Alderman, P.D., Boote, K.J., Jones, J.W., Bhatia, V.S., 2015. Adapting the CSM-CROPGRO model for pigeonpea using sequential parameter estimation. Field Crops Res. 181, 1–15.

Assefa, Y., Roozeboom, K.L., Thompson, C., Schlegel, A., Stone, L., Lingenfelser, J., 2013. Corn and Grain Sorghum Comparison: All Things Considered. Academic Press.

Boote, K., Jones, J.W., Batchelor, W., Nafziger, E., Myers, O., 2003. Genetic coefficients in the CROPGRO–soybean model. Agron. J. 95, 32–51.

Christopher Frey, H., Patil, S.R., 2002. Identification and review of sensitivity analysis methods. Risk Anal. 22, 553–578.

Cooper, M., Technow, F., Messina, C., Gho, C., Totir, L.R., 2016. Use of Crop Growth Models with Whole-Genome Prediction: Application to a Maize Multienvironment Trial. Crop Sci.

DeJonge, K.C., Ascough II, J.C., Ahmadi, M., Andales, A.A., Arabi, M., 2012. Global sensitivity and uncertainty analysis of a dynamic agroecosystem model under different irrigation treatments. Ecol. Model. 231, 113–125. doi:10.1016/j.ecolmodel.2012.01.024

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C., 2010. Food security: the challenge of feeding 9 billion people. science 327, 812–818.

Hammer, G.L., Kropff, M.J., Sinclair, T.R., Porter, J.R., 2002. Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. Eur. J. Agron., Process Simulation and Application of Cropping System Models 18, 15–31. doi:10.1016/S1161-0301(02)00093-X

He, J., Dukes, M.D., Jones, J.W., Graham, W.D., Judge, J., 2009. Applying GLUE for estimating CERES-Maize genetic and soil parameters for sweet corn production. Trans. ASABE 52, 1907–1921.

He, J., Jones, J.W., Graham, W.D., Dukes, M.D., 2010. Influence of likelihood function choice for estimating crop model parameters using the generalized likelihood uncertainty estimation method. Agric. Syst. 103, 256–264. doi:10.1016/j.agsy.2010.01.006

Hoogenboom, G., Jones, J.W., Wilkens, P.W., Porter, C.H., Hunt, L.A., Singh, U., Lizaso, I., White, J., Uryasev, O., Ogoshi, R.M., Koo, J., Shelia, V., Tsuji, G.Y., 2015. Decision Support System for Agrotechnology Transfer (DSSAT) version 4.5 (htttp://dssat.net)., DSSAT Foundation. Prosser, Washington.

Hoogenboom, G., White, J.W., Messina, C.D., 2004. From genome to crop: integration through simulation modeling. Field Crops Res. 90, 145–163.

Hunt, L., White, J., Hoogenboom, G., 2001. Agronomic data: advances in documentation and protocols for exchange and use. Agric. Syst. 70, 477–492.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. Eur. J. Agron. 18, 235–265.

Jones, J.W., Naab, J., Fatondji, D., Dzotsi, K., Adiku, S., He, J., 2012. Uncertainties in simulating crop performance in degraded soils and low input production systems, in: Improving Soil Fertility Recommendations in Africa Using the Decision Support System for Agrotechnology Transfer (DSSAT). Springer, pp. 43–59.

J. P. Pabico, G. Hoogenboom, R. W. McClendon, 1999. DETERMINATION OF CULTIVAR COEFFICIENTS OF CROP MODELS USING A GENETIC ALGORITHM: A CONCEPTUAL FRAMEWORK. Trans. ASAE 42, 223–232. doi:10.13031/2013.13199

Klepper, O., Hendrix, E.M., 1994. A method for robust calibration of ecological models under different types of uncertainty. Ecol. Model. 74, 161–182.

Klepper, O., Rouse, D.I., 1991. A procedure to reduce parameter uncertainty for complex models by comparison with real system output illustrated on a potato growth model. Agric. Syst. 36, 375–395.

Koduru, P., Welch, S.M., Das, S., 2007. A particle swarm optimization approach for estimating parameter confidence regions, in: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation. ACM, pp. 70–77.

Mavromatis, T., Boote, K., Jones, J., Irmak, A., Shinde, D., Hoogenboom, G., 2001. Developing genetic coefficients for crop simulation models with data from crop performance trials. Crop Sci. 41, 40–51.

Mavromatis, T., Boote, K., Jones, J., Wilkerson, G., Hoogenboom, G., 2002. Repeatability of model genetic coefficients derived from soybean performance trials across different states. Crop Sci. 42, 76–89.

McRae, G.J., Tilden, J.W., Seinfeld, J.H., 1982. Global sensitivity analysis—a computational implementation of the Fourier amplitude sensitivity test (FAST). Comput. Chem. Eng. 6, 15–25.

Messina, C.D., Jones, J., Boote, K., Vallejos, C., 2006. A gene-based model to simulate soybean development and yield responses to environment. Crop Sci. 46, 456–466.

Pathak, T.B., Jones, J.W., Fraisse, C.W., Wright, D., Hoogenboom, G., 2012. Uncertainty analysis and parameter estimation for the CSM-CROPGRO-Cotton model. Agron. J. 104, 1363–1373.

Ray, D.K., Mueller, N.D., West, P.C., Foley, J.A., 2013. Yield trends are insufficient to double global crop production by 2050. PloS One 8, e66428.

Reymond, M., Muller, B., Leonardi, A., Charcosset, A., Tardieu, F., 2003. Combining quantitative trait Loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiol. 131, 664–675. doi:10.1104/pp.013839

Román-Paoli, E., Welch, S.M., Vanderlip, R.L., 2000. Comparing genetic coefficient estimation methods using the CERES-Maize model. Agric. Syst. 65, 29–41.

Royce, F.S., Jones, J.W., Hansen, J.W., 2001. Model–based optimization of crop management for climate forecast applications. Trans. ASAE 44, 1319.

Saltelli, A., Chan, K., Scott, E.M., others, 2000. Sensitivity analysis. Wiley New York.

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

Staggenborg, S.A., Vanderlip, R.L., 2005. Crop simulation models can be used as dryland cropping systems research tools. Agron. J. 97, 378–384.

Taranatola, S., Salteli, A., 2003. Methodological Advances and Innovative Applications of Sensitivity Analysis. Reliab Eng Syst Saf SAMO 79, 121–122.

Technow, F., Messina, C.D., Totir, L.R., Cooper, M., 2015. Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PloS One 10, e0130855.

Thorp, K.R., DeJonge, K.C., Kaleita, A.L., Batchelor, W.D., Paz, J.O., 2008. Methodology for the use of DSSAT models for precision agriculture decision support. Comput. Electron. Agric. 64, 276–285.

Thorp, K.R., Wang, G., West, A.L., Moran, M.S., Bronson, K.F., White, J.W., Mon, J., 2012. Estimating crop biophysical properties from remote sensing data by inverting linked radiative transfer and ecophysiological models. Remote Sens. Environ. 124, 224–233. doi:10.1016/j.rse.2012.05.013

Van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M., Srinivasan, R., 2006. A global sensitivity analysis tool for the parameters of multi-variable catchment models. J. Hydrol. 324, 10–23.

Wallach, D., Buis, S., Lecharpentier, P., Bourges, J., Clastre, P., Launay, M., Bergez, J.-E., Guerif, M., Soudais, J., Justes, E., 2011. A package of parameter estimation methods and implementation for the STICS crop-soil model. Environ. Model. Softw. 26, 386–394. doi:10.1016/j.envsoft.2010.09.004

Wallach, D., Goffinet, B., Bergez, J.-E., Debaeke, P., Leenhardt, D., Aubertot, J.-N., 2001. Parameter Estimation for Crop Models. Agron. J. 93, 757. doi:10.2134/agronj2001.934757x

Wang, J., Li, X., Lu, L., Fang, F., 2013. Parameter sensitivity analysis of crop growth models based on the extended Fourier Amplitude Sensitivity Test method. Environ. Model. Softw. 48, 171–182.

Welch, S.M., Wilkerson, G., Whiting, K., Sun, N., Vagts, T., Buol, G., Mavromatis, T., 2002. Estimating soybean model genetic coefficients from private–sector variety performance trial data. Trans. ASAE 45, 1163.

White, J., 2009. Combining ecophysiological models and genomics to decipher the GEM-to-P problem. NJAS-Wagening. J. Life Sci. 57, 53–58.

White, J.W., Hoogenboom, G., 2010. Crop response to climate: ecophysiological models, in: Climate Change and Food Security. Springer, pp. 59–83.

Wilczek, A.M., Roe, J.L., Knapp, M.C., Cooper, M.D., Lopez-Gallego, C., Martin, L.J., Muir, C.D., Sim, S., Walker, A., Anderson, J., Egan, J.F., Moyers, B.T., Petipas, R., Giakountis, A., Charbit, E., Coupland, G., Welch, S.M., Schmitt, J., 2009. Effects of genetic perturbation on seasonal life history plasticity. Science 323, 930–934. doi:10.1126/science.1165826

Xinli, W., Futang, W., Guowang, Q., 1995. APPLICATION OF OPTIMIZING THEORY TO DETERMINING GENETIC PARAMETERS INVOLVED IN CERES-SOYBEAN MODEL [J]. Q. J. Appl. METEORLOLGY S 1.

Xu, C., Gertner, G., 2011. Understanding and comparisons of different sampling approaches for the Fourier Amplitudes Sensitivity Test (FAST). Comput. Stat. Data Anal. 55, 184–198.

Yin, X., Stam, P., Kropff, M.J., Schapendonk, A.H., 2003. Crop modeling, QTL mapping, and their complementary role in plant breeding. Agron. J. 95, 90–98.

CHAPTER 2 - A SENSITIVITY ANALYSIS OF THE CERES-SORGHUM MODEL VIA FOURIER-BASED GLOBAL METHODS

Abstract

Kansas is the second largest sorghum producing state in the U.S. and both dryland and irrigated sorghum production is common. It has been proven that cropping system models can be used as a research tool in predicting production outcomes under different climate, soil and management conditions. Crop model predictive skill depends strongly on the accuracy of many input parameters and thus, one would ideally wish to estimate all model parameters. However, it is not likely to be computationally feasible in a big data world, because ..... Thus, identifying the parameters which has greater influence in output can greatly assist the estimation effort. Therefore, the main objective of this study is to determine the most influential model (CERES-Sorghum) parameters under wet and dry conditions. A global sensitivity analysis [Extended Fourier Amplitude Sensitivity Test (eFAST)] method was used to identify the parameter (cultivar, soil, agronomic) that are influential to model output (yield, anthesis days, maturity days, and leaf area index). Results showed that cultivar, soil and agronomic parameters can shift their influence dominance pattern relative to different output responses. Furthermore, it is revealed that CERES-Sorghum output responses were highly sensitive to genetic parameters in wet environment and highly sensitive to soil parameters in dry environment. Result also demonstrated that eFAST can be very useful for detecting both individual and interaction effect of model input parameters.

2.1 Introduction

Sorghum (*Sorghum bicolar (L.) Moench)* is an important cereal crop grown for food and fodder security on semi-arid tropics of African and Asian continents (Reddy et al., 2013). It is a genetically diverse crop and is grown in many countries all around the world (Mutava et al., 2011). Due to its drought tolerance, resistance to mycotoxins and fungi, and survivability in relatively adverse climatic conditions, sorghum production is increasing world-wide. In the United States, Sorghum represents the third-largest cereal grain and Kansas is the second highest producer of sorghum . In terms of production area, both dryland and irrigated cropping is common, though dryland sorghum production is far greater than irrigated sorghum production in Kansas (Assefa et al., 2013). It has been proven that cropping system models can be used as a research tool in predicting production outcomes under different climate, soil and management conditions (Jones et al., 2003; Staggenborg and Vanderlip, 2005), but uncertainty associated with model input parameters for different lines and environmental conditions restricts their wide application (Staggenborg and Vanderlip, 2005; Wang et al., 2013).

In general, models play a crucial role in understanding and predicting the potential behaviors of many systems ranging from physics and chemistry to biology, the environmental sciences, the social sciences and beyond (Xu and Gertner, 2011). Crop modeling is a powerful tool to quantify future cropping trends and to identify targets for improvements (Semenov and Shewry, 2011). White (2009) stated that process-based models are powerful tools for predicting how plant performance varies in response to genetic, environmental and agronomic management parameters (Hammer et al., 2002; Jones et al., 2012; Lobell and Ortiz-Monasterio, 2007). The Crop Estimation through Resource and Environment Synthesis (CERES)-Sorghum model is one of the oldest, crop simulation models for Sorghum (Virmani et al., 1989). We used the version incorporated in CSM

4.5 Cropping System Model; (Hoogenboom et al., 2015; Jones et al., 2003). CERES-Sorghum is a complex, nonlinear, dynamic system which simulates Sorghum growth and development as a function of a large number of input variables and parameters describing climatic, genetic, soil, and crop management factors (Ritchie and Alagarswamy, 1989a; Virmani et al., 1989)

Crop model have been extensively used in research aimed at predicting the outcomes of different production systems (e.g., irrigated vs. dryland) (Rosenzweig, 1990; Saseendran et al., 2008; Staggenborg and Vanderlip, 2005; Xie et al., 2001; Ziaei and Sepaskhah, 2003). Crop model predictive skill depends strongly on the accuracy of many input parameters that describe the fixed properties of varieties and soils and which are commonly based on field experiments (DeJonge et al., 2012). However, not all of these parameters are equally influential on model outputs, and determining which ones are the most important necessitates a sensitivity analysis (SA). In addition, the influence of parameters can vary with environmental conditions; thus, parameters that are important to one production system might not have the same impact elsewhere (Confalonieri et al., 2010; Wang et al., 2013). Thus, care is warranted when using parameters at one location that were estimated at another. This makes it desirable to identify parameter sensitivities at multiple sites.

Sensitivity analysis (SA) is a fundamental tool for supporting mathematical model development and use (Taranatola and Salteli, 2003). SA addresses the question of "how the variation of the output of a model can be apportioned to different input variation" (Saltelli et al., 2000). It is the most sensitive parameters that demand the greatest experimental accuracy. Along with understanding a model's behavior, SA also helps to reduce and simplify the modeling process by avoiding over parameterization and redundant model structure (Taranatola and Salteli, 2003;

Van Griensven et al., 2006). Spear and Hornberger (1980) further mentioned that SA helps to reduce the number of parameters that require fitting with output data.

Crop models have many input parameters (indeed, they are over-parameterized), many output variables, and may be used in many environments. This can generate many output/input × environment combinations for which sensitivities might be desired, making SA computationally demanding. Therefore, most such analyses have been commonly carried out by focusing only some of the general factors known *a priori* in a general manner to significantly affect the crop growth and development such as water, fertilizer and climate. Thus, it remains quite possible that there might be other factors which also impact crop growth and development to an unexpected degree (Jones et al., 2012). To uncover these, should they exist, requires an SA method that can accommodate the large numbers of combinations.

There are primarily two different kinds of SA methods (Saltelli et al., 2000): local and global. Local SA methods calculate sensitivity by varying one parameter at a time in a small neighborhood of some nominal set of values, keeping the rest of the parameters constant. The result of local SA heavily depends on the base value of the input parameter. The results may not adequately inform when applied to non-linear models because of interaction effects and parameter sensitivities that can easily change with the value of the parameter (i.e., its location in parameter space; Frey and Patil, 2002). Global SA methods overcome this drawback by examining the overall response of model outputs to variation across the parameter space (McRae et al., 1982). These methods compute the sensitivity of specific output variables to both single parameters and multiple, interacting parameters.

To the best of our knowledge, very little SA literature exists on CERES-Sorghum model. The most notable one is White et al. (2005), who studied the sensitivity of only a single variable

(temperature) and measured responses to several temperature levels. There do not seem to be any papers on the global sensitivity of CERES-Sorghum input parameters although such studies do exist for other crops [e.g. CERES-Maize (Jones et al 2012), STICS-wheat (Varella et al., 2010) and APSIM-wheat (Zhao et al., 2014)]. In addition, SA results examining overall sensitivity of a model cannot be adapted to different treatments (e.g. irrigated vs. dryland) because sensitivity of input parameters with respect to different response variables differ across location and climatic condition (Staggenborg and Vanderlip, 2005). Thus, the main objective of this study is to determine the sensitivity of the CERES-Sorghum output variables to many of its input parameters in both dryland and wetland cropping systems.

2.2 Background

The literature contains several global sensitivity analysis techniques including: (1) commonly used variance-based methods such as Fourier Amplitude Sensitivity Test (FAST; Cukier et al., 1973; Saltelli et al., 1999), Extended FAST (eFAST; Saltelli et al., 1999), and Sobol's method (Sobol, 2001) ; (2) the screening or elementary effect method (Morris, 1991; Compolongo et al., 2007) (3) regression based methods (Helton et al., 2006; Tondel et al., 2013), (4) Mckay's one-way ANOVA method (McKay, 1997), and (5) the moment independent approach (Borgonovo and Tarantola, 2008; Park and Ahn, 1994).

The Screening method aims at identifying parameters as either having (1) negligible influence, (2) linear additive effects, or (3) nonlinear/interaction effects perhaps in concert with other parameters. It constructs trajectories of parameter sampling points based on a randomized, one-at-a-time selection of parameters for which model runs are then done using a set of predetermined levels. The results of each run are compared to a base run and sensitivities are calculated by a finite difference approximation to the partial derivative. These values are assumed

to be samples from a probability distribution whose sample mean, μ, and standard deviation, σ, are the sensitivity measures. The first describes the overall linear additive influence of a parameter (which might be negligible) and the second quantifies the extent of non-linear (interaction) effects with other parameters (Morris, 1991; Campolongo et al., 2007). The main limitation of this method is that it can commit Type II errors and fail to identify parameters with considerable influence on the model – the price of its robustness against making a Type I error by declaring a parameter to be important when it is not (Campolongo and Saltelli, 1997).

The regression based method is based on the computation of standard or partial regression coefficients. In this approach, the parameters are viewed as regressors and the model outputs as response variables. The model is run with multiple combinations of parameter values upon which the outputs are regressed. The regression coefficients quantify the sensitivity of each parameter (Tondel et al., 2013). This method performs well when the parameter values are statistically independent of each other (Peck and Devore, 2011) but lacks robustness when key assumptions of regression are not met. In addition, the method is critically dependent on the regression model used and the range of variation of the inputs (Christopher Frey and Patil, 2002). The ANOVA method is a parametric method that assesses whether there is a statistical association between an output and one or more inputs (Krishnaiah, 1981). This method is computationally intensive when applied for large number of parameters, lacks robustness when there is significant departure of the response variable from normality, and has difficulty quantifying the sensitivities of individual parameters when they are correlated.

Variance-based SA methods, also known as global sensitivity approaches, assess how the uncertainty in the outputs are distributed across the inputs. It decomposes the variance of the model outputs into fractions which can be attributed to each of the input parameters. Except for the

classical FAST method (Cukier et al., 1973), these algorithms all calculate the total and first order sensitivity for each parameter (Confalonieri et al., 2012). The Sobol method requires the numerical integration techniques such as Monte Carlo, and thus is highly computationally intensive (Mokhtari et al., 2005).

FAST is one of the most popular approaches to compute global sensitivity and was introduced by Cukier et al. (1973). The method explores large parts of a multidimensional parameter space by drawing input samples along a space-filling periodic curve constructed by assigning different frequencies to different parameters. The notion is that if a particular model output is sensitive to a given input, then a Fourier spectrum analysis of the outputs will reveal a component at that parameter's corresponding frequency. In this paper we have used the Extended (eFAST) version that adds to FAST the capacity to calculate both main and total (interaction) sensitivities (Saltelli et al., 1999). Due to the good convergence ability with relative small sample size, eFAST is significantly less computationally expensive than other global sensitivity (Sobol sensitivity) analysis methods (Saltelli et al., 1999; Zhao et al., 2014). Even though, Screening method is often considered as low computation cost, but result is more considered as qualitative (by ranking parameters based on sensitivity) then quantitative (Dejonge et al., 2007). The mathematical description of this method is explained in methodology section below.

2.3 Materials and Methods

2.3.1 Site and Experiment Description

This study was located in Kansas (central US), where the elevation increases from the southeast (207 m) to the northwest (1231 m). Mean annual precipitation varies from east (>114 cm) to west (<46 cm), and the yearly average temperature gradient ranges from 9^{o}C (north) to >14^{o}C (south), thus framing the prevailing challenges affecting agricultural production in Kansas

(Tomilnson and Knapp, 2012). In response, Ottawa, Hutchinson, Hays, and Tribune (Table 2.1) were selected to represent all precipitation, elevation, and temperature gradients. To select dry and wet years, precipitation throughout the sorghum growing season was totaled for each year from 1950 to 2014. The years with closest to the 90th and 10th percentiles were selected to be the wet and dry years, respectively, for each site. The soil information required by CERES-Sorghum was obtained from the National Resources Conservation Service (NRCS) Web Soil Survey database (http://websoilsurvey.nrcs.usda.gov/). Dryland sorghum performance trial reports for each site-year combination (obtained from the Kansas State University Research and Extension Service) provided all required model agronomic inputs. Table 2.1 gives the detailed description of each site. All model runs were performed assuming rainfed conditions and using crop parameters present in DSSAT cultivar template file for the RS160 sorghum variety as base values (Hoogenboom et al., 2015).

2.3.2 Crop Model Description

The Cropping System Model (CSM) is one of the oldest, the most advanced, and the most widely used crop simulation models (Quiring and Legates, 2008). The sensitivity of several CERES-Sorghum (v. 4.5) outputs were measured with respect to a number of agronomic, genetic, and soil input parameters. CSM-CERES-Sorghum is based on the CERES-Sorghum model described by Singh et al. (1993). The model requires input data such as daily weather (maximum and minimum temperatures, solar radiation, relative humidity, and precipitation), soil characteristics (soil texture, pH, and soil water related parameter etc.), cultivar parameters (e.g., phenological parameter (P1, PHINT), and agronomic management practices (e.g., planting date, plant population, fertilizer application date).

CERES-Sorghum simulates crop development, growth, and yield based on environmental (meteorological and soil) and cultivar specific parameters for each individual phenological phase over time until harvest. Plant phenological development rates are calculated based on temperature and photoperiod. The model assumes that from end of juvenile phase to panicle initiation, photoperiods longer than the critical short day length slow the development. Similarly, the durations of particular crop growth stages are directly related to temperature; specifically, to the sums of mean daily air temperature above a base value (cumulative growing degree days).

The total crop biomass (expressed as dry matter) is calculated as product of average growth rate and the growth duration, which is broken up into daily time steps. The biomass increments are initially partitioned to leaves, stems, roots and (after transition to the reproductive stage) ear and grain growth (Ritchie, 1998). The CERES-Sorghum model computes daily dry matter increments based on radiation use efficiency and light interception (Ritchie, 1998) and calculates a deduction for respiration that is based on the amount of biomass existing at each time step. Light interception is estimated assuming an homogenous canopy and using a canopy-level radiation extinction coefficient that is adjusted for row width (White et al., 2005). In the CERES-models, a potential leaf expansion value is calculated that depends on the proportion of the daily dry matter increment that is allocated to leaves. However, actual leaf growth is determined by scaling this potential to reflect the impact of various stressors. Specifically, the potential expansion rate is multiplied by a 0-to-1 value that depends on temperature extremes, water deficits, and/or nitrogen insufficiency. Stress can affect the growth rate of other tissues (e.g. grain) via a similar mechanism.

The model also assumes that the ear and panicle of sorghum expand rapidly after the end of the leaf growth. Final grain yield is estimated as the product of grain numbers per plant, the individual kernel grain weight, and the number of plants per unit area. Reductions in grain yield,

if any, result from decreases in dry matter production during grain filling or stress effects that impact kernel growth rates (Ritchie, 1998). The ability to capture soil, environment and plant interaction across a wide range of environmental condition is the major strength of this model (Staggenborg and Vanderlip, 2005).

2.3.3 Input parameters and output responses:

Input parameters related to genetic, soil, and management are important for crop growth and development. The input parameters chosen for this study and their description are presented in Table 2.2 We categorized input parameters in to the following groups: a. genotypic-specific parameters (GSP's), b. soil parameters, and c. agronomic parameters. GSP's are key traits that enable a generic model to mimic the phenotypic outcomes of particular varieties grown in different regions and years (Ritchie and Alagarswamy, 1989; Jones et al., 2003). The GSP's chosen for this study are P1, P2, P2O, PHINT, P5, G1, G2, RUE, and TBASE whose definitions are given in Table 2.2. These parameters are important in such way that they mimic the phenological and reproductive characteristics of crop (Alagarswamy and Ritchie, 1991; Folliard et al., 2004; Virmani et al., 1989). Soil parameters such as DLL, DUL, SSAT, SLRO, and SLDR were taken as they influence the water content of the soil. In addition, other parameters such as SLOC, SBDM, SLPF, and SLU1 were also considered for soil parameter sensitivity. Previous studies have also demonstrated that soil water related parameters play a crucial role in simulating crop growth and development. Similar parameters were also used by DeJonge et al. (2012) to evaluate sensitivity of the CERES-Maize model whose modular structure is quite similar to CERES-Sorghum. PPOP was the only parameter taken from the agronomic category to evaluate sensitivity. Upper and lower bounds for each parameter (Table 2.2) were selected based on the experience of the authors and other CERES-Maize and Sorghum based literature (Alagarswamy and Ritchie, 1991; Bert et al.,

2007; DeJonge et al., 2012; Folliard et al., 2004; Ritchie and Alagarswamy, 1989a). Major output responses (Grain yield (Yield), Anthesis date (ADAT), Maturity date (MDAT), and Leaf area index (LAI)) were selected for this study.

2.3.4 Mathematical Basis of the Extended Fourier Amplitude Sensitivity Test

The eFAST algorithm was developed by Saltelli et al. (1999). Consider a crop model output $y = \mathrm{f}(\mathbf{x})$ where $\mathbf{X} = [x_1, \ldots \ldots, x_N]$ is a vector of the *N* model parameters of interest. The main idea of eFAST is to introduce a frequency signal into *y* for each parameter by generating periodic samples of model parameters. A subsequent Fourier transformation of *y* will reveal the partial variances of *y* contributed by different model parameters. In order to generate periodic samples for specific parameterX_i, the eFAST method uses a sinusoidal sampling function $\mathbf{x}(s)$ defined by making the i^{th} component of $\mathbf{x}$ equal to

$$x_i = x_{i,\min} + \left\{\frac{1}{2} + \frac{1}{\pi}\arcsin\left[\sin\left(\omega_i s + \varphi_i\right)\right]\right\}\left(x_{i,\max} - x_{i,\min}\right)$$

where *s* is a scalar that varies over the range (- π to + π), ω_i is a set of different angular frequencies each of which is assigned to a parameter, φ is a random phase shift chosen uniformly between [0 to 2π], and $\left[x_{i,\max} - x_{i,\min}\right]$ is the range of x_i. This yields a set of straight lines oscillating between 0 and 1; i.e. each x_i oscillates periodically at the corresponding frequency w_i. Fig. 2.1 shows the parameter oscillations before transforming to their respective ranges. Using the properties of Fourier series expansion (Saltelli et al., 1999),

$$\mathrm{E}[y] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{f}\left(\mathbf{x}(s)\right) ds$$

$$\mathrm{Var}[y] = -\mathrm{E}[y]^2 + \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{f}\left(\mathbf{x}(s)\right)^2 ds$$

$$\approx \sum_{j=-\infty}^{\infty}\left(A_j^2 + B_j^2\right) - \left(A_0^2 + B_0^2\right) \approx 2\sum_{j=1}^{\infty}(A_j^2 + B_j^2)$$

where E(y) and Var(y) are, respectively, the expected value and variance of the output y; and A_j and B_j are the Fourier coefficients over the domain of integer frequencies $j \epsilon \{-\infty, \ldots . -1, 0, 1, \ldots . . +\infty\}$. The Fourier coefficients are calculated as follows:

$$A_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} f\left(\mathbf{x}(s)\right)\cos(js)\, ds$$

$$B_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} f\left(\mathbf{x}(s)\right)\sin(js)\, ds$$

A separate sensitivity analysis is conducted for each x_i in $\mathbf{x}$. The sample size, *i.e*, the total number, N_l of runs done for the i^{th} parameter involves a tradeoff between computational capabilities and the number of parameters to be analyzed. In particular, the Nyquist criteria (Saltelli et al., 1999) requires that there be at least twice as many samples as the highest frequency to be resolved. (Stated another way, there has to be at least one more sample than the number of A_j's and B_j's.) Accurate resolving the sensitivity during the analysis of the i^{th} parameter is best achieved by both assigning it the highest frequency and also recovering several, say M, even higher harmonics of it. Thus, one can solve for $\omega_{\max}$, which is the maximum frequency to be used from the equation $N_l = 2M\omega_{\max} + 1$. This frequency is rounded down to the nearest integer and assigned to x_i.

Next frequencies must be assigned to the other parameters during the sensitivity analysis of x_i. We will refer collectively to these other parameters as the *complementary set* of x_i. This set

will be denoted by a subscript of "-*i*". First, the maximum allowable frequency with the -*i* set was calculated as

$$\omega_{-i,\max} = (1/M)(\omega_{\max}/2)$$

also rounded down to the nearest integer. Then frequencies were assigned to the -*i* parameters in a way that exhaust the range 1 to $\omega_{-i,\max}$ according to methods detailed in Saltelli et al. (1999).

2.3.5 Main effects

One measure of sensitivity of *y* to an individual input x_i is the estimated conditional variance of the i^{th} factor. Letting J_i be the set of frequency indices for parameter *i* and its *M* harmonics in the SA for x_i, then

$$\mathrm{Var}_i[y] = 2\sum_{j\in J_i}\left(A_j^2 + B_j^2\right)$$

After dividing by the total unconditional variance Var*(y)*, whose formula was given earlier, we obtain the *first order sensitivity index*

$$S_i = \frac{\mathrm{Var}_i[y]}{\mathrm{Var}[y]}$$

2.3.6 Total effects

The expression $\mathrm{Var}[y] - \sum_k \mathrm{Var}_k[y]$ is the residual variance not accounted for by any first order effects and include the interactions of any order between the parameters. We can define the total sensitivity of *y* to parameter *i* as

$$ST_i = \left\{\mathrm{Var}[y] - \sum_{k=1}^{N}\mathrm{Var}_i[y]\right\} \Big/ \mathrm{Var}[y]$$

ST_i takes into account both S_i and the interactions between the i^{th} and all other parameters. The interaction effects between the i^{th} parameter and all others can therefore be calculated as

$$ST_{-i} = \left\{ \mathrm{Var}[y] - \sum_{\substack{k=1 \\ k \neq i}}^{N} \mathrm{Var}_i[y] \right\} \Big/ \mathrm{Var}[y] = ST_i - S_i$$

2.3.7 Characteristics and Interpretation of the eFAST Sensitivity Indices:

- ST_i is always greater than or equal to S_i. If it is equal then x_i is not involved in any interactions with other input variables. Otherwise, the difference between ST_i and S_i tells the magnitude of the interactions between x_i and the other input variables.
- The sum of all S_i is always less than 1 (for non-additive models) or equal to 1 if the model is perfectly additive (no interactions). This follows from the rules governing the variances of the sums of random variables.
- The sum of all ST_i is greater than 1 (for non-additive model) or equal to 1 for perfectly additive model.
- Higher values of S_i mean that the i^{th} parameter has more influence on the model output *y*.
- A very low values of S_i signify that parameter *i* has a negligible influence on *y*.

2.3.8 Statistical Analysis

Analysis of variance (ANOVA) was done to determine the effect of site years (wet and dry) on sensitivity index of each parameter. Since the sensitivity indices of all parameter in both dry and wet years for Tribune were almost identical, Tribune was excluded in the ANOVA test. One-way ANOVA for individual parameters for each response variable was conducted using the least significant difference (LSD) procedure at 0.05 probability level. The *p*-value of each parameter was used to test whether the observed indices were significantly different between dry and wet years.

2.4 Results

The CERES-Sorghum yield output was highly sensitive to the soil parameter DLL in all dry site-years. For example, it accounted for 70% of yield variability in the dry year at Tribune. G2 and P2O were other parameters to which yield exhibited elevated sensitivity across all dry site-

years (Fig. 2.2). Similar results were also observed in the wet years for Tribune, due to the limited amount of rainfall even in nominally wet years (Table 2.1). However, a notable difference was observed in wet years at other sites, where yield was found to be most sensitive to SLRO, accounting up to 35% of yield variability. In addition, yield was slightly sensitive to both P2O, PHINT, G2, SLPF, and P1 main and interaction effects. *P*-values indicate that the sensitivity indices of parameters P2O, SLRO and DLL were significantly different across dry and wet site-years (Table 2.3). Furthermore, the SLRO S_i for yield was significantly higher in wet years whereas sensitivity indices of DLL and P2O were significantly higher in dry site-years. In all dry years, yield also showed slight sensitivity to the interaction between P1, P2O, G2, SLPF, SLRO and DLL. For other parameters (SSAT, SSKS, SBDM, SLOC, and PPOP), yield was found to be insensitive in terms of both first order and total sensitivity for any of the site years.

Anthesis date (ADAT) was highly sensitive to genetic parameter P2O accounting for about 80% followed by P1 at 17% variability (Fig. 2.3). There was very minimal interaction effect between these parameters. Although the sensitivity to PHINT was small, there was a significant difference between its effects in wet vs. dry years. No other parameter, including the highly influential P2O and P1, showed this pattern.

Maturity date (MDAT) was primarily sensitive to P2O accounting for up to 45% of variability in all dry and wet sites except Tribune (Fig. 2.4). In addition, parameters DLL, P5, P1, and SLPF are also influential to MDAT in all dry site-years except Tribune. In Tribune, MDAT was most sensitive to DLL accounting up to 62% of variability followed by P2O and P1 (20%) for both dry and wet years. In contrast, MDAT was primarily sensitive to P2O accounting 50% variability followed by P5 and P1 (20%) in Hutchinson and Ottawa wet years. Similar to yield, MDAT also indicated notable sensitivity differences between dry and wet years. The most

significantly different MDAT sensitivities were observed for P2O and SLPF (Table 2.3). The sensitivity indices of SLPF were significantly higher in dry site-years and the sensitivity indices of P2O were significantly higher in wet site-years. It was also observed that MDAT had very minimal first order sensitivity to PHINT, G1, G2, TBASE, RUE, DUL, SSAT, SSKS, SBDM, SLOC, and PPOP, which, in total accounted for <10% of the variation in any site-year. However, these parameters showed interaction effect on MDAT. Furthermore, higher interaction effects on MDAT were observed for the same parameters (DLL, P2O, P1, and P5) that also have high first order sensitivity.

LAI was primarily sensitive to DLL for all dry site-years and at Tribune in wet years (Fig. 2.5). DLL itself account more than 60% to 90% of LAI variation. The S_i of DLL is significantly higher in dry site-years than wet site-years. Except for Tribune, LAI was also sensitive to SLRO, SLPF, and P2O and P1, PHINT for all dry and wet site years. However, sensitivity indices of P1, P2O and PHINT, P5, and SLPF are significantly higher in wet site-years than dry ones. In Tribune, LAI was most sensitive to DLL accounting 90% of the variation followed by SLPF (5%) in both wet and dry years. In comparison to dry sites, wet sites experienced a slightly higher interaction effect between P1, P2O, PHINT, G1, and SLPF. Furthermore, LAI was not sensitive to parameters such as SSAT, SSKS, SBDM, SLOC, and PPOP for any dry or wet site-years.

2.5 Discussion

The comparatively higher yield and LAI S_i values for soil-related parameters vs. genetic and agronomic inputs across both dry and wet site-years (Fig. 2.2 and 2.5) suggest that primary attention should be paid to measure soil related parameters. Additionally, the notable sensitivity differences of LAI and yield to DLL and SLRO between dry and wet years document that S_i values can be highly dependent on the environment. The importance of soil-related parameters during

model calibration was also seen in Chapter 2 of this dissertation. The higher sensitivity of yield and LAI to DLL in dry site-years likely results because water levels have a greater chance of reaching the DLL when water is limited. In contrast, when water is adequate, the exact value of DLL is of much less influence because moisture levels always exceed it (Fig. 2.2 bdf). The high S_i of SLRO for yield in wet site-year's (Fig. 2.2 bdf) results because that parameter controls the proportion of precipitation that infiltrates the soil and becomes available to the crop (Bert et al., 2007). Dejonge et al. (2012) and Xie et al. (2001) have also reported that the yield and LAI predictions from CERES-Maize were highly sensitive to soil water related parameters in drier conditions.

In wet site-years, relatively higher sensitivity of LAI to genetic parameters (P1, P2O, and PHINT) suggests that when water is adequate, primary importance has to be given to genetic plant characteristics when simulating LAI. As described earlier, CERES-Sorghum calculates leaf expansion by multiplying the potential expansion rate (which is only a function of intercepted light and RUE) by a fraction varying between 0 to 1 related to temperature extremes, water deficit, and/or nitrogen deficit (Ritchie, 1998). The total leaf growth over time is the summed product of rate times duration. Thus, when stress is absent, leaf expansion is mainly controlled by temperature- and photoperiod-related genetic parameters such as P1, PHINT, and P2O.

The sensitivity of ADAT to only P1 and P2O in both dry and wet site-years contradicts the experimental result reported by Craufurd et al., (1993), who observed flowering delays under soil water deficit conditions. However, a similar simulation result was reported by Dejonge et al., (2012) for CERES-Maize. This is because in both CERES-Maize and CERES-Sorghum, ADAT is not a function of soil water availability. Instead, ADAT is determined based on the thermal time

from emergence to juvenile phase (P1), critical photoperiod hour (P2O), rate of delay of growth (P2R) as day length increase P2O (Ritchie and Alagarswamy, 1989a, 1989b).

The large difference between total and first order sensitivities (i.e., the interaction effects $ST_i - S_i = ST_{-i}$ when *i* refers to MDAT and, to a lesser extent for yield, indicates the existence of large non-linear relationships and interactions between many model processes. This strongly suggests that purely local sensitivity analysis approaches cannot quantify the response behavior of this model to multiple parameter interaction. In contrast, global sensitivity methods like eFAST can detect such effects because they (1) fully explore the uncertainty range of the parameters analyzed and (2) perturb multiple parameters simultaneously, instead of one at a time changes and (3) compute the main and interaction effect among parameters. In addition, this method is highly efficient in terms of computation time. This result further suggests that accurate prediction of a variable such as MDAT requires accurate estimates for a very large number of interacting parameters.

The miniscule sensitivity of yield to PPOP (Fig. 2.2) in both dry and wet site-years was surprising because CERES Sorghum calculates yield as a product of grain weight and plant population. However, this might occur because end yield is mostly driven by light interception and RUE and minimal effect of plant population is possible when model simulates reasonably higher LAI. Similar results of minimal effect of plant population were observed in both sorghum simulation (Baumhardt et al. 2005) and field experiments (Conley et al. 2005).

For all output responses, very low ST_i and Si were found for the soil parameters SSAT, SSKS, SBDM, and SLOC; the genetic parameters TBASE and RUE; and the agronomic parameter PPOP. This suggests that these can be set to nominal values during calibration. This result will help to reduce the total cost of computation during parameter estimation.

The results of this study will be especially helpful to researchers who use crop model at different locations and have interest in multiple response variables. In particular, ecophysiological models have a great many parameters and it is highly impractical to estimate all of them. We have categorized the parameters that need to be used in the estimation process based on their influence on response variables.

2.6 Conclusion

The influence of soil, genetic, and agronomic parameters on simulated yield, ADAT, MDAT, and LAI in eight different environmental conditions were presented in this study. The eFAST global SA approach was applied to estimate the partial variances contributed by both main and interaction effects of model parameters.

This study determined that cultivar parameters, soil parameters, and agronomic parameter can differ and shift their influence dominance patterns relative to simulated yield, ADAT, MDAT, and LAI depending on the production situation studied. This study also gave insight to some of the parameters that do not have high first order sensitivities, but have major impacts on model outputs via interactions involving other parameters. The results demonstrated that depending on the target environment and the response variable of interest, cases exist where (1) relatively few parameters might require accurate estimates (e.g. ADAT) or (2) alternatively, a great many (e.g. MDAT at Hays). Results showed that CERES-Sorghum output responses were mostly sensitive to genetic parameters in wet environments but highly sensitive to soil parameters, especially DLL, in dry land conditions. This result will reduce the computational cost and time of parameter estimation in future.

2.7 Acknowledgement

Author would like to acknowledge Dr. Sanjoy Das for providing the initial concept on the method used in this study. Furthermore, we would like to thank service climatologist of Department of Agronomy, KSU, Mrs. Mary Knapp for her help on accessing weather data.

2.8 References

Alagarswamy, G., U. Singh, and D. Godwin. 1989. Modeling nitrogen uptake and response in sorghum and pearl millet. In: S.M. Virmani, H.L.S. Tandon, and G. Alagarswamy, editors, Modeling the growth and development of sorghum and pearl millet. Res. Bull.12. ICRISAT, Patancheru, Andhra Pradesh. http://oar.icrisat.org/955/ (accessed 10 Feb. 2015). p. 11–12.

Alagarswamy, G., Ritchie, J.T., 1991. Phasic development in CERES-Sorghum model. Predict. Crop Phenol. 143–152.

Assefa, Y., Roozeboom, K.L., Thompson, C., Schlegel, A., Stone, L., Lingenfelser, J., 2013. Corn and Grain Sorghum Comparison: All Things Considered. Academic Press.

Baumhardt, R.L., Tolk, J.A., Winter, S.R., 2005. Seeding practices and cultivar maturity effects on simulated dryland grain sorghum yield. Agron. J. 97, 935–942.

Bert, F.E., Laciana, C.E., Podestá, G.P., Satorre, E.H., Menéndez, A.N., 2007. Sensitivity of CERES-Maize simulated yields to uncertainty in soil properties and daily solar radiation. Agric. Syst. 94, 141–150.

Borgonovo, E., Tarantola, S., 2008. Moment independent and variance-based sensitivity analysis with correlations: An application to the stability of a chemical reactor. Int. J. Chem. Kinet. 40, 687–698.

Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environ. Model. Softw. 22, 1509–1518.

Campolongo, F., Saltelli, A., 1997. Sensitivity analysis of an environmental model: an application of different analysis methods. Reliab. Eng. Syst. Saf. 57, 49–69.

Christopher Frey, H., Patil, S.R., 2002. Identification and review of sensitivity analysis methods. Risk Anal. 22, 553–578.

Confalonieri, R., Bellocchi, G., Bregaglio, S., Donatelli, M., Acutis, M., 2010. Comparison of sensitivity analysis techniques: A case study with the rice model WARM. Ecol. Model. 221, 1897–1906. doi:10.1016/j.ecolmodel.2010.04.021

Confalonieri, R., Bregaglio, S., Cappelli, G., Francone, C., Carpani, M., Acutis, M., Zhiming, W., 2012. SENSITIVITY ANALYSIS.

Conley, S.P., Stevens, W., Dunn, D.D., 2005. Grain sorghum response to row spacing, plant density, and planter skips. Crop Manag. 4, 0–0.

Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., Schaibly, J.H., 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. J. Chem. Phys. 59, 3873–3878.

DeJonge, K.C., Ascough II, J.C., Ahmadi, M., Andales, A.A., Arabi, M., 2012. Global sensitivity and uncertainty analysis of a dynamic agroecosystem model under different irrigation treatments. Ecol. Model. 231, 113–125. doi:10.1016/j.ecolmodel.2012.01.024

Folliard, A., Traoré, P.C.S., Vaksmann, M., Kouressy, M., 2004. Modeling of sorghum response to photoperiod: a threshold–hyperbolic approach. Field Crops Res. 89, 59–70.

Hammer, G.L., Kropff, M.J., Sinclair, T.R., Porter, J.R., 2002. Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. Eur. J. Agron., Process Simulation and Application of Cropping System Models 18, 15–31. doi:10.1016/S1161-0301(02)00093-X

Helton, J.C., Johnson, J.D., Sallaberry, C.J., Storlie, C.B., 2006. Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliab. Eng. Syst. Saf. 91, 1175–1209.

Hoogenboom, G., Jones, J.W., Wilkens, P.W., Porter, C.H., Hunt, L.A., Singh, U., Lizaso, I., White, J., Uryasev, O., Ogoshi, R.M., Koo, J., Shelia, V., Tsuji, G.Y., 2015. Decision Support System for Agrotechnology Transfer (DSSAT) version 4.5 (htttp://dssat.net)., DSSAT Foundation. Prosser, Washington.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. Eur. J. Agron. 18, 235–265.

Jones, J.W., Naab, J., Fatondji, D., Dzotsi, K., Adiku, S., He, J., 2012. Uncertainties in simulating crop performance in degraded soils and low input production systems, in: Improving Soil Fertility Recommendations in Africa Using the Decision Support System for Agrotechnology Transfer (DSSAT). Springer, pp. 43–59.

Krishnaiah, P.., 1981. Analysis of variance (Hand book of Statistics). Elsevier, New York.

Lobell, D.B., Ortiz-Monasterio, J.I., 2007. Impacts of day versus night temperatures on spring wheat yields. Agron. J. 99, 469–477.

McKay, M.D., 1997. Nonparametric variance-based methods of assessing uncertainty importance. Reliab. Eng. Syst. Saf. 57, 267–279.

McRae, G.J., Tilden, J.W., Seinfeld, J.H., 1982. Global sensitivity analysis—a computational implementation of the Fourier amplitude sensitivity test (FAST). Comput. Chem. Eng. 6, 15–25.

Mokhtari, A., Frey, H.C., Durham, N.C., 2005. Review and Recommendation of Methods for Sensitivity and Uncertainty Analysis for the Stochastic Human Exposure and Dose Simulation (SHEDS) Models. Alion Sci. Technol. Durh. NC 1, 1–99.

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

Mutava, R.N., Prasad, P.V.V., Tuinstra, M.R., Kofoid, K.D., Yu, J., 2011. Characterization of sorghum genotypes for traits related to drought tolerance. Field Crops Res. 123, 10–18.

Park, C.K., Ahn, K.-I., 1994. A new approach for measuring uncertainty importance and distributional sensitivity in probabilistic safety assessment. Reliab. Eng. Syst. Saf. 46, 253–261.

Peck, R., Devore, J.L., 2011. Statistics: The exploration & analysis of data. Cengage Learning.

Quiring, S.M., Legates, D.R., 2008. Application of CERES-Maize for within-season prediction of rainfed corn yields in Delaware, USA. Agric. For. Meteorol. 148, 964–975.

Reddy, R.N., Madhusudhana, R., Mohan, S.M., Chakravarthi, D.V.N., Mehtre, S.P., Seetharama, N., Patil, J.V., 2013. Mapping QTL for grain yield and other agronomic traits in post-rainy sorghum [Sorghum bicolor (L.) Moench]. Theor. Appl. Genet. 126, 1921–1939.

Ritchie, J., 1998. Soil water balance and plant water stress, in: Understanding Options for Agricultural Production. Springer, pp. 41–54.

Ritchie, J.T., Alagarswamy, G., 1989a. Genetic Coefficient for the CERES Models, in: Modeling the Growth and Development of Sorghum and Pearl Millet. ICRISAT, India, p. 27.

Ritchie, J.T., Alagarswamy, G., 1989b. Simulation of Sorghum and Pearl Millet Phenology, in: Modeling the Growth and Development of Sorghum and Pearl Millet. ICRISAT, India.

Rosenzweig, C., 1990. CROP RESPONSE TO CLIMATE CHANGE IN THE SOUTHERN GREAT PLAINS: A SIMULATION STUDY∗. Prof. Geogr. 42, 20–37.

Saltelli, A., Chan, K., Scott, E.M., others, 2000. Sensitivity analysis. Wiley New York.

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

Saseendran, S.A., Ahuja, L.R., Nielsen, D.C., Trout, T.J., Ma, L., 2008. Use of crop simulation models to evaluate limited irrigation management options for corn in a semiarid environment. Water Resour. Res. 44, W00E02. doi:10.1029/2007WR006181

Semenov, M.A., Shewry, P.R., 2011. Modelling predicts that heat stress, not drought, will increase vulnerability of wheat in Europe. Sci. Rep. 1.

Singh, U., Ritchie, J.T., Alagarswamy, G., 1993. A User's Guide to CERES Sorghum, V2. 10. International Fertilizer Development Center.

Sobol, I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. 55, 271–280.

Spear, R.C., Hornberger, G.M., 1980. Eutrophication in Peel Inlet—II. Identification of critical uncertainties via generalized sensitivity analysis. Water Res. 14, 43–49.

Staggenborg, S.A., Vanderlip, R.L., 2005. Crop simulation models can be used as dryland cropping systems research tools. Agron. J. 97, 378–384.

Taranatola, S., Salteli, A., 2003. Methodological Advances and Innovative Applications of Sensitivity Analysis. Reliab Eng Syst Saf SAMO 79, 121–122.

Tomilnson, j. P., Knapp, M., 2012. Introduction and Kansas Climate Overview, in: Efficient Crop Water Use in Kansas. K-State Research and Extension, Kansas, USA, pp. 1–2.

Tondel, K., Vik, J.O., Martens, H., Indahl, U.G., Smith, N., Omholt, S.W., 2013. Hierarchical multivariate regression-based sensitivity analysis reveals complex parameter interaction patterns in dynamic models. Chemom. Intell. Lab. Syst. 120, 25–41.

Varella, H., Guérif, M., Buis, S., 2010. Global sensitivity analysis measures the quality of parameter estimation: The case of soil parameters and a crop model. Environ. Model. Softw. 25, 310–319.

Virmani, S.M., Tandon, H.L.S., Alagarswamy, G., 1989. Modeling the growth and development of sorghum and pearl millet. International Crops Research Institute for the Semi-Arid Tropics.

Wang, J., Li, X., Lu, L., Fang, F., 2013. Parameter sensitivity analysis of crop growth models based on the extended Fourier Amplitude Sensitivity Test method. Environ. Model. Softw. 48, 171–182.

White, J., 2009. Combining ecophysiological models and genomics to decipher the GEM-to-P problem. NJAS-Wagening. J. Life Sci. 57, 53–58.

White, J.W., Hoogenboom, G., Hunt, L.A., 2005. A structured procedure for assessing how crop models respond to temperature. Agron. J. 97, 426–439.

Xie, Y., Kiniry, J.R., Nedbalek, V., Rosenthal, W.D., 2001. Maize and sorghum simulations with CERES-Maize, SORKAM, and ALMANAC under water-limiting conditions. Agron. J. 93, 1148–1155.

Xu, C., Gertner, G., 2011. Understanding and comparisons of different sampling approaches for the Fourier Amplitudes Sensitivity Test (FAST). Comput. Stat. Data Anal. 55, 184–198.

Zhao, G., Bryan, B.A., Song, X., 2014. Sensitivity and uncertainty analysis of the APSIM-wheat model: Interactions between cultivar, environmental, and management parameters. Ecol. Model. 279, 1–11.

Ziaei, A.N., Sepaskhah, A.R., 2003. Model for simulation of winter wheat yield under dryland and irrigated conditions. Agric. Water Manag. 58, 1–17.

Table 2.1 Detail description of experimental sites.

	Ottawa		**Hutchinson**		**Hays**		**Tribune**	
Soil type	Woodson Silt loam		OST loam		Harney Silt loam		Ulysses Silt loam	
Lat/long	45.2N,75.69W		38.06N,97.92W		39.87N,99.32W		38.47N,101.75W	
Elevation (m)	205		470		609		1101	
PDATE	5/29		5/7		6/8		6/16	
Fertilizer (N/acre)	135		112		90		100	
Row Spacing (cm)	75		75		75		75	
	Dry	**Wet**	**Dry**	**Wet**	**Dry**	**Wet**	**Dry**	**Wet**
Year	1988	1993	1994	1993	1994	1993	2002	1997
Precipitation (mm)	4677	10624	3319	8996	2840	7672	2285	5222

Table 2.2 CERES-Sorghum input parameters and output responses for sensitivity analysis (SA).

SN	Variable	Definition	Unit	Lower Bound	Upper Bound
		Genotype-Specific Parameters			
1	**P1**	Thermal time from emergence to end of juvenile phase	Degree day	150	500
2	**P2**	Thermal time from end of juvenile stage to tassel initiation	Degree day	90	110
3	**P2O**	Critical Photoperiod hour	Hour	11	16
4	**P5**	Thermal time from flowering to physiological maturity	Degree day	400	700
5	**PHINT**	Phylochron interval	Degree day	30	90
6	**G1**	Leaf size Coefficient	-	0	30
7	**G2**	Panicle Size partitioning coefficient	-	4	7
8	**TBASE**	Base temperature	^{o}C	4	9
9	**RUE**	Radiation use efficiency	g MJ^{-1}	3	6
		Soil specific parameter			
10	**SLPF**	Soil fertility factor	-	0.7	1.0
11	**SLU1**	Evaporation limit	cm	5	12
12	**SLDR**	Drainage rate	Day^{-1}	0	1
13	**SLRO**	Runoff curve number	-	60	95
14	**DLL**	Drained lower limit (wilting point)	Mm^3mm^{-3}	0.11	0.20
15	**DUL**	Drained upper limit (field capacity)	Mm^3mm^{-3}	0.25	0.42
16	**SSAT**	Saturated water limit	Mm^3mm^{-3}	0.42	0.51
17	**SSKS**	Saturated hydraulic conductivity	cmh^{-1}	0.3	2.0
18	**SBDM**	Bulk density	g cm^{-3}	1.2	1.5
19	**SLOC**	Soil organic carbon	%	0.5	2.0
		Agronomic management parameter			
20	**PPOP**	Plant population	Number m^{-2}	10	20
Output Variable					
1	**Yield**	Grain yield	Kg ha^{-1}		
2	**ADAT**	Anthesis days	DAP		
3	**MDAT**	Maturity days	DAP		
4	**LAI**	Leaf area index	---		

Table 2.3 *P*-value from ANOVA test of sensitivity index between dry and wet years.

Output					PHI			TBA		SL	SL	SL	SL	SL	SD	SS	SS	SB	SL	PP
Var	**P1**	**P2**	**P2O**	**P5**	**NT**	**G1**	**G2**	**SE**	**RUE**	**PF**	**U1**	**DR**	**RO**	**L**	**UL**	**AT**	**KS**	**DM**	**OC**	**OP**
Yield	.16	.26	**.04**	.76	.21	.61	.26	.33	.23	.30	.28	.52	**.01**	**.01**	.14	.34	.25	.32	.26	.28
ADAT	.87	.54	.21	.33	**.03**	.33	.92	.22	.77	.07	.50	.19	.79	.92	.48	.22	.90	.29	.56	.85
MDAT	.24	.60	**.03**	.07	.35	.40	.67	.39	.25	**.03**	.58	.38	.62	.05	.18	.32	.75	.77	.79	.95
LAI	***.00***	.10	***.00***	***.01***	***.00***	.07	.09	**.03**	**.02**	**.02**	**.03**	.73	.27	***.00***	**.04**	.14	.06	.20	.57	.22

Note: Bold & Italic = 99% significant, Bold: 95% significant

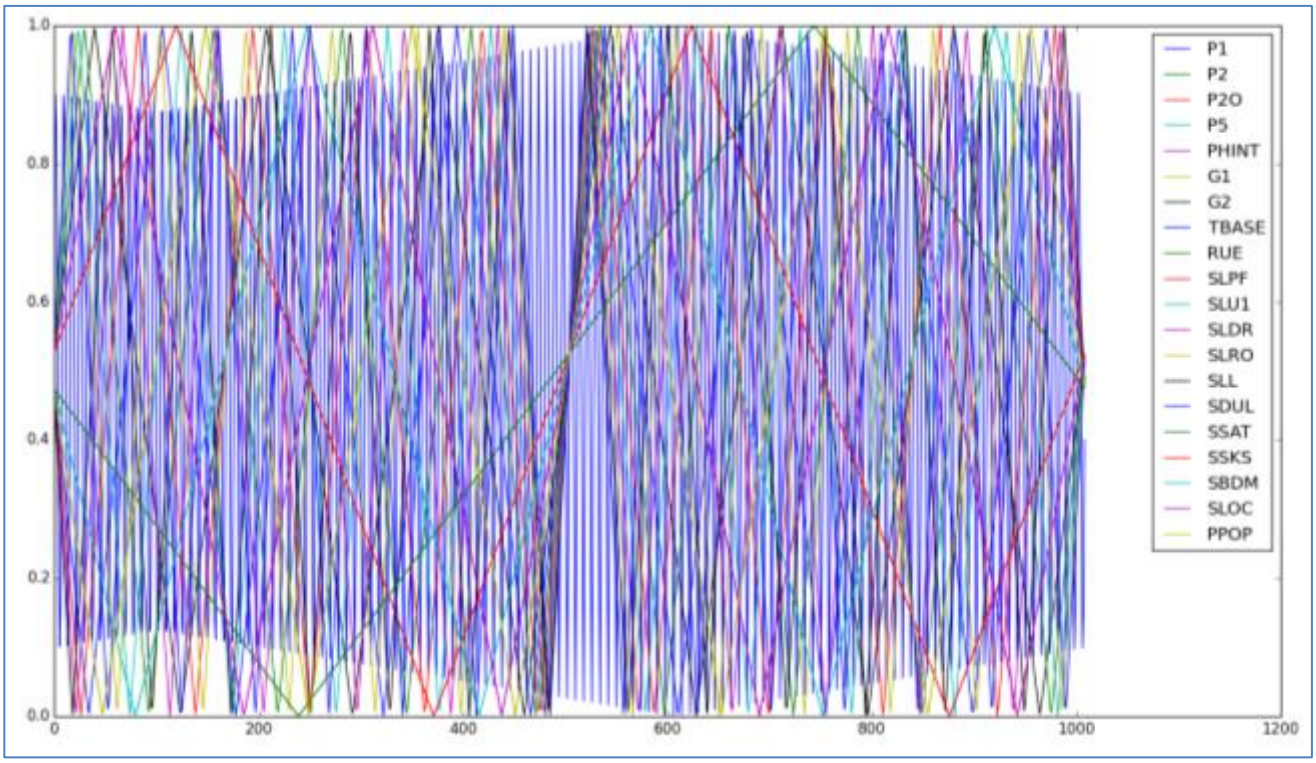

Fig. 2.1 Sampling curves generated using eFAST for all 20 parameters used in the study.

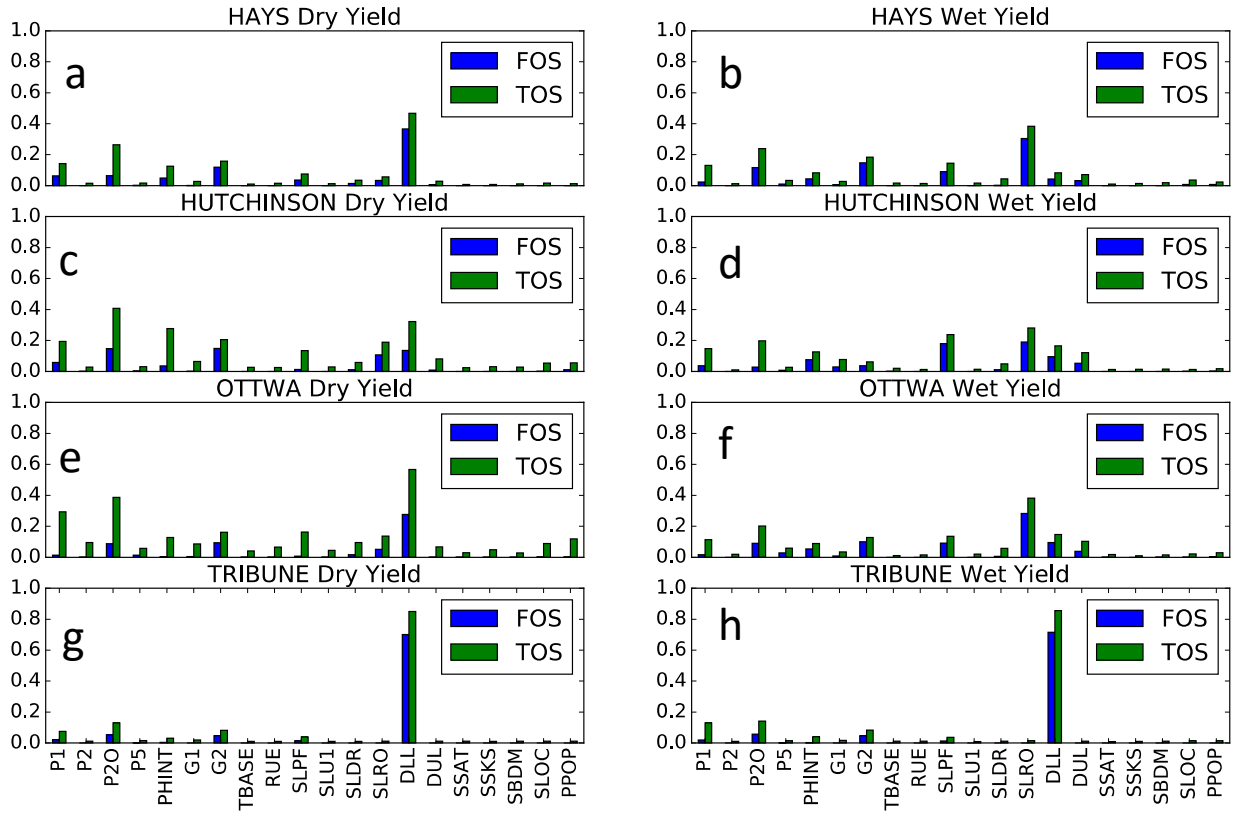

Fig. 2.2 First Order Sensitivity (S_i) and Total Sensitivity (ST_i) indices for CERES-Sorghum input parameter in response to grain yield.

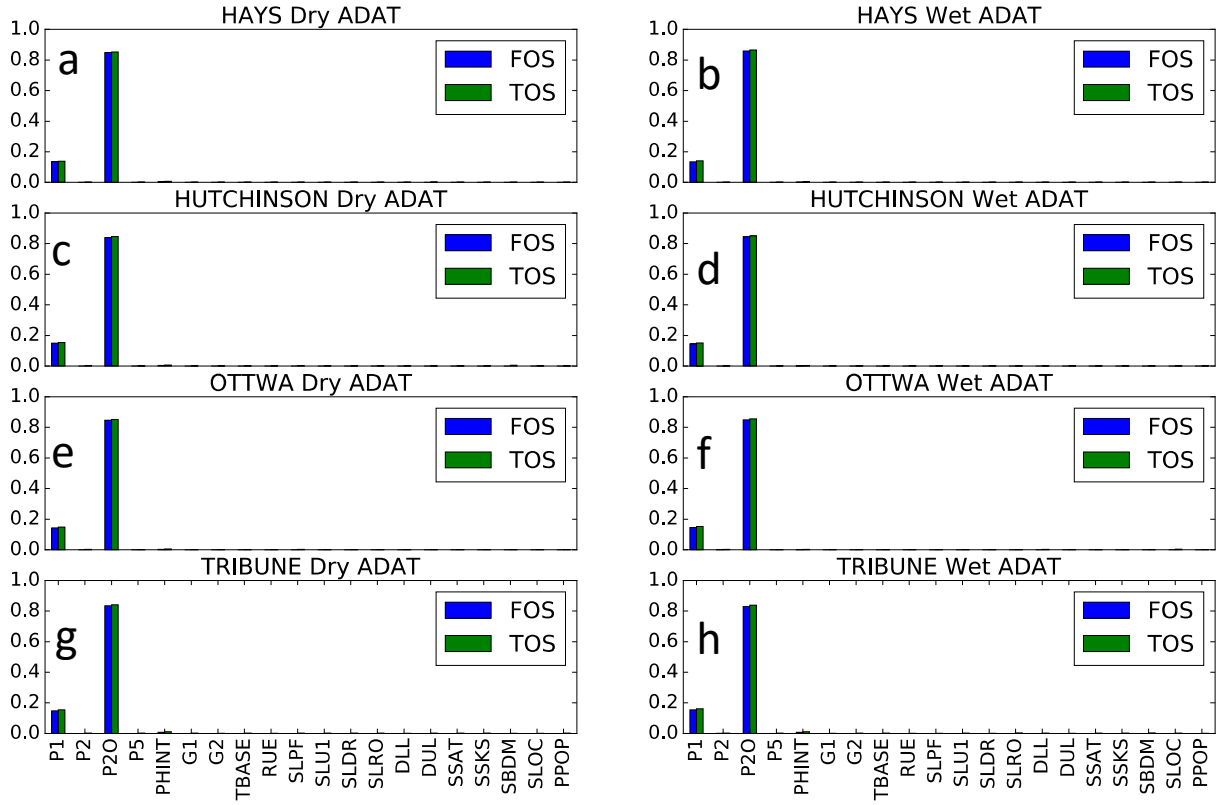

Fig. 2.3 First Order Sensitivity (S_i) and Total Sensitivity (ST_i) indices for CERES-Sorghum input parameter in response to anthesis days (ADAT).

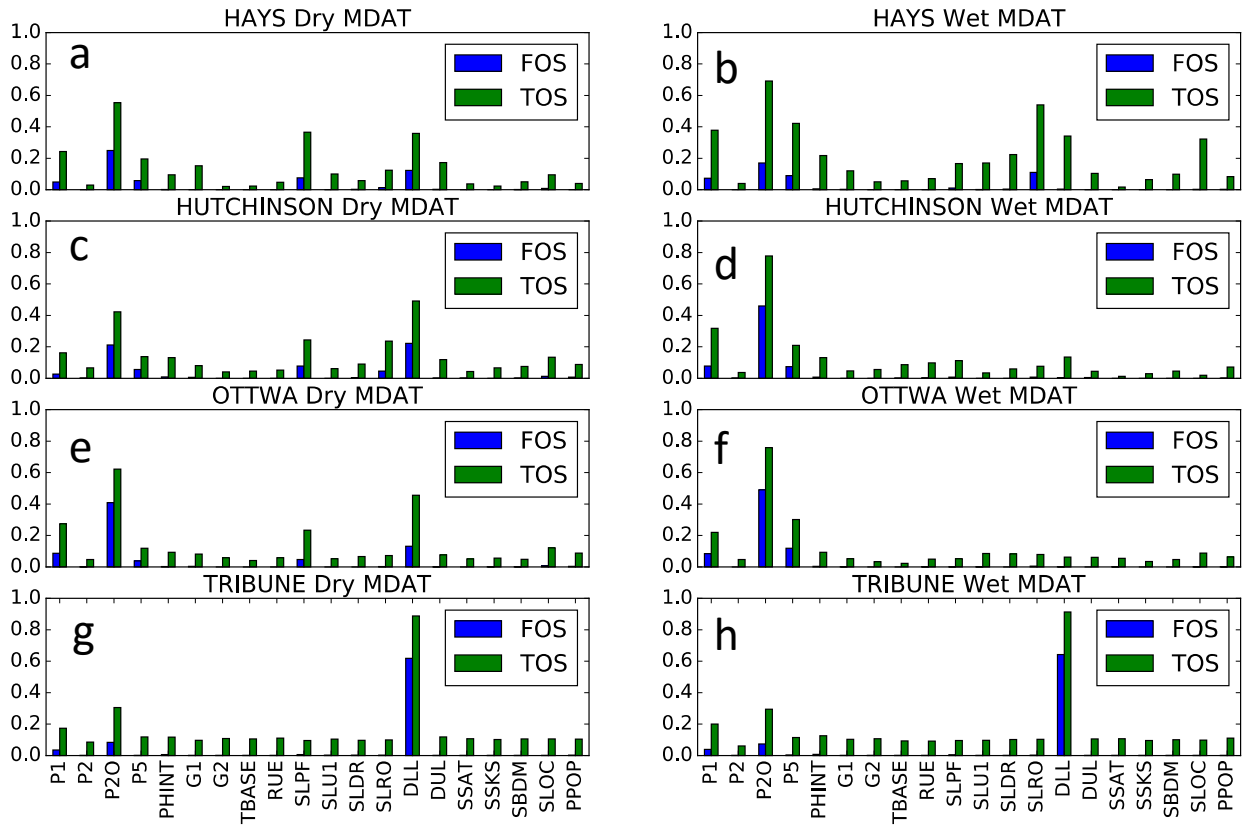

Fig. 2.4 First Order Sensitivity (S_i) and Total Sensitivity (ST_i) indices for CERES-Sorghum input parameter in response to Maturity Days (MDAT).

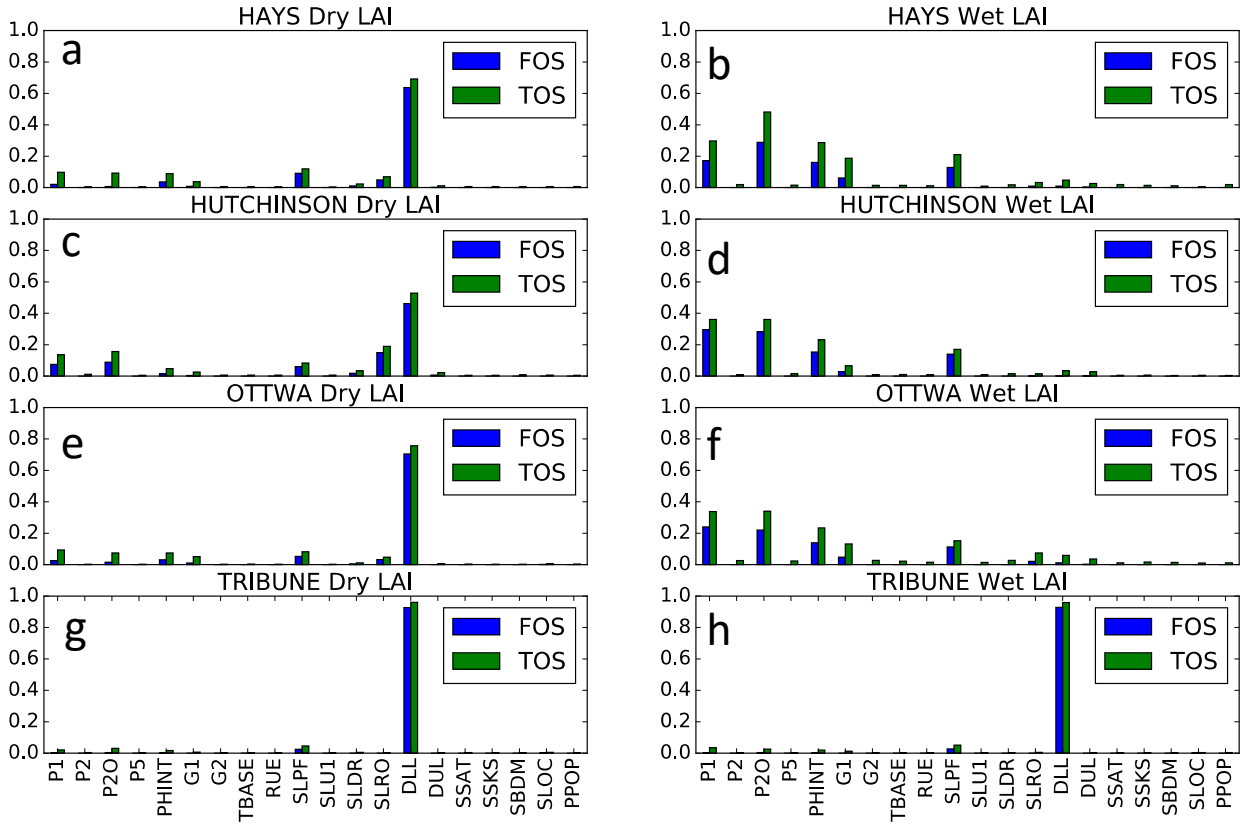

Fig. 2.5 First Order Sensitivity (S_i) and Total Sensitivity (ST_i) indices for CERES-Sorghum input parameters in response to leaf area index (LAI).

CHAPTER 3 - EFFICIENT CROP MODEL PARAMETER ESTIMATION AND SITE CHARACTERIZATION USING LARGE BREEDING TRIAL DATA SETS

Abstract

Global crop production needs to double by 2050 to supply the demand for food, feed, and fuel. To reach this goal, novel methods are needed to increase breeding rates of gain as well as on-farm yields through enhanced management strategies. Both of these tasks require the ability to predict plant performance in multiple, dynamic environments based on a knowledge of cultivar characteristics (critical short day lengths, maximum leaf photosynthetic rates, pod fill durations, etc.) that are ultimately linked to genetics. Because of this linkage, we refer to such traits as genotype-specific parameters (GSP's). Using industry-provided yield and weather data from 353 site-years, we estimated seven primary CROPGRO-Soybean GSP's for each of 182 varieties. The data set had two shortcomings. First, no planting dates were supplied, rendering unknowable the environment actually experienced by the crop. Second, soil data were provided only for the top 20 cm, which is inadequate to specify the root environment. Therefore, additional soil information was acquired. A novel optimization algorithm was developed that simultaneously estimated GSP's and planting dates, while tuning layered soil water holding properties. The optimizer, which we have named the holographic genetic algorithm (HGA), used both externally supplied constraints and its own analysis of data structure to reduce what would otherwise be a search over 3000+ dimensions to a much smaller number of overlapping 1- to 3-D problems. Two types of runs were performed. The first was preceded by an independent component analysis (ICA) of published GSP's. The subsequent training sought good component scores rather than the GSP's themselves.

The second allowed all GSP's to vary separately. This makes the parameters less constrained and more evenly distributed than ICA. Results showed that HGA works quite well with the CROPGRO-Soybean model to estimate the cultivar and site-specific parameters from breeding trial data. The quality of the calibrations and evaluations were similar across both run types with RMSE values being ca. 5.2x % of the maximum yields. Moreover, the GSP's for a variety can be used to predict its yield in trials not used in that cultivar's calibration. Finally, despite high dimensionality, GSP's, planting dates, and soil properties for all lines and sites converged concurrently in <58 iterations, demonstrating great utility for use with big data sets.

3.1 Introduction

Scientists currently estimate that global crop production must double by 2050 to meet the world's need for food, fiber, and fuel resources (Ray et al., 2013). To meet this enormous challenge, novel crop improvement and management methods are needed. Central to this task are reliable quantitative methods for predicting the behavior of differing crop cultivars in novel, time-varying environments. In the context of this report, *time-varying* refers to the time series of daily weather events, which varies from year to year at any location, within a year at different locations, and with either quasi-cyclic or secular trends on the scale of decades or longer. Ecophysiological crop growth simulation models provide a means to make such predictions. Crop models use differential equation-based descriptions of plant physiological processes (i.e. photosynthesis, transpiration, respiration, growth, development, and assimilate partitioning), along with chemical and physical processes (e.g. soil chemical transformations, energy flows, gas diffusion in leaves). Incorporating biotic and abiotic information helps to explain observed biological processes along with relationships between fixed plant properties and predicted response variables. Cropping system models have been used to explain processes at the level of genotype, crop, farming system,

region, and global environment (Matthews et al., 2002). Because crop models are process-based, they are powerful for predicting growth and yield in response to different environment and agronomic management practices, which opens the possibility of using models for crop improvement (White, 1998).

Currently, ecophysiological models exist for all major crops, many minor ones, and some weed species. CROPGRO-Soybean is an ecophysiological model that has been widely tested since its initial release by Wilkerson et al. (1983) and has been shown to predict accurate yield responses to weather and management for different maturity groups throughout the USA (Boote et al., 2003, 1997; Boote and Tollenaar, 1994). It has seen application in precision agriculture (Irmak et al., 2006, 2002; Paz et al., 2003), yield prediction (Jones et al., 1991; S. Welch et al., 2002), and water management (Calmon et al., 1999; Swaney et al., 1983), among other topics. It is one of the models in the Decision Support System for Agrotechnology Transfer (DSSAT) software suite (Jones et al., 2003). The model considers mechanistic carbon balance with photosynthesis inputs at leaf-level, with hedgerow canopy assimilation using canopy LAI, canopy height and width for light capture (Boote and Pickering, 1994). It was tested and shown by Alagarswamy et al. (2006) to accurately predict photosynthesis for soybean. The model has explicit nodule growth and N-fixation and considers energy costs for synthesis of protein and oil and other compounds in vegetative and seed structures (Boote et al., 1998; K. J. Boote et al., 2008). Addition of pods and seeds is based on assimilate supply and carrying capacity. The model considers N mobilization and canopy self-senescence as features during seed-filling. It has a well-tested phenology subroutine that is sensitive to day length and temperature that can mimic developmental maturity groups (MG) from 00 to IX (Grimm et al., 1994, 1993; Jones et al., 1991; Mavromatis et al., 2002, 2001).

The model has been tested intensively with time-series growth analyses as well as extensive final yield data sets under different environments (Boote et al., 1997; Piper et al., 1998; Sau et al., 1999) including insect defoliation (Timsina et al., 2007). It has been used in climate change studies beginning as early as 1995 (Curry et al., 1995). The model uses a tipping bucket soil water balance based on (Ritchie, 1998) and its sensitivity to water stress was reviewed by Boote et al. (1998) and Boote et al. (2008). The ability of the soybean model to accurately simulate N-fixation was documented by Boote et al. (2008) and Sexton et al. (1998). Use of CROPGRO-Soybean to study genetic improvement in soybean yield has been illustrated in several papers (Boote et al., 2003, 2001; Boote, 2011; Boote and Tollenaar, 1994).

The model encodes the responses of different genetic lines to the environment, nutritional, and management conditions via a set of numeric constants. Called *genotype-specific-parameters* (GSP's), these constants are organized in a hierarchical fashion wherein some 18 parameters are specific to each cultivar and an additional 16 describe different ecotypes. These fixed, innate traits specify the sensitivities of soybean crop processes to environmental factors such as temperature, solar radiation, carbon dioxide, N, as well as plant initializations and tissue compositions. The cultivar traits vary frequently across lines, whereas ecotype traits are more stable and describe groups of cultivars with similar behaviors.

With the advances in plant genomics and the falling costs of locating genetic markers, efforts are being made to link the GSP's to actual genes and/or quantitative trait loci (QTLs), not just for soybean (Boote et al., 2003; Messina et al., 2006; Wilczek et al., 2009) but for many other crops as well (Cooper et al., 2016; Hammer et al., 2006; Technow et al., 2015). For example, recent research in common bean has made progress in developing a gene-based ecophysiological model for common bean, based on phenotyping and genotyping of 190 recombinant inbred lines, creating

QTL-based modules for rate of leaf appearance, leaf area expansion, and progress toward anthesis and maturity (Boote et al., 2016).

However, obtaining sufficient data to quantify model constants has been an issue from the earliest days of crop simulation. Direct measurement of so many traits for more than a few varieties has never been practical. Various papers have demonstrated indirect GSP estimation from field data on phenology, yield, and seed size (Alderman et al., 2015; Mavromatis et al., 2002, 2001; Pathak et al., 2012; Welch et al., 2002) or reproducing gene effects on crop development and yield (Messina et al., 2006). When soil data was inadequate, edaphic properties had to be estimated simultaneously, a process that can limit predictive skill depending on the amount of data available (Welch et al., 2002; Wilkerson et al., 1983). Over time, interest in high throughput phenotyping and, perhaps, better soil sensors will accelerate the acquisition of needed data. However, at the present time, there are relatively few large data sets that have been used in estimation studies. Mavromatis et al. (2002) used 393 location-year-line combinations and Welch et al., (2002) used 1155 location-year-lines. Here we report results of calibrating and testing the CROPGRO-Soybean model employing a new algorithm specially designed for use with large multi-location breeding trials. The test data set (7426 location-year-lines) was posted by Syngenta AG as part of a predictive soybean modeling contest conducted in 2015-16 in collaboration with the Institute for Operations Research and the Management Sciences (INFORMS).

3.2 Methodology and theory

Crop models have been used to successfully predict crop yield over many global locations and climates but their accuracy is dependent on the quality of their environmental inputs and the accuracy and completeness with which the plant material is characterized. The latter is a driving force behind efforts at developing *high throughput phenotyping* and *inverse modeling* methods of

estimating plant physiological traits from the resulting data. Basically, uncertainty and prediction error are minimized by improving how individual processes are represented mathematically and by accurately estimating the GSP's embodied in the model equations (White, 2009). Inverse modeling is an automated training procedure in which estimates of crop characteristics (GSP's) are iteratively adjusted to bring model behavior into alignment with observed yields and/or other measurements. To make accurate parameter estimations, we followed five basic procedures given by Welch et al. (2002).

1. Minimum data required to run model simulations are crucial. Therefore, because the information provided was insufficient to run the model, additional public data were collated. These included more extensive soil data and geographically adjusted ranges of likely planting dates.
2. Computer programs were developed to automate the preparation of model input data, thus greatly reducing the amount of manual labor entailed along with the potential for human error given the large number of cultivars and site-years.
3. A suitable optimization algorithm was developed that exploited both known biological constraints on model parameters and the implicit constraints resulting from the structure of the data. As an example of the latter, the maximum potential leaf photosynthetic rate of a specific variety is a genetically determined trait, and so the algorithm constrained it to be the same at all sites where the variety is grown. (Of course, a variety's *realized* photosynthetic rates will differ across sites as influenced by local environments in ways described mathematically within the model.)
4. Due to the volume of the data and the large number of model runs required in the estimation process large-scale parallel processing was applied.
5. Finally, the calibration quality was evaluated by comparing predicted and observed yields for site-year-line combinations not used in the calibration process.

3.2.1 Assembling the minimum data set

The CROPGRO-Soybean model requires information on soil environment, daily weather data, agronomic management practices, and genetic information to develop yield prediction. We

used the given information augmented along with additional publicly available data to develop a minimum data set needed for model runs.

3.2.2 Soil Data

DSSAT models the soil as a series of layers, each of which is characterized by parameters that govern water holding capacities and suitability for root growth. These include the drained upper limit (DUL), the drained lower limit (DLL), saturated water content (SAT), and the soil root growth factor (SRGF). DLL and DUL are required to establish how much water the soil will hold by capillarity, and how much will drain out to gravity. In addition, surface runoff curve numbers (SLRO) control how much of a heavy rain actually infiltrates the soil. Unfortunately, the soil data supplied by Syngenta only described the top 20 cm, which is inadequate to specify fully the environments of roots that can extend to depths of 2 m or more.

Given accurate location information, the needed values can be estimated based on publicly available data but there was also an impediment in that regard. The provided geographic coordinates of field sites were only guaranteed to be within 1 km of the actual field location but soils can be quite variable over such distances. Indeed, in some cases (e.g., Fig. 3.1) the geographic coordinates did not even map to agricultural sites. Additionally, a *site* as described in the data was not a particular field but, rather, a set of locations in proximity to one another that were used in different years.

Therefore, as a first step, Google Earth was used to improve the accuracy of as many locations as possible by exploiting the highly diagnostic appearance of breeding trial plot structures. Google Earth does not image all surface points each year but we were able to find the exact locations for 57% of the 354 site-years of data provided. There were also some instances where exact imagery was not found on particular year but found on next or previous year within

the given 1km of radius and similar soil map unit. These locations, which comprised about 16% of site-years, were used as proxies for the unobserved fields. The remaining 27% site-years were not found. In what follows, the three different location categories are, respectively, called “Exact”, “Exact-Conditional” and “Missing”. The supplied geographic coordinates were used for the Missing soil location types. Tools provided within Google Earth (e.g. KMZ files) were used to automate the handling of location data.

With better location data in hand, soil information was obtained via IBM's PAIRS Technology (Klein et al., 2015) with data from SSURGO (SSURGO, 2015). Soil texture (percentage of sand, silt and clay), bulk density, soil organic matter, hydraulic conductivity, pH, cation exchange capacity was extracted for each layer. Using these data DUL, SAT, and SLRO were estimated via the pedotransfer functions in Cronshey (1986), Saxton and Rawls (2006), and Singh et al. (2014). DLL and SGRF values were obtained as part of the estimation process.

3.2.3 Agronomic Management and Weather Data

While the data provided by Syngenta was comprehensive, it provided few management details. We assumed that all sites were rainfed, and used statewide averages for plant populations and row spacing. Planting dates were estimated during the model calibration stage.

We used the weather data that was provided by Syngenta, which supplied all needed weather variables for CROPGRO-Soybean, in particular daily maximum and minimum air temperature, precipitation, and solar radiation.

3.2.4 Genotype Specific Parameters

As shown in Table 3.1, CROPGRO-Soybean represents the characteristics of individual cultivars in terms of 18 genotypic specific parameters (GSPs). It would be quite demanding to try to estimate all of these independently for each variety. This is especially true because there are

tradeoffs between them such that raising the value of some can be offset by lowering the value of others, creating a lack of parameter identifiability. However, biological experience with the soybean modeling has resulted a degree of understanding about parameters relationships.. For example, not all life-cycle intervals scale uniformly. For example, as the interval from first seed to physiological maturity (SDPM) increases, some subintervals (e.g. seed fill duration, SFDUR) do as well but others (e.g. the time between first flower and first pod, FLSH) either do not or such changes as they might undergo have little effect on yield. In this study, we have chosen 7 important parameters for estimation. Among them, CSDL is the day length sensitive traits which account the influence of day length on growth of soybean. EMFL, FLSD and SDPM are the important life cycle "phase" durations determining traits, LFMAX represents the maximum photosynthetic rate, and SFDUR, PODUR represents the reproductive traits of Soybean affecting grain yield. The latter might as well be set to constants, and the former can be approximated as linear functions within the intervals containing them.

X-factor approach used in this study was first introduced by Mavromatis et al. (2001) and Welch et al. (2002). The idea of this approach is to reduce the parameter search space dimensionality by making groups of parameters into a linear function of these X factors.

The objective of this study was to examine two different approaches to estimate cultivar coefficients, which are based on i) independent component analysis (ICA) (Comon, 1992) and ii) treating each GSP factor separately (SF)

3.2.4.1 ICA Approach

ICA is a computational method that is used to identify and separate independent hidden factors that are linearly mixed in variables. ICA is more useful when data are non-Gaussian. It separates the multivariate signal to independent separate factors. Thus, ICA was used to search for

factors that might underlie CROPGRO-Soybean GSP's. All the component obtained from ICA are statistically independent factors. Total of 82 already published CROPGRO-Soybean GSP's from different parts of the world were collected. ICA was applied to this data set for each model parameter and the results are shown in Table 3.1.

3.2.4.2 Separate Factor (SF) approach

In this approach, each of the seven target parameters were allowed to vary separately. The allowed ranges were the same as in the ICA approach. In order to simplify the notation to be used below in describing the optimizer (see below) X-factor designations (X8-X10) were also given to the variables driving the search for planting dates and selected soil characteristics. These are discussed next.

3.2.5 Planting Date

To constrain planting date searches, crop insurance deadlines were used for each state. These dates specify the earliest and latest planting dates that farmers must follow to obtain insurance coverage. These dates were deemed to be reasonable, realistic constraints because of the importance of insurance in crop production and, therefore, the grower incentives to qualify. We also examined public sources of planting date information for each location including university websites, the USDA Risk Management Agency, and private insurance providers using Google searches. This helped to confirm and narrow down the planting date range obtained from crop insurance for most of the locations

3.2.6 Estimated soil characteristics

The soil root growth factor (SRGF) is a scalar whose variation with depth influences the spatial distribution of roots as their simulated development progresses (Singh et al., 2014). In real plants the realized root distribution depends on a detailed interaction between soil mechanical,

nutrient, and water status and plant allocation of growth resources to different morphological components. However, CROPGRO-Soybean summarizes this complexity into a single curve, whose shape is defined by a governing mathematical curve and associated parameter. A variety of curves have been used to compute SRGF. A common one is:

$$\mathrm{SRGF}(\mathrm{L}) = \left[1.0 - \frac{Z(\mathrm{L})}{5}\right]^{X_9}$$

where Z(L) is the depth midpoint of soil layer L and X9 is a parameter to be estimated from data. Typical value of X9 is $5.5 \leq X9 \leq 6.5$. A problem with this formula is that it declines monotonically with depth, whereas real soybean plants often have relatively constant root densities near the surface. Therefore, the function was modified to reflect this behavior in the top 30 cm. The exponent above was set to six and the curve was reparametrized to express the soil-specific depth (cm) below which it would be highly unlikely to ever find soybean roots; that is, the rooting depth (X9). The new expression and a graph for different values of X9 are shown in Fig. 3.2.

$$SRGF = \begin{cases} 1 & if\ 0.0 \leq d \leq 0.3 \\ a\left[1 - \dfrac{(d - 0.3)}{5}\right]^6 + b & if\ 0.3 < d \leq X9 \\ 0 & if\ X^9 < d \end{cases}$$

where,

$$b = \frac{\left\{0.02 - \left[1 - \dfrac{X9 - 0.3}{5}\right]^6\right\}}{\left\{1.00 - \left[1 - \dfrac{X9 - 0.3}{5}\right]^6\right\}}$$

$$a = 1 - b$$

d=Depth of the midpoint of soil layers in meters

X9= Rooting depth of the soil in meters

The second estimated parameter, X10, describes the soil water holding capacity. Soil water potential relates to the amount of energy required to extract the next unit of moisture from the soil with negative numbers indicating the greater difficulty that arises as soil dries. Atmospheres (i.e., bars) is one (of many) units used to quantify soil water potential. Typically, the range of potentials within which plants can extract water is between -0.3 and -15 bars. This range can also be delimited by the corresponding fractions of soil volume occupied by water. Soil texture (i.e., the percentages of sand, silt, and clay) strongly influences these limiting fractions. Thus, in very sandy soil, the *lower limit* (DLL; i.e., the volumetric water fraction at -15 bars) can be as small as 0.04. It will be much larger in clay soils that retain water more tenaciously. Note the fact that more water is present in clay soils at -15 bars (0.16) does not help the plant, which still has to expend the same (large) amount of energy to extract it. By analogy, the *drained upper limit* (DUL) is the volumetric water fraction present at -0.3 bars and the difference between them, DUL-DLL, is the key quantity of interest when modeling plant water availability.

Because CROPGRO-Soybean uses soil layers, the thickness of these layers must also be taken into account. This is because, even with the same value of (DUL-DLL), a thicker layer will hold more water than a thinner one. That is, the total profile water content is

$$\mathrm{TPWC} \equiv \sum_{i}^{\mathrm{L}} z_i \left(\mathrm{DUL}_i - \mathrm{DLL}_i \right)$$

where, z_i is the thickness (m) of layer *i*.

DUL, DLL, and SAT (the volumetric fraction of the soil occupied by water at saturation, a value used elsewhere in the model) can be roughly estimated from the soil texture using the pedotransfer functions in Saxton and Rawls (2006).

$$\text{DUL} = \theta_{33} = \theta_{33t} + \left[1.283\left(\theta_{33t}\right)^2 - 0.374\left(\theta_{33t}\right) - 0.015\right] \text{ where}$$
$$\theta_{33t} = -0.251S + 0.195C + 0.011OM + 0.006(S \times OM) - 0.027\left(C \times OM\right) + 0.452\left(S \times C\right) + 0.299$$

$$\text{DLL} = \theta_{1500} = \theta_{1500t} + \left(0.14 \times \theta_{1500t} - 0.02\right)$$
$$\theta_{1500t} = -0.024S + 0.487C + 0.006OM + 0.005\left(S \times OM\right) - 0.013\left(C \times OM\right) + 0.068\left(S \times C\right) + 0.031$$

$$\theta_{(\text{S}-33)} = \theta_{(\text{S}-33)t} + \left(0.636\theta_{(\text{S}-33)t} - 0.107\right)$$
$$\theta_{(\text{S}-33)t} = 0.278S + 0.034C + 0.022OM - 0.018\left(S \times OM\right) - 0.027\left(C \times OM\right) - 0.584\left(S \times C\right) + 0.078$$

$$\text{SAT} = \theta_{\text{S}} = \theta_{33} + \theta_{(\text{S}-33)} - 0.097S + 0.043$$

Here, S, C, and OM are the percentages by weight of sand, clay, and organic material for each layer, respectively.

However, pedotransfer functions are only approximations and, for refined work, it is necessary to tune their outputs, which is what X10 does. An initial estimate of total profile water content is $X_{10} = TPWC_0$, where the subscript "0" indicates a value calculated from the summation above using the Saxton and Rawls (2006) equations. Reasonable search limits are assumed to be, $\alpha TPWC_0 \leq X_{10} \leq \beta TPWC_0$ with $\alpha = 0.2$ and $\beta = 1.6$.

However, different trial values of X10 need to be converted into specific DUL and DLL values for each layer. CROPGRO-Soybean only considers the layer-by-layer (DUL-DLL) differences, so there is some freedom as to how to do this. However, the model creates more stress for water uptake when the DUL is quite high, for instance with a clay soil, even when (DUL-DLL) is the same. Therefore, X10 is mapped into proportionate changes in each layer's DLL such that the overall effect is to change total profile water content as desired.

That is, at each iteration, *j*, k_j is solved for in the equation $X_{10} = \sum_i^{\mathrm{L}} z_i \left[\mathrm{DUL}_{i,0} - \left(1 + k_j\right) \mathrm{DLL}_{i,0} \right]$, and used to calculate the values $\mathrm{DLL}_{i,j+1} = \left(1 + k_j\right) \mathrm{DLL}_{i,0}$. Naturally, k_j must be restrained so that $\Theta \le \mathrm{DLL}_{i,j+1} \le \mathrm{DUL}_{i,0}$, where $\Theta = 0.02$ was assumed to be the smallest plausible DLL. A minor issue is that these limits must be expressed as constraints on X10 rather than k_j. Some algebra reveals that

$$k_j = \left(\mathrm{TPWC}_0 - X_{10,j}\right) \Big/ \sum_i^L z_i \, \mathrm{DLL}_i$$

and

$$\max\left(\alpha \, \mathrm{TPWC}_0 \;,\; -\min_i \left(\frac{\mathrm{DUL}_{i,0} - \mathrm{DLL}_{i,0}}{\mathrm{DLL}_{i,0}} \right) \sum_i^L z_i \, \mathrm{DLL}_i + \mathrm{TPWC}_0 \right)$$

$$\le X_{10} \le$$

$$\min\left(\beta \, \mathrm{TPWC}_0 \;,\; -\max_i \left(\frac{\Theta - \mathrm{DLL}_{i,0}}{\mathrm{DLL}_{i,0}} \right) \sum_i^L z_i \, \mathrm{DLL}_i + \mathrm{TPWC}_0 \right)$$

where the outer limits on X10 only need be calculated only once during preprocessing.

The current frontier of genotype-to-phenotype modeling is discovering methods to merge statistical genetic approaches with those of ecophysiological modeling (Cooper et al., 2016; Hammer et al., 2006; Technow et al., 2015). This problem is being worked on from two directions. The first, the roots of which trace to Reymond et al. (2003 and Yin et al. (1999), is to use ecophysiological models in a genomic discovery mode. By fitting what are now referred to as GSPs to the individuals in mapping populations one can identify genomic regions that may contain genes controlling the traits individual GSPs quantify. The reverse of this, which dates to White and Hoogenboom (2010) is to express GSPs as linear functions of genetic states. When (*i*) the genetic states represent the possible results of (possibly only contemplated) crosses and (*ii*) one has weather, soil, and management data from some (possibly hypothetical) set of sites/conditions,

one can forecast the performance of novel genotypes in novel environments (Tardieu et al., 2005). The linkage between these forward and reverse approaches is that both involve expressing GSPs as the dependent variables in linear equations and using them in models.

There is another form of equality relationship that is crucially important to understand. It can be hard for field biologists to believe that there are *any* plant traits that do not vary with the environment. And yet, it must be so. When two individuals of a given variety are planted in different environments, they still have the same DNA, and, therefore, at some level of reductionism, have attributes that remain equal across environments. It is the modeling assertion that GSPs quantify that identity. Indeed, if a putative GSP is shown to vary with the environment then it instantly loses its status as a GSP and a new researchable question emerges, namely "What is the mechanism that causes it to vary?" Examples include a QTL analysis of specific leaf area (SLA) by Yin et al. (1999), which found that QTL of SLA are non stable across the environment suggesting estimation of GSP are affected due to G*E. In contrast, Reymond et al. (2003) found that estimation of GSP didn't suffer from G*E.

It may well be the case that some of the GSPs used herein might someday be demoted. Absent such a demonstration, however, the assumption is that their values are independent of the combination of environments used to estimate them.

A second equality assumption concerns planting dates. Breeding trials aim to expose all lines to exactly the same environment at each point in their respective life cycles. This is only possible when all lines at a site are planted on exactly the same day. In practice this is not always possible but the difference between the first and last line planted is seldom more than one or two days. It was therefore deemed an acceptable approximation to assume that all of the lines in a trial were planted on the same but unknown date.

The third assumption is on location. Out of 354 location-year lat-long, it was found that there were 344 different lat-long used only once and 10 which were repeated twice and not found on google imagery i.e. falls under missing category. Thus, we treated these 10 repeated locations as a separate 20 locations. This made all 354 locations –year as separate locations.

Putting these notions together, the estimation problem can be seen to have the structure of what is formally called a tri-partite graph. For any given variety (green circle), the objective is to find the set of X1-X7 values that best explain all the yields obtained across all the site-planting date combinations at which it was planted. For each blue circle, the goal is to find the planting date, X8, within the given year for which the varieties linked to it would have the yields they did. Finally, for each brown circle, the aim is to find the soil root growth factor distribution, X9, and profile available water capacity, X10, that produced the observed set of yields for the varieties planted there.

An ideal set of solutions will accomplish this simultaneously. Toward this end a novel algorithm was developed in which a set of optimizers, one per circle above, operates concurrently to achieve the desired result. Each optimizer is responsible for solving one part of the problem – for example, the GSPs of a particular variety or the date of a particular planting or the soil characteristics of a given site. In addition, each optimizer has access to all the data pertinent to its task. Thus, each soil property optimizer is aware of the yields of all the varieties that were planted there. Similarly, a GSP optimizer has access to all yield data for its variety no matter where or when that line was planted. This distributed pattern seemed somewhat analogous to a holographic plate, each small piece of which contains information about the whole scene. Therefore, the estimation scheme was named the *holographic genetic algorithm* (HGA).

Each individual optimizer is an implementation of the GENOCOP genetic algorithm (GA) (Michalewicz and Janikow, 1996). These authors showed that GA's offer a powerful approach to highly nonlinear problems have linear constraints, as is the case here. Overall, GENOCOP algorithm is a rather standard GA grounded in elite tournament selection. Its developers' watershed advance was to employ a set of ingenious random operators with the property that their application to feasible trial solution yields another feasible solution. Each of the circles in Fig. 3.3 contains a mechanistically exact copy of the GENCOP method.

HGA incorporates two novel features. The first is how multiple copies of this algorithm work together on problems structured as in Fig. 3.3. This is grounded in the manner by which the parameter vectors in the populations operated on by each GA are processed into the inputs for model runs. Consider one variety/planting_date/location (VPL) combination of present in the data – that is, one green-blue-brown sequence of three circles linked by black lines in Fig. 3.3. Any combination of parameter vectors from the corresponding three optimizers defines a model run that might be done. The number of such combinations is the product of the population sizes and would not be feasible for a data set this large. HGA uses a very simple approach to reduce this number. The corresponding populations are stacked next to other as shown in Fig. 3.4 where the horizontal stripes in each rectangle correspond to individual trial parameter vectors. Each resulting extended row constitutes one augmented parameter vector that generates one model run. Thus the total number of model runs per generation is only the product of the population size across the set of optimizers times the number of unique VPL combinations in the data.

As the algorithm was initially conceptualized it was recognized that tournament selections within each optimizer would make an important contribution to the exploration of the problem's very large parameter space. As the generations progressed a byproduct of tournament selection

would be that the vertical position of a surviving solution and its offspring change. The resulting horizontal alignments with diverse parameter vectors from the other optimizers increase exploration. As each optimizer progresses toward convergence, the vectors in its population will become more and more similar as will, therefore, the extended vectors. Thus, the whole ensemble of optimizers, it was thought, would converge with each element realizing its part of the overall solution.

Within the context of any individual problem, an ideal optimizer will properly balance parameter space exploration with *exploitation*, i.e., the ability to narrow the search focus when indications of a potentially promising subregion are found. Unfortunately, the initial version of the algorithm failed in this regard. The problem was that the exploration mechanism just described was too strong. Portions of potentially good augmented parameter vectors would be overwritten and destroyed by individual optimizers. In response, two elements were introduced. First, no matter what else happened, a certain number of the best augmented vectors found would be preserved from one generation to the next. This is called *elitist* selection. Second, a conceptual element was borrowed from simulated annealing. Each generation produced new sets of augmented vectors. However, the augmented vectors (and their objective function values) were retained elsewhere in memory. If the new augmented vector produced a better objective function value, it was carried forward into selection. However, if the new augmented vector scored worse than the previous solution in the same vertical position, it was retained only with a probability that declined rapidly with the degree of its inferiority.

The search strategy was an iterative process that was repeated until all estimated yield value were successfully fitted with observed. Root Mean square error was computed using the following equation and used as evaluation criteria.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_p - Y_o)^2}$$

where,

N= Number of observation

Y_p= Predicted yield

Y_o=Observed yield

There were 34,212 individual yield observations from 182 cultivars planted in 354 location-years. One location-year had only 30 cm of soil profile data and so was excluded, reducing the total number of observations to 34,052. Averaging the observations for each variety within each planting produced 7,426 yield means; i.e., the number of VPL combinations. The GSP optimizers each have to solve a 7D problem so, somewhat arbitrarily, their population sizes were set to 72. For simplicity, the same size was used for all optimizers. Table 3.2 documents the resulting scale of the estimation problem, which is the same for both the SF and ICA approach.

3.2.7 Model Evaluation

Ideally, one would calibrate soil parameters and GSP's using data from one set of years and then validate with data from other years. However, this was not possible with this data set because no location was ever used more than once. Instead, a procedure grounded in graph theory was adopted. As a first step, all cultivars grown in less than five plantings and all site-years with less than five cultivars were excluded. Then the tripartite graph in Fig. 3.3 was divided into two parts as illustrated by the dotted line in Fig. 3.5. All segments that cross the line are colored red and the remainder blue. The VPL's linked by blue edges comprise the calibration set while those defined by red segments (termed a *cutset* in graph theory) form the validation set. It is clear by

inspection that no yield data used in calibration are also used in validation. The downside, of course, is that the same cannot be said for weather data – an unavoidable consequence of the no-location-reuse property of the data.

An automated process was used to examine different cutsets seeking one that comprised a ca. 10% sample of the data. In the end a set of 568 observations was used for validation and 6617 for calibration. The model was evaluated using observed and simulated yield data. To evaluate the model performance, coefficient of correlation (r), and root mean squared error (RMSE) (Willmott et al., 1985) were used. RMSE is one of the best statistics that summarizes the mean difference in the units of simulated and observed value. Coefficient of correlation measure the strength and direction of the relationship between observed and predicted value.

3.3 Results and Discussion

Because of their proprietary nature, Syngenta required data be destroyed at the conclusion of the contest. Therefore, all yield information presented here is rescaled to a [0,1] interval and derived statistics (e.g. RMSE values) are adjusted proportionately.

3.3.1 Quality of Fit

Fig. 3.6 shows the simulated and observed yield of 34052 total observations and 7426 (mean observations) from 182 cultivars and 353 site-years using two different optimization approaches. The overall RMSE of ca. 5% of total yield for both approaches suggests that both (ICA and SF) methods are equally good and can be used for estimation process. This goodness of fit value is consistent with the results of Irmak et al. (2000), Mavromatis et al.(2001), and Welch et al. (2002). Regressing observed on predicted mean values yields a slope close to 1.00 and R^2 of .80, which are indicative of a good fit. Slightly lower RMSE and R^2 values (~7%, 0.65, respectively) were observed when all 34,052 observations were used in the fit. This resulted

because of the high levels of variation in those lines that were planted more than once in each trial. Fig. 3.7 shows that about 40% of observations have coefficients of variation in the range of 10% to 28%. This is likely due to within-planting variation between replications. (However, it should be noted that no orderly pattern of replication numbers was detected.).

The results showed that out of 182 cultivars, about 10% of cultivars (18) had RMSE of < 2% of maximum yield, and about 12% (21) with RMSE >8% of maximum. However, remaining about 78 % (143) were in between (Fig. 3.8). The lower RMSE's occurred when the samples were quite small (<15) and probably represent overfitting. Over fitted result mostly occurs when optimizer tries to estimate too many parameters from a sample that are too small. The result of overfitting in our case was because of few number of observations (Fig. 3.9). Overfitting makes any individual too optimistic about the performance of the model which in fact not trustworthy result. Thorp et al. (2008) also suggested that if users select too many parameters to optimize, they may get very good fit in optimization process but get a poor fit for validation due to overfitting.

Result also showed that about 12% of lines were poorly fitted or under fitted (Fig. 3.9). Under fitting generally occurs when an estimator is not flexible enough to capture the underlying variation in the observed data. Our result showed that poorly fitted lines were due to the fewer number of observation along with high yield variability across the locations. High variability across the locations for same line suggest that variability has to be determined only from site characteristic since the estimated GSP's would be the same for all observations. But, at the same time, it was also observed that there are also other many lines that were planted in that same locations, and their yield value is largely different than the poor fitted lines. Thus, the optimizer in general estimated the site characteristics for other large number of lines to minimize the overall

objective function and miss to accurately simulate yield for the poor fitted lines and increase RMSE.

3.3.2 Parameter Stability

Fig. 3.10 and 3.11 shows the stability of estimated parameters (GSP and site parameters, respectively) obtained from different optimization approach used in this study. Each point represents the values of one parameter as estimated by the ICA (vertical axis) vs. the Separate factor (horizontal axis) approaches. The range limits of each plot equal as the imposed optimization constraints. The excellent goodness-of-fit values obtained from both optimization methods created an expectation of parameter stability but that was clearly not observed. It appears that the point scatter in Fig. 3.10 might be more constrained in the horizontal (ICA) direction as compared to the vertical (SF). This might be due to the fact that, unlike ICA approach, the SF method didn't have any linear constraints allowing the estimates to be more widespread. In Fig. 3.11, planting date estimates appear stable (upper left panel) but this impression disappears when one zooms into a single year (bottom right). Some stability is apparent in the calculated DLL values (lower left) but not in the SRGF parameter (upper right). Similar instability of parameter across sites were also observed by Thorp et al., (2015).

This degree of parameter instability demonstrates a significant degree of model *equifinality*, or, as termed in engineering and statistical fields, a *lack of parameter identifiability*. These terms describe the condition wherein different parameter values produce the same model predictions, rendering alternative estimates indistinguishable (Franks et al., 1997; Medlyn et al., 2005). Equifinality arises in any situation where changes in model outputs that result from altering one parameter can be exactly offset by adjusting some other one. Although its presence complicates model use, equifinality is not, *per se*, evidence of model misspecification.

3.3.3 Estimation effects of location determination method

Fig. 3.12 shows the cumulative distribution of the observation's residuals obtained from each of the Exact, Exact-Conditional, and Missing location assignment methods. A Kolmogorov-Smirnov test indicated that the Exact and Exact-Conditional distributions differ significantly ($p=0.02$). However, the test was found non-significant for other combinations. Comparatively higher residuals from the Exact category of site-years is not because of any biological reason but an artifact of having a larger number of observations (Fig. 3.13) with wide range of yield value. In addition, it was also revealed that extreme high and low yield observations were also linked to the higher and lower soil water holding capacity of each soil type. This result is demonstrated in Fig. 3.13 and 3.14.

Fig. 3.13 shows the convex hull of a scatter plot of predicted and observed yield which has residuals (>3 and <12 bu/ac) and (<-3 and >-10 bu/ac) from Exact, Exact-Conditional and Missing location types. Within these three types, site-years of the Exact category has 1580 observations, 608 and 1068 from Exact-Conditional and Missing soil types. Fig. 3.14 shows the soil water holding capacity (DUL-DLL) of each site-years from three different soil location types. Black and Magenta color dot in each figure are the same observations that were seen in Fig. 3.13 with the same color.

3.3.4 Computational Performance

The HGA runs were executed on two Linux clusters: BEOCAT at Kansas State University (https://www.cis.ksu.edu/beocat); and Stampede at the Texas Advance Computing Center (TACC) (https://www.tacc.utexas.edu). Run lengths were 100 generations. In each generation, there were total 534672 number of model runs to be executed. Each model run takes about 0.14 seconds to finish so, total model run time for one generation would take about 20.8 CPU-hours in local

machine. Thus, total number of model simulation were evenly distributed across 136 different Linux clusters. Therefore, the whole wall-clock time to complete all 100 generations is about 1000 min (~16 hr.) or total of ~2176 CPU hrs. Fig. 3.15 shows that the convergence curves for the SF and ICA methods are essential identical. The objective function improves rapidly in the first 10 generations and plateaus after 60. It is worth noting that although 534,672 model runs per generation yielded by the stacking method is 620 times less that the 3.31×10^6 runs per generation that would be needed if all parameter vector combinations were run. Considering the high problem dimensionality (Table 3.2) this is impressive performance. Furthermore, our optimization algorithm almost reached the minimum before a maximum number of generation is reached for both the optimization methods used. This suggests that even with a different route that different optimization methods took through different parameter space and terminates at different places, it is clear that neither routing entailed any barriers that HGA was less able to handle.

3.3.5 Validation

The model was calibrated using 6617 grain yield observations and evaluated for 568 independent data sets for both ICA and SF methods. Fig. 3.16 shows the predicted and observed yields from the calibration and validation processes. The validation RMSE of about 9.00% from both (SF and ICA) methods shows that the model can predict yields quite well in situations with the degree of independence permitted by the structure of this data set (i.e., no location used more than once).

A common test for simulation model evaluation is by looking at the linear regression line of observed and predicted values, and a perfect model is assumed to have unit slope and zero intercept. Looking at the validation regression line (Fig. 3.16b and 3.16d), it seems that the regression line deviated from 1:1 line. This kind of fitting regression line has very checkered

history in validation. This is because of type II error and is mostly escalated with amount data taken so that it can reject a good model. There is a statistical identity that relates the slope of regression line to the correlation coefficient and standard error of predicted and observed. Thus, a much better number to focus on in these figure is RMSE and which is about 9% and is really good. Because of the existing structure of data (no location used repeatedly), validation didn't incorporate the variation due to weather. Thus, in future, breeding trial data for use in modeling should be structured such that locations are at least periodically reused to enable the incorporation of weather variation in model evaluation.

Mavromatis et al. (2001) and Welch et al. (2002) showed that good characterization of soil information is essential for better model performance. Now, there are many progressive farmers who are using precision agriculture methods such as getting yield monitor data and keeping track of their crop performance, but they are not using these data for management decisions. This result suggests that those farmers can utilize those records using this model to characterize their site and come to this level of predictability.

3.4 Conclusion

Our results showed that a large number of breeding trial yield data obtained from a wide range of environments can be successfully used to estimate the cultivar parameters for the CROPGRO-Soybean model. Furthermore, model yield predictions for independent situations (no yield data used in calibration used in validation) were as good as in estimation. However, because of the structure of the existing data (no location were repeated), weather information used in calibration might also be used in validation.

The optimization algorithm developed for this study (HGA) showed its potential on estimating three different type of parameters (cultivar, management, and site) at once in very few generations.

It was also concluded that soil information was very critical for model simulations. The number of observations used in the estimation process is always critical because of which estimation might end up with an over or under fitted result.

Lack of stability on estimated parameters from different approaches was due to the equifinality problem which increases the model uncertainty. Although equifinality doesn't affect model prediction, it creates problems when anyone tries to link parameter value in to its genetics.

3.5 Acknowledgments

Invaluable assistance with cluster computing was provided by David Turner of the Kansas State University Department of Computing and Information Science, Manhattan, KS, and by John Fonner of the Texas Advanced Computing Center (TACC) at the University of Texas, Austin, TX. The development of certain concepts instrumental to this undertaking would not have been possible without sabbatical support provided to Welch by the NSF-funded iPlant Collaborative and by TACC.

3.6 References

Alagarswamy, G., Boote, K., Allen, L., Jones, J., 2006. Evaluating the CROPGRO–soybean model ability to simulate photosynthesis response to carbon dioxide levels. Agron. J. 98, 34–42.

Alderman, P.D., Boote, K.J., Jones, J.W., Bhatia, V.S., 2015. Adapting the CSM-CROPGRO model for pigeonpea using sequential parameter estimation. Field Crops Res. 181, 1–15.

Boote, K.J., 2011. Improving soybean cultivars for adaptation to climate change and climate variability. Crop Adapt. Clim. Change 370–395.

Boote, K.J., Hoogenboom, G., Jones, J.W., Ingram, K.T., 2008. Modeling nitrogen fixation and its relationship to nitrogen uptake in the CROPGRO model. Quantifying Underst. Plant Nitrogen Uptake Syst. Model. Press Florence USA 13–46.

Boote, K., Jones, J., Hoogenboom, G., Pickering, N., 1998. The CROPGRO model for grain legumes, in: Understanding Options for Agricultural Production. Springer, pp. 99–128.

Boote, K., Jones, J., Hoogenboom, G., Wilkerson, G., 1997. Evaluation of the CROPGRO-Soybean model over a wide range of experiments, in: Applications of Systems Approaches at the Field Level. Springer, pp. 113–133.

Boote, K., Jones, J.W., Batchelor, W., Nafziger, E., Myers, O., 2003. Genetic coefficients in the CROPGRO–soybean model. Agron. J. 95, 32–51.

Boote, K., Kropff, M., Bindraban, P., 2001. Physiology and modelling of traits in crop plants: implications for genetic improvement. Agric. Syst. 70, 395–420.

Boote, K., Pickering, N., 1994. Modeling photosynthesis of row crop canopies. HortScience 29, 1423–1434.

Boote, K., Sau, F., Hoogenboom, G., Jones, J.W., 2008. Experience with water balance, evapotranspiration, and predictions of water stress effects in the CROPGRO model. Response Crops Ltd. Water Underst. Model. Water Stress Eff. Plant Growth Process. 59–103.

Boote, K., Tollenaar, M., 1994. Modeling genetic yield potential. Physiol. Determ. Crop Yield 533–565.

Calmon, M., Batchelor, W., Jones, J., Ritchie, J., Boote, K., Hammond, L., 1999. Simulating soybean root growth and soil water extraction using a functional crop model. Trans. ASAE 42, 1867.

Comon, P., 1992. Independent Component Analysis, in: J-L.Lacoume (Ed.), Higher-Order Statistics. Elsevier, pp. 29–38.

Cooper, M., Technow, F., Messina, C., Gho, C., Totir, L.R., 2016. Use of Crop Growth Models with Whole-Genome Prediction: Application to a Maize Multienvironment Trial. Crop Sci.

Cronshey, R., 1986. . Urban Hydrol. Small Watersheds.

Curry, R.B., Jones, J.W., Boote, K.J., Peart, R., Allen, L.H., Pickering, N.B., 1995. Response of soybean to predicted climate change in the USA. Clim. Change Agric. Anal. Potential Int. Impacts 163–182.

Franks, S., Beven, K.J., Quinn, P., Wright, I., 1997. On the sensitivity of soil-vegetation-atmosphere transfer (SVAT) schemes: equifinality and the problem of robust calibration. Agric. For. Meteorol. 86, 63–75.

Grimm, S.S., Jones, J.W., Boote, K.J., Herzog, D., 1994. Modeling the occurrence of reproductive stages after flowering for four soybean cultivars. Agron. J. 86, 31–38.

Grimm, S.S., Jones, J.W., Boote, K.J., Hesketh, J.D., 1993. Parameter estimation for predicting flowering date of soybean cultivars. Crop Sci. 33, 137–144.

Hammer, G., Cooper, M., Tardieu, F., Welch, S., Walsh, B., van Eeuwijk, F., Chapman, S., Podlich, D., 2006. Models for navigating biological complexity in breeding improved crop plants. Trends Plant Sci. 11, 587–593.

Irmak, A., Jones, J., Batchelor, W., Irmak, S., Paz, J., Boote, K., 2006. Analysis of spatial yield variability using a combined crop model-empirical approach. Trans. ASABE 49, 811–818.

Irmak, A., Jones, J., Batchelor, W., Paz, J., 2002. Linking multiple layers of information for diagnosing causes of spatial yield variability in soybean. Trans. ASAE 45, 839.

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

Jones, J.W., Boote, K., Jagtap, S., Mishoe, J., 1991. Soybean development. Model. Plant Soil Syst. 71–90.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. Eur. J. Agron. 18, 235–265.

Klein, L.J., Marianno, F.J., Albrecht, C.M., Freitag, M., Lu, S., Hinds, N., Shao, X., Rodriguez, S.B., Hamann, H.F., 2015. PAIRS: A scalable geo-spatial data analytics platform, in: 2015 IEEE International Conference on Big Data (Big Data). Presented at the 2015 IEEE International Conference on Big Data (Big Data), pp. 1290–1298. doi:10.1109/BigData.2015.7363884

Matthews, R., Stephens, W., Hess, T., Middleton, T., Graves, A., 2002. Applications of crop/soil simulation models in tropical agricultural systems. Adv. Agron. 76, 31–124.

Mavromatis, T., Boote, K., Jones, J., Irmak, A., Shinde, D., Hoogenboom, G., 2001. Developing genetic coefficients for crop simulation models with data from crop performance trials. Crop Sci. 41, 40–51.

Mavromatis, T., Boote, K., Jones, J., Wilkerson, G., Hoogenboom, G., 2002. Repeatability of model genetic coefficients derived from soybean performance trials across different states. Crop Sci. 42, 76–89.

Medlyn, B.E., Robinson, A.P., Clement, R., McMurtrie, R.E., 2005. On the validation of models of forest CO2 exchange using eddy covariance data: some perils and pitfalls. Tree Physiol. 25, 839–857.

Messina, C.D., Jones, J., Boote, K., Vallejos, C., 2006. A gene-based model to simulate soybean development and yield responses to environment. Crop Sci. 46, 456–466.

Michalewicz, Z., Janikow, C.Z., 1996. GENOCOP: a genetic algorithm for numerical optimization problems with linear constraints. Commun. ACM 39, 175.

Pathak, T.B., Jones, J.W., Fraisse, C.W., Wright, D., Hoogenboom, G., 2012. Uncertainty analysis and parameter estimation for the CSM-CROPGRO-Cotton model. Agron. J. 104, 1363–1373.

Paz, J., Batchelor, W., Jones, J., 2003. Estimating potential economic return for variable soybean variety management. Trans. ASAE 46, 1225.

Piper, E., Boote, K., Jones, J., 1998. Evaluation and improvement of crop models using regional cultivar trial data. Appl. Eng. Agric. 14, 435–446.

Ray, D.K., Mueller, N.D., West, P.C., Foley, J.A., 2013. Yield trends are insufficient to double global crop production by 2050. PloS One 8, e66428.

Reymond, M., Muller, B., Leonardi, A., Charcosset, A., Tardieu, F., 2003. Combining quantitative trait Loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiol. 131, 664–675. doi:10.1104/pp.013839

Ritchie, J., 1998. Soil water balance and plant water stress, in: Understanding Options for Agricultural Production. Springer, pp. 41–54.

Sau, F., Boote, K.J., Ruız-Nogueira, B., 1999. Evaluation and improvement of CROPGRO-soybean model for a cool environment in Galicia, northwest Spain. Field Crops Res. 61, 273–291.

Saxton, K., Rawls, W., 2006. Soil water characteristic estimates by texture and organic matter for hydrologic solutions. Soil Sci. Soc. Am. J. 70, 1569–1578.

Sexton, P., Batchelor, W., Boote, K., Shibles, R., 1998. Evaluation of CROPGRO for prediction of soybean nitrogen balance in a Midwestern environment. Trans. ASAE 41, 1543.

Singh, P., Nedumaran, S., Ntare, B., Boote, K., Singh, N., Srinivas, K., Bantilan, M., 2014. Potential benefits of drought and heat tolerance in groundnut for adaptation to climate change in India and West Africa. Mitig. Adapt. Strateg. Glob. Change 19, 509–529.

SSURGO, 2015. Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Soil Survey Geographic (SSURGO) Database.

Swaney, D., Mishoe, J., Jones, J., Boggess, W., 1983. Using crop models for management: Impact of weather characteristics on irrigation decisions in soybeans. Trans. ASAE 26, 1808–1814.

Tardieu, F., Reymond, M., Muller, B., Granier, C., Simonneau, T., Sadok, W., Welcker, C., 2005. Linking physiological and genetic analyses of the control of leaf growth under changing environmental conditions. Crop Pasture Sci. 56, 937–946.

Technow, F., Messina, C.D., Totir, L.R., Cooper, M., 2015. Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PloS One 10, e0130855.

Thorp, K.R., DeJonge, K.C., Kaleita, A.L., Batchelor, W.D., Paz, J.O., 2008. Methodology for the use of DSSAT models for precision agriculture decision support. Comput. Electron. Agric. 64, 276–285.

Thorp, K. R., Hunsaker, D. J., French, A. N., Bautista, E., & Bronson, K. F. (2015). Integrating geospatial data and cropping system simulation within a geographic information system to analyze spatial seed cotton yield, water use, and irrigation requirements. *Precision Agriculture*, *16*(5), 532-557.

Timsina, J., Boote, K., Duffield, S., 2007. Evaluating the CROPGRO soybean model for predicting impacts of insect defoliation and depodding. Agron. J. 99, 148–157.

Welch, S., Jones, J., Brennan, M., Reeder, G., Jacobson, B., 2002. PCYield: model-based decision support for soybean production. Agric. Syst. 74, 79–98.

Welch, S.M., Wilkerson, G., Whiting, K., Sun, N., Vagts, T., Buol, G., Mavromatis, T., 2002. Estimating soybean model genetic coefficients from private–sector variety performance trial data. Trans. ASAE 45, 1163.

White, J., 2009. Combining ecophysiological models and genomics to decipher the GEM-to-P problem. NJAS-Wagening. J. Life Sci. 57, 53–58.

White, J., 1998. Modeling and crop improvement, in: Understanding Options for Agricultural Production. Springer, pp. 179–188.

White, J.W., Hoogenboom, G., 2010. Crop response to climate: ecophysiological models, in: Climate Change and Food Security. Springer, pp. 59–83.

Wilczek, A.M., Roe, J.L., Knapp, M.C., Cooper, M.D., Lopez-Gallego, C., Martin, L.J., Muir, C.D., Sim, S., Walker, A., Anderson, J., Egan, J.F., Moyers, B.T., Petipas, R., Giakountis, A., Charbit, E., Coupland, G., Welch, S.M., Schmitt, J., 2009. Effects of genetic perturbation on seasonal life history plasticity. Science 323, 930–934. doi:10.1126/science.1165826

Wilkerson, G., Jones, J., Boote, K., Ingram, K., Mishoe, J., 1983. Modeling soybean growth for crop management. Trans. ASAE 26, 63–0073.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'donnell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models.

Yin, X., Kropff, M.J., Stam, P., 1999. The role of ecophysiological models in QTL analysis: the example of specific leaf area in barley. Heredity 82, 415–421.

Table 3.1 CROPGRO-Soybean genotype specific parameters. Linear ICA equations and ranges shown for the seven targeted parameters along with the constants used for non-targets.

GSP	Unit	Value/Equation	Range
CSDL	hr	12.904-0.49X1-.132X2+.3318X3-.411X4+.13X5+.10X6 Critical short day length below which the development rate is not affected by day length	[13.04 to 13.84]
EMFL	days	X7 Time between plant emergence and first flowering	[14.5 to 21.5]
SDPM	days	34.117+.72X1-2.27X2-.41X3+.03X4+1.681X5-1.455X6 Time between first seed and physiological maturity	[32 to 28]
FLSD	days	12.991-.98X1+1.3X2+1.45X3-1.031X4-.88X5-1.435X6 Time between first seed and physiological maturity	[11.5 to 16.5]
SFDUR	days	26.709+.801X1+1.043X2-3.06X3-2.73X4+1.3X5-2.94X6 Seed filling duration for pod cohort in standard growth conditions	[22.0 to 25.4]
LFMAX	-	1.044-.043X1-.006X2-.003X3+.005X4+.013X5+.007X6 Maximum leaf photosynthesis rate at 30^{o}C	[1.0 to 1.2]
PODUR	days	12.02-0.773X1-0.196X2-1.30`X3-3.43X4-.463X5-2.97X6 Time required for cultivar to reach final pod load under optimal conditions	[8.0 to 12]
FLSH	days	Time between first flower and first pod	6.0
FLLF	days	Time between first flower and end of leaf expansion	26.0
SLAVR	cm^2/g	Specific leaf area of cultivar under standard growth conditions	370.0
SIZLF	cm^2	Maximum size of full trifoliate leaf	180.0
XFRT		Maximum fraction of daily growth that is apportioned to seed and shell	1.0
WTPSD	gm	Maximum weight per seed	0.165
SDPDV	#pod	Average seed per pod under standard growing conditions	2.20
THRSH	%	Threshing percentage.	78
SDPRO	g / g	Fraction protein in seeds	0.405
SDLIP	g /g	Fraction oil in seed	0.205
PPSEN	1/hr	Slope of the relative response of development to photoperiod with time	0.129

Table 3.2 Problem and cluster statistics used in optimization approach.

Definition	**ICA**	**SF**
Number of parameter vectors to be estimated (equals the number of optimizers: 182 varieties + 353 planting dates + 353 soils)	888	888
Problem Dimensionality (equals the total number of all X-factors; 182 varieties × 7 GSP's + 353 site-years × (1 planting date + 2 soil parameters)	2,333	2,333
Total number of CROPGRO-Soybean runs per HGA generation (equals the number of VPS combinations times the optimizer population size: 7426 × 72)	534,672	534672
Total HGA population size (equals the number of optimizers times the population size: 888 × 72)	63,936	63936

Fig. 3.1 Example of one site based on the latitude, longitude provided in trial data. The provided information corresponds to the yellow pin, which is actually located in a residential area; however, the trial location can be inferred from the image. Field plot trials have identifying features such as many parallel alleys that can be used to identify their location – in this case marked with an "X". *Google Earth, 43°53'38.19"N,91°05'50.56"W. 9/28/15*.

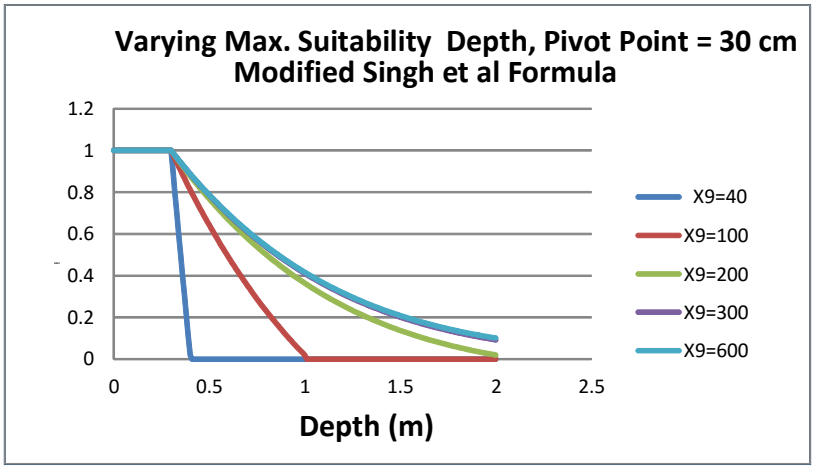

Fig. 3.2 Soil root growth factor for a variety of maximum suitable depth (X9) values. Note: that the horizontal axis is in meters but the parameter values are specified in centimeters. The search range is $40 \leq X9 \leq 500$ cm.

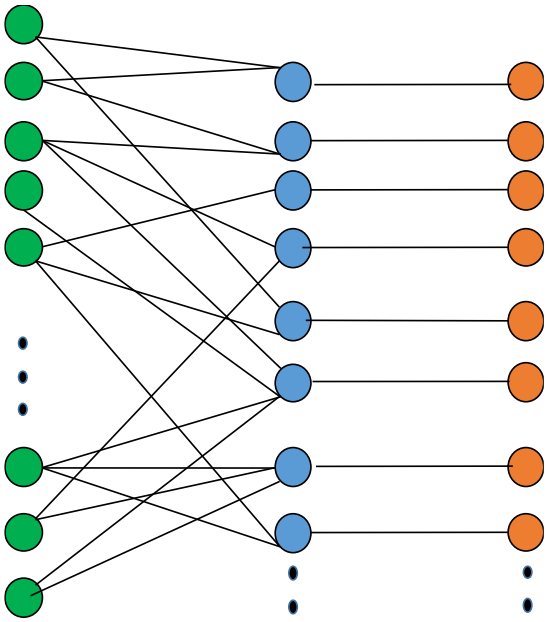

Fig. 3.3 Estimation problem structure. Green circles are varieties, brown circles are sites, and blue circles are particular planting dates. The black lines tell which cultivars were planted on which dates at which sites. As discussed in the text, each site has only one planting date in a given year.

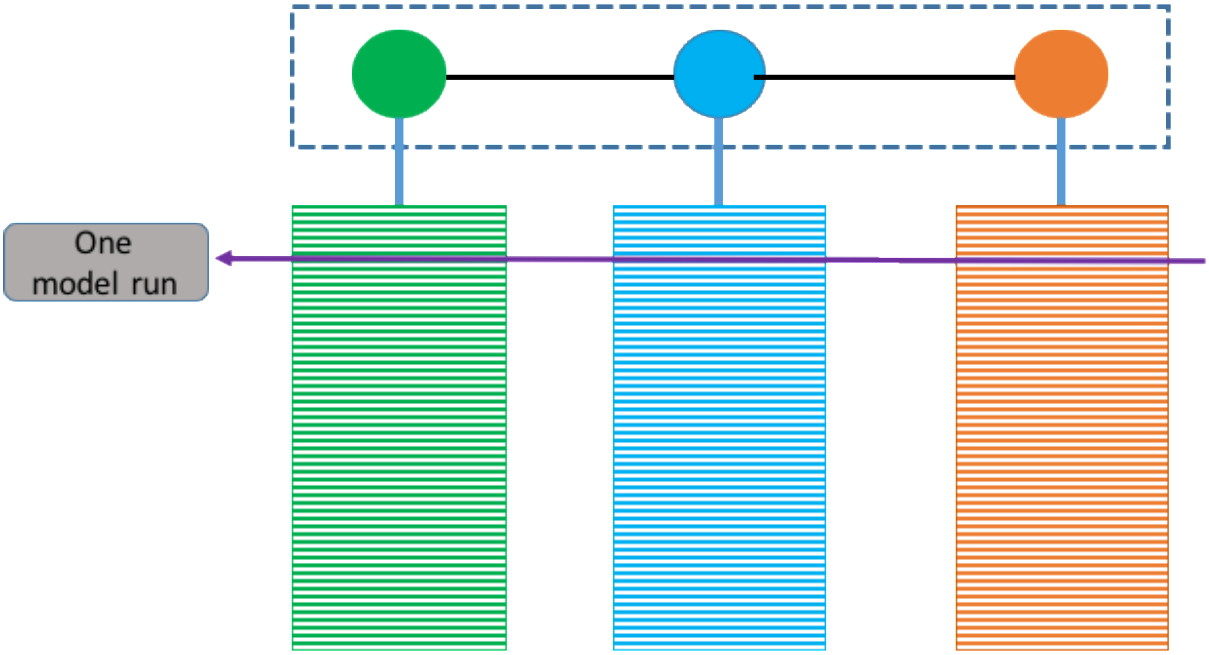

Fig. 3.4 Population structure used in HGA. Green, blue and brown circles are the optimizer for varieties, planting date, and site characteristics respectively. Stacked horizontal stripes are the population used in each optimizers and correspond to individual trial parameter vectors.

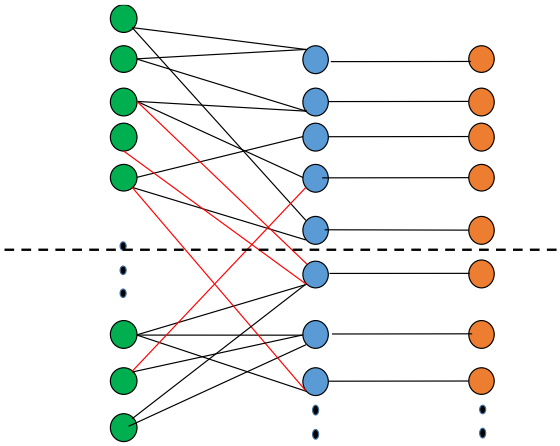

Fig. 3.5 Approach to separate calibration and validation data sets. Green circles are varieties and blue circles are particular location-year.

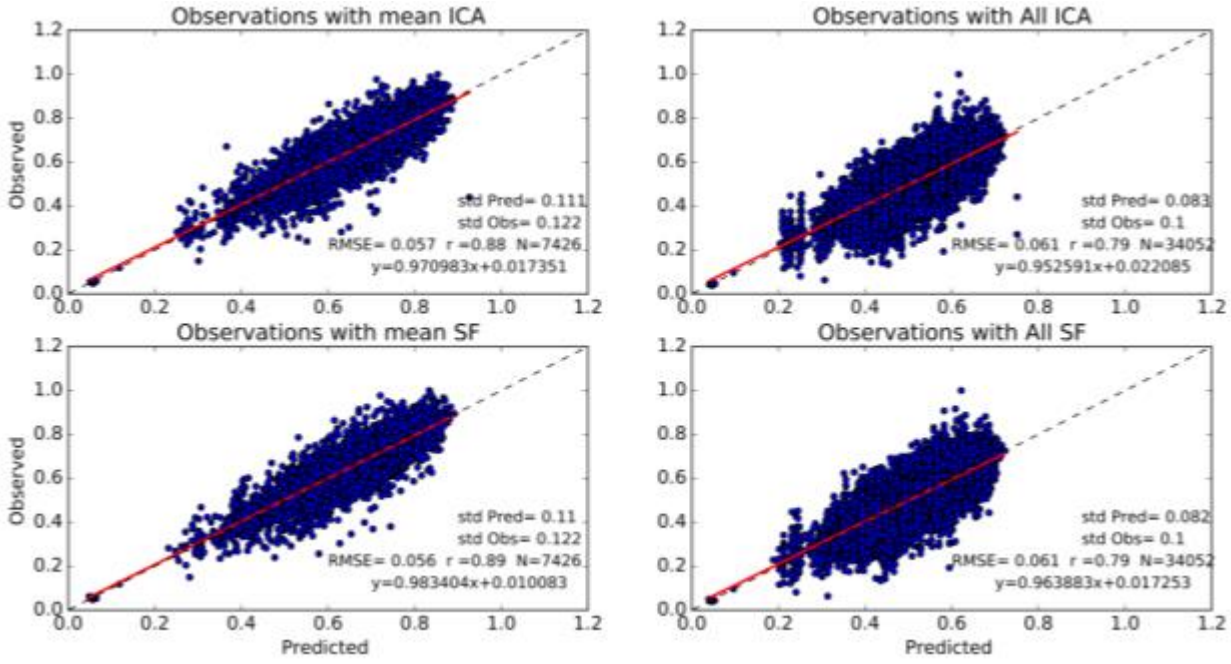

Fig. 3.6 Observed yield compared with predicted from ICA and SF optimization approach taken from all observation and observations with mean. All yield data are rescaled to a relative, [0,1] scale.

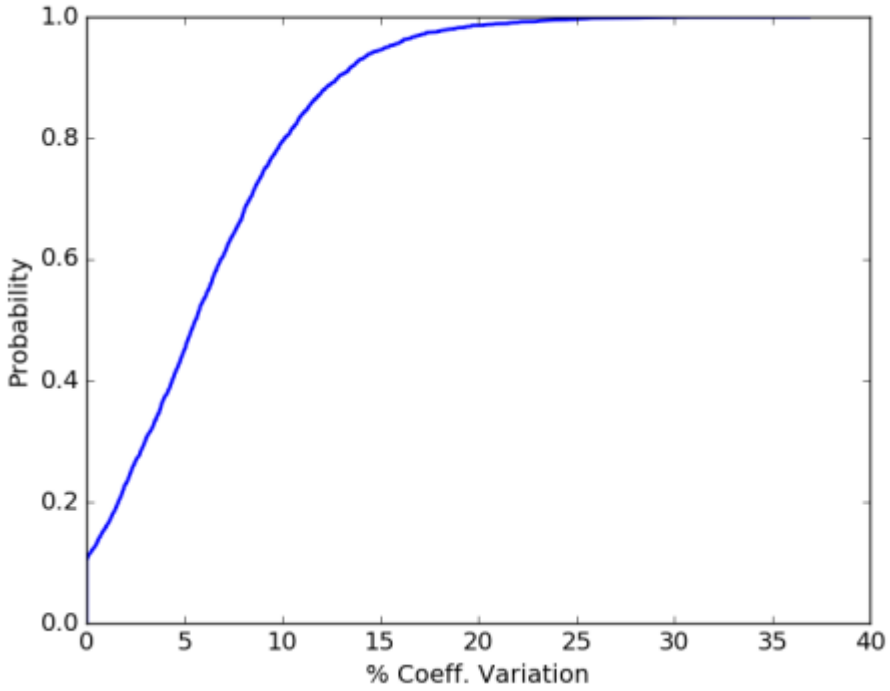

Fig. 3.7 Cumulative distribution of coefficient of variation of observed yield of each observation.

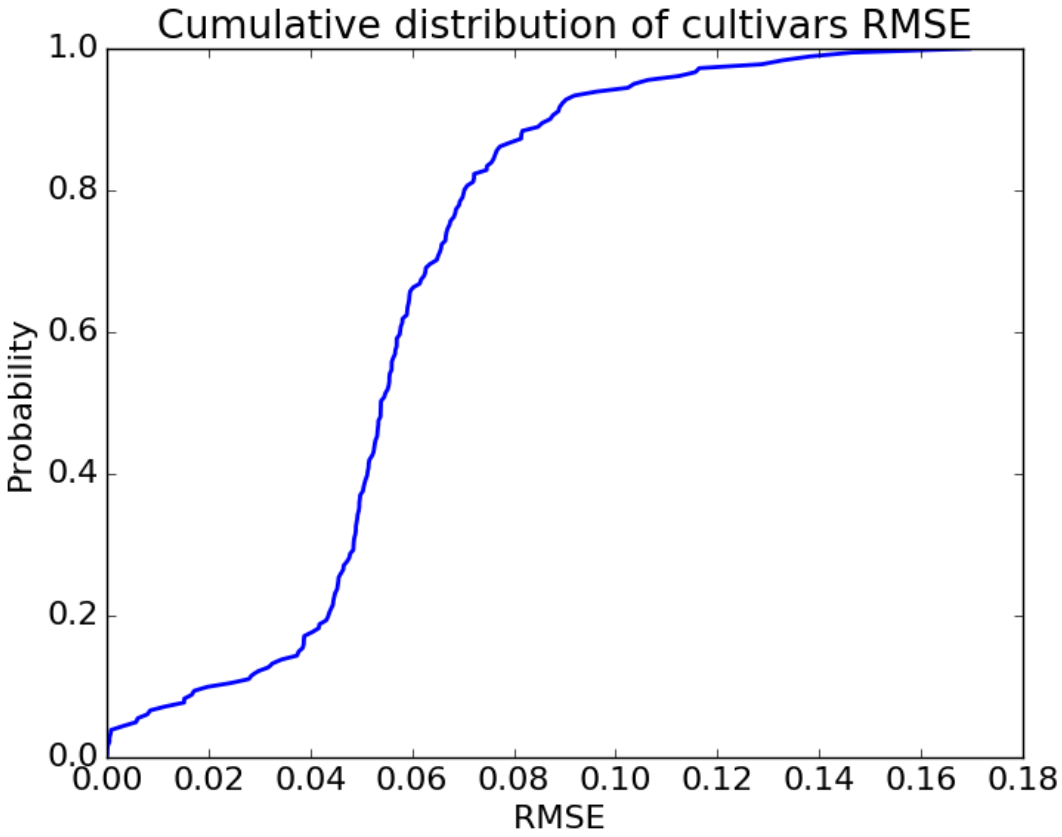

Fig. 3.8 Cumulative distribution of RMSE obtained from each 182 cultivars. RMSE value was calculated from rescaled data relative to [0,1] scale.

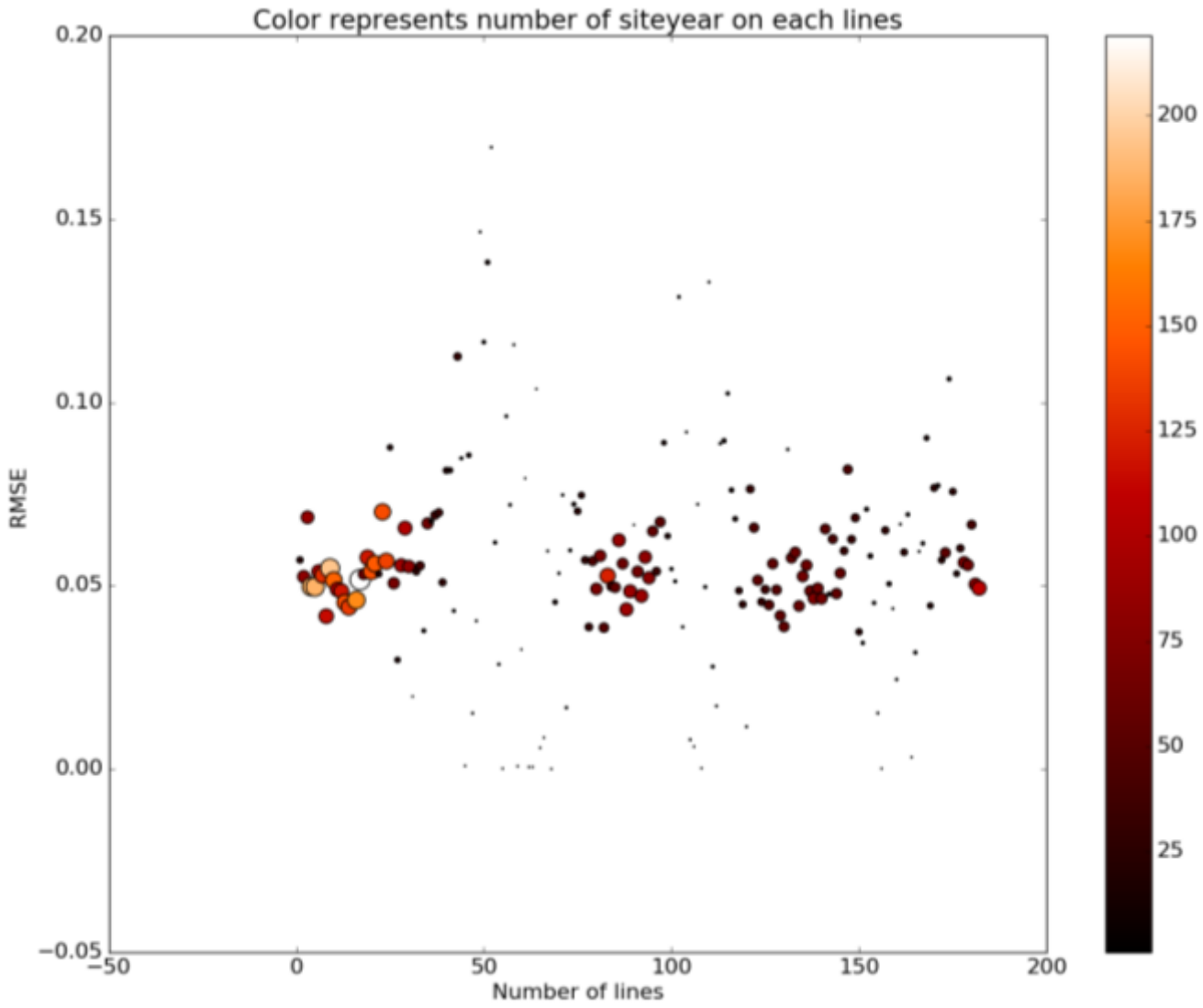

Fig. 3.9 RMSE for each lines. Each dot represents individual lines and size/color of each dots represents the number of site-year present in each line. RMSE value was calculated from rescaled data relative to [0,1] scale.

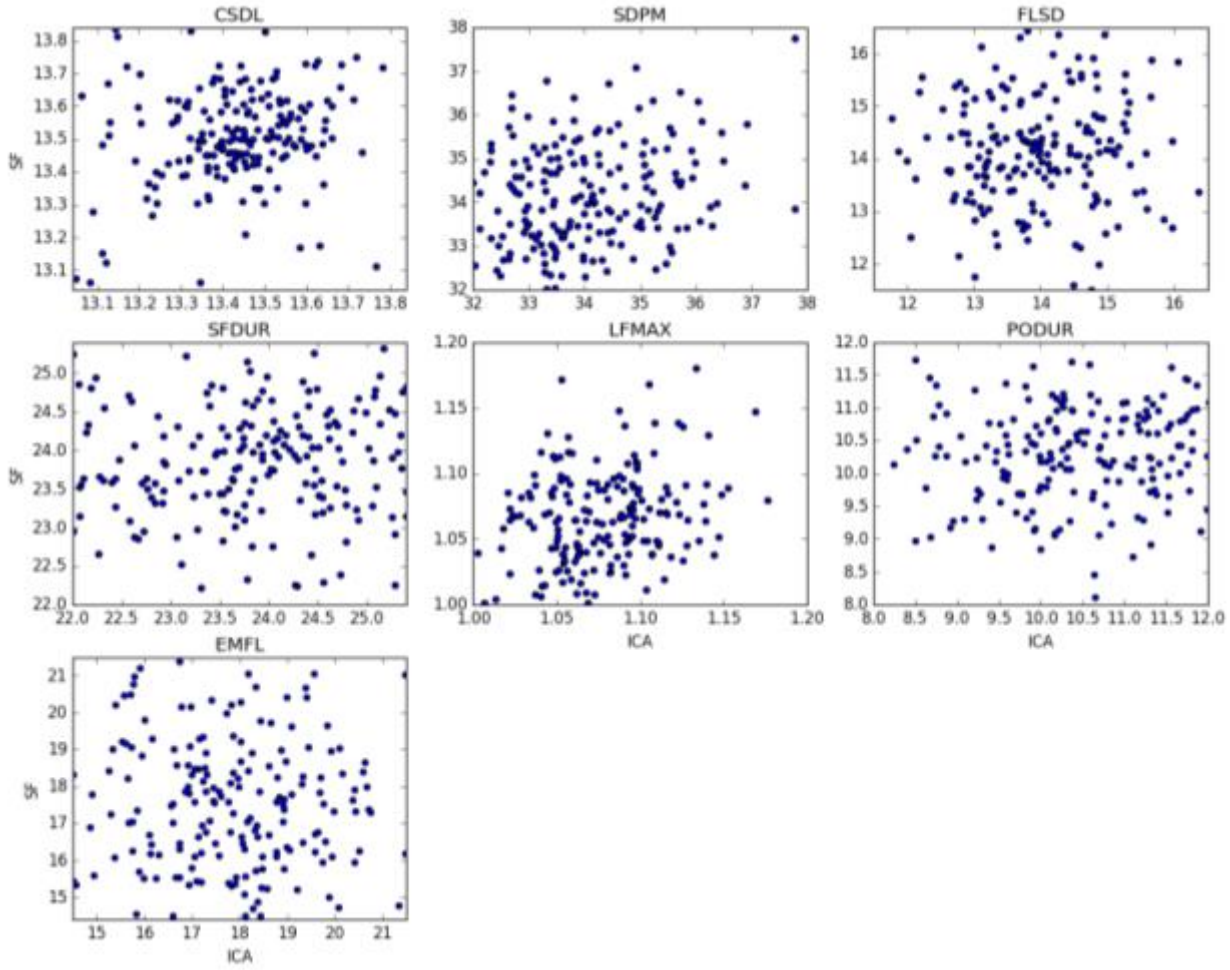

Fig. 3.10 Genotype specific parameters (GSP's) value obtained from estimation compared with SF and ICA optimization approach.

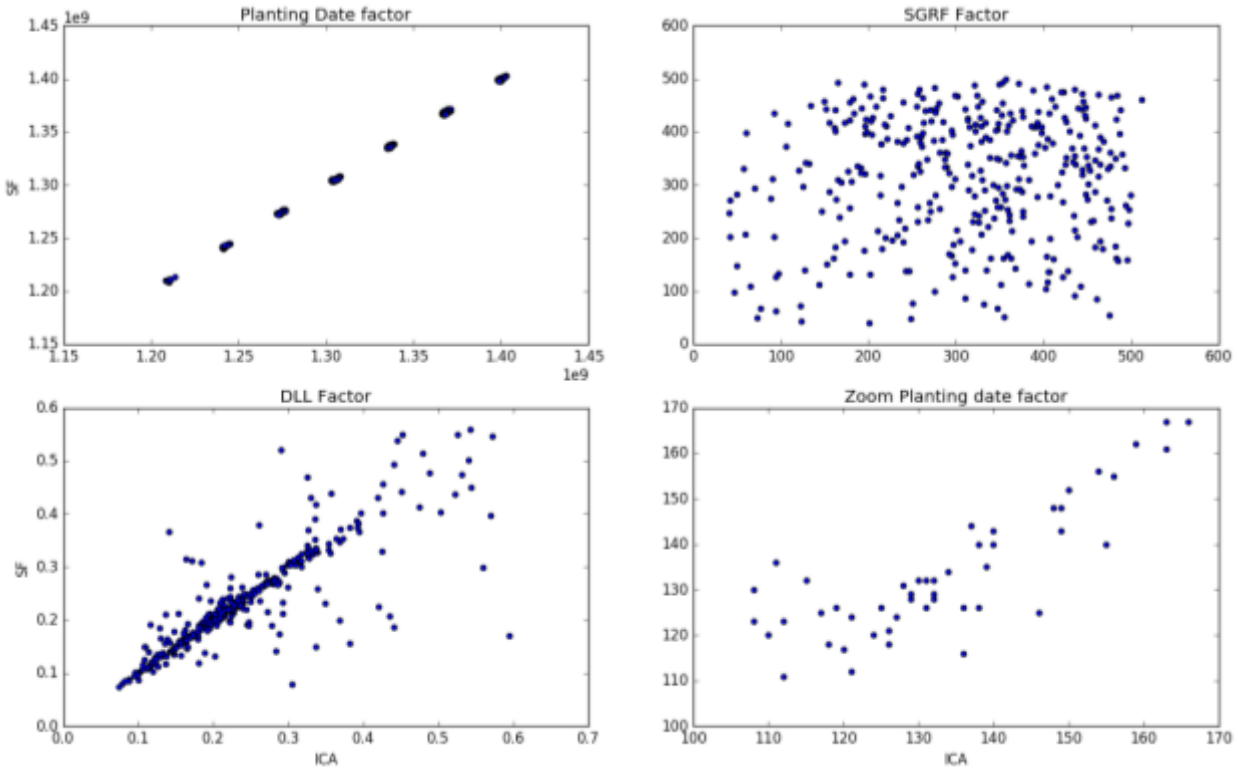

Fig. 3.11 Site parameters (Planting date (a), Soil root growth factor (b), Soil water factor (c) value obtained from estimation compared with ICA and SF optimization approach. d. Zoom section of planting date.

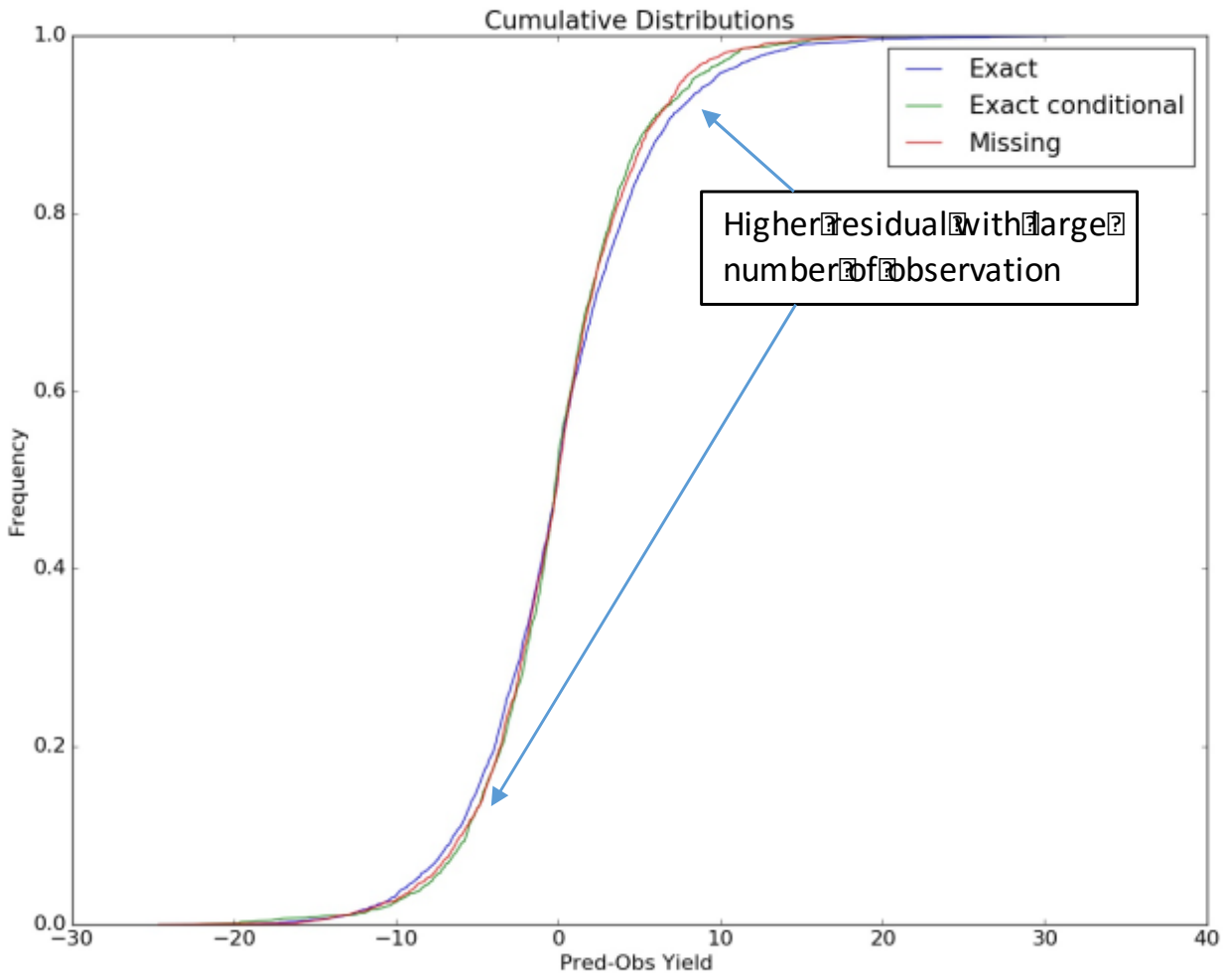

Fig. 3.12 Cumulative distribution of yield residuals obtained from three different soil location types.

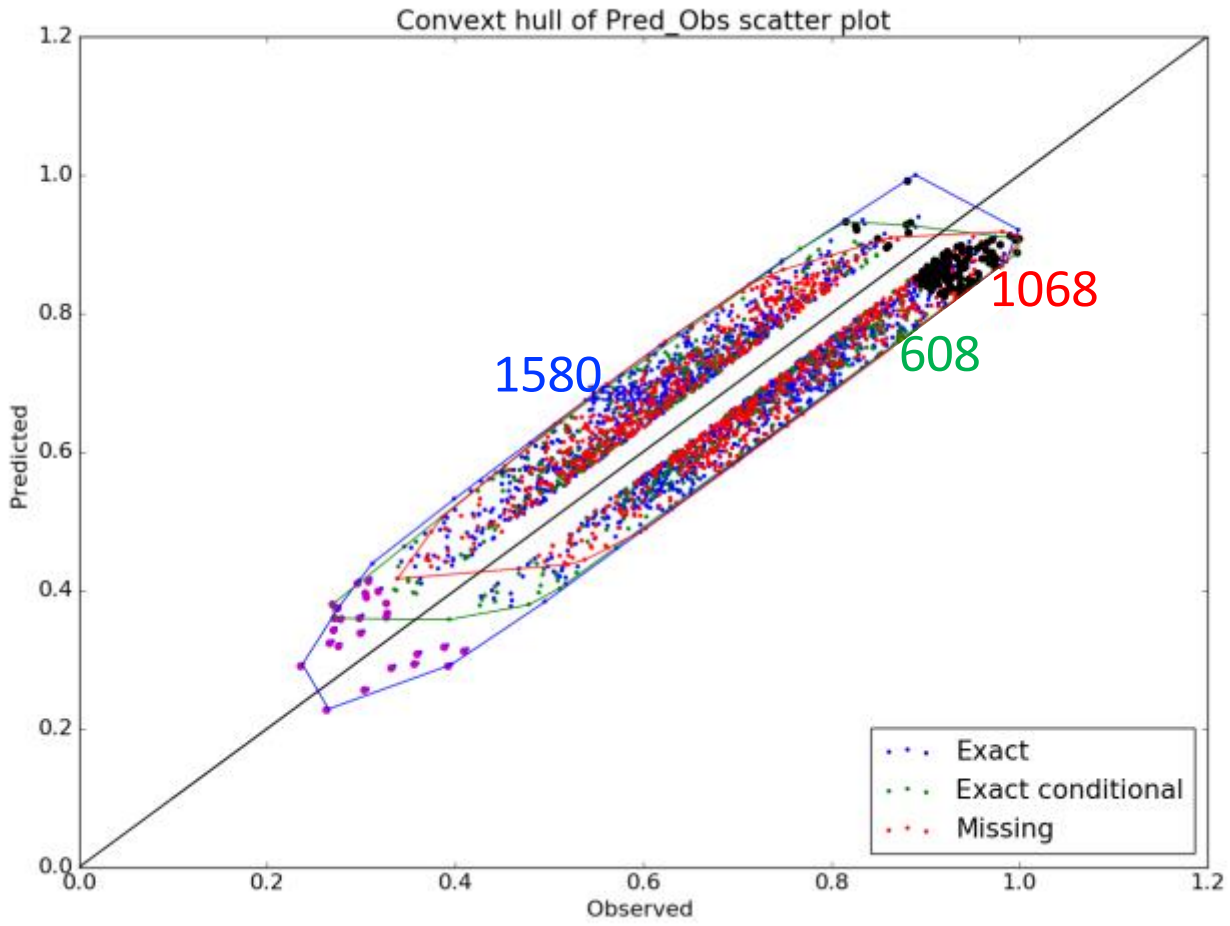

Fig. 3.13 Convex hull from each observation's predicted and observed yield for three different soil types.

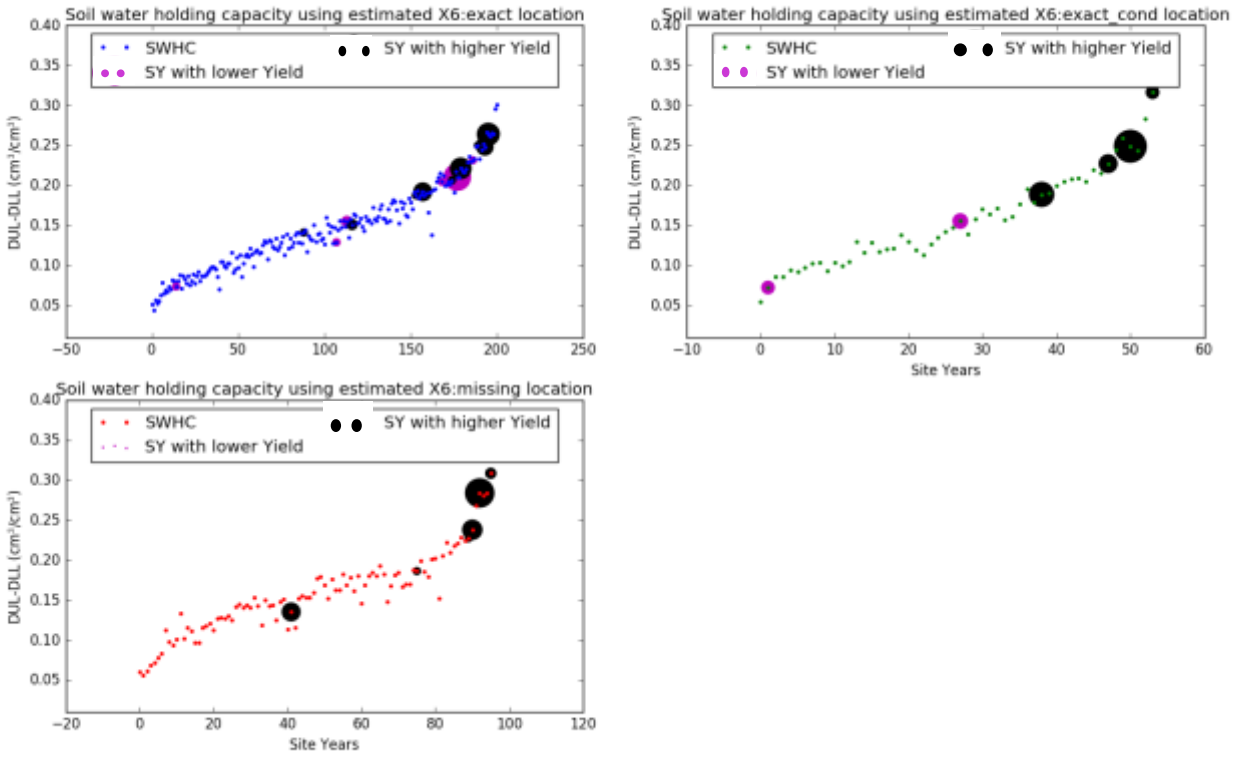

Fig. 3.14 Soil Water Holding Capacity (SWHC) from each location-year present in three different soil location types.

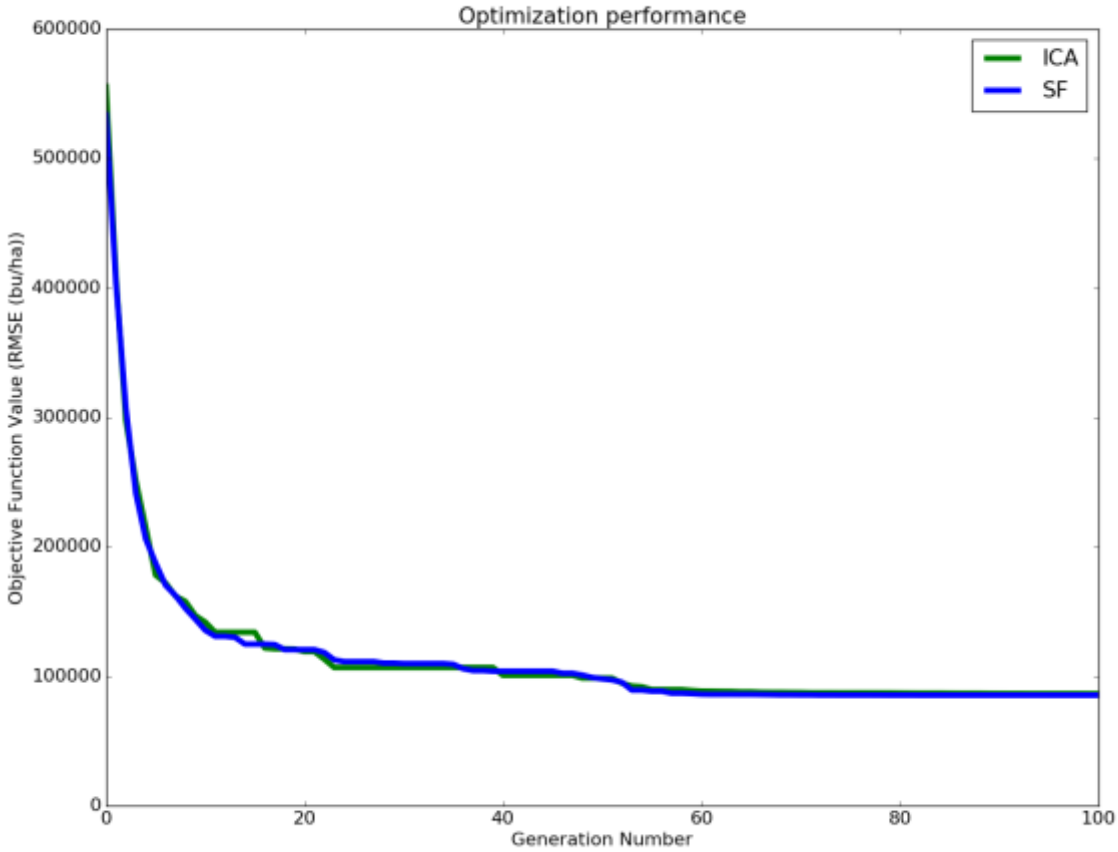

Fig. 3.15 Optimization performance throughout each generation from ICA and SF approach. Objective function value is the total sum of RMSE estimated from all 888 optimizers.

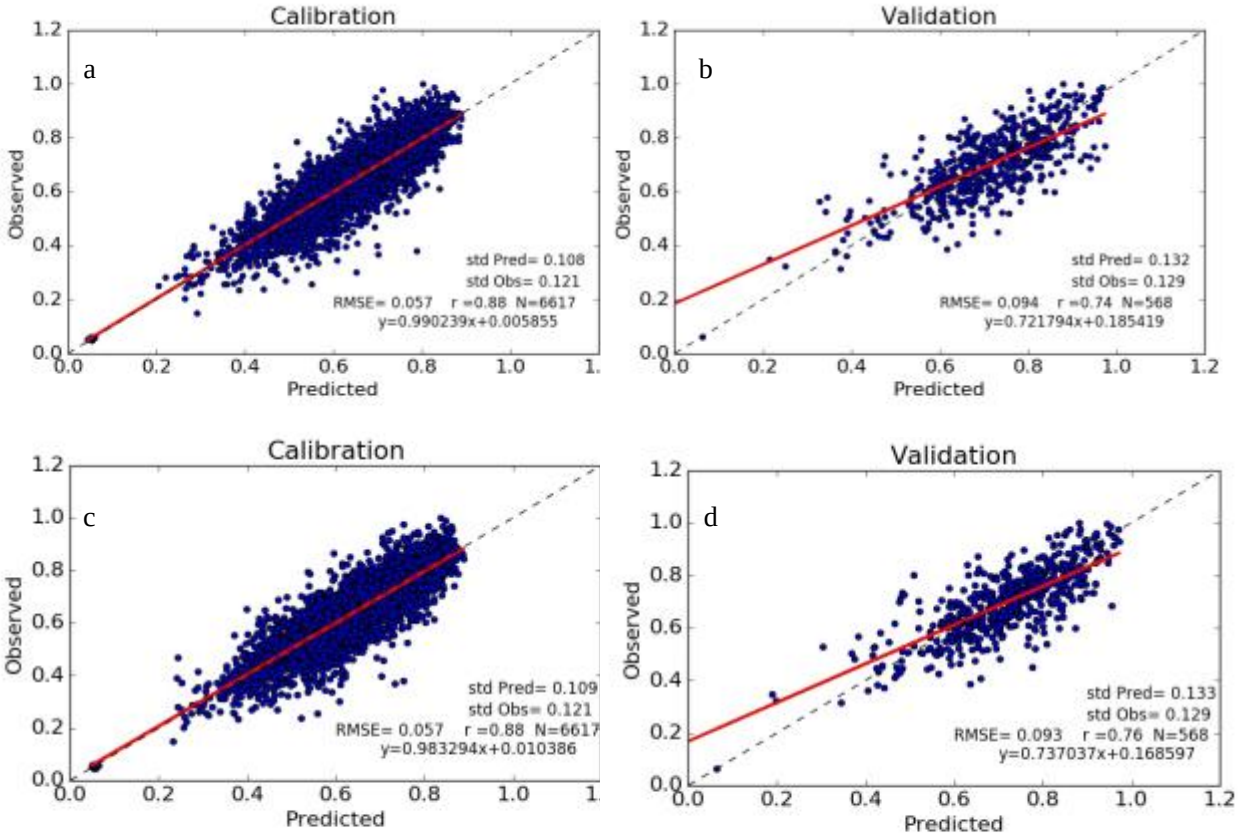

Fig. 3.16 Observed and predicted yield compared for calibration and validation data sets for a) ICA and b) SF approach. Model was validated using 568 independent observations obtained from 17 different cultivars and 271 different site-years and calibrated with 6617 observations. Values were scaled to 0 to 1.

CHAPTER 4 - PROBLEMS WITH ESTIMATING ANTHESIS PHENOLOGY PARAMETERS IN *ZEA MAYS*: CONSEQUENCES FOR COMBINING ECOPHYSIOLOGICAL MODELS WITH GENETICS

Abstract

Ecophysiological crop models encode intra-species behaviors using constant parameters that are presumed to summarize genotypic properties. The accurate estimation of these parameters is crucial because much recent work has sought to link them to genotypes. The original goal of this study was to fit the anthesis date component of the CERES-Maize model to 5266 genetic lines grown at 11 site-years and then genetically map the resulting parameter estimates. However, despite the high predictive quality of the values obtained, numerous artifacts emerged during estimations. The constraining issues fall into two categories. The first arose in situations where the model was unable to express the observed data for many lines, which ended up sharing the same parameter value. In the second (2254 lines), the model reproduced the data but there were often many parameter sets that did so equally well (equifinality). These artifacts made our original goal of genetic mapping completely unachievable.

4.1 Introduction

In the opening sentences of the 1968 book, The Population Bomb, Paul Ehrlich (and his wife Anne, uncredited at publisher behest) wrote, “The battle to feed all of humanity is over. In the 1970s hundreds of millions of people will starve to death in spite of any crash programs embarked upon now” and, in a subsequent chapter, “I don't see how India could possibly feed two hundred million more people by 1980." Fortunately, research started in India by Norman Borlaug

before 1968 created high yielding dwarf wheat varieties that, worldwide, are credited with averting one billion deaths from famine. India also introduced IR8, the so-called "miracle rice" developed at the International Rice Research Institute in the Philippines and the predicted human catastrophe was averted.

Now nearly 50 years later, the specter of global disruption is again upon us. This time the challenges are not only increasing human population (which has doubled since 1970) but also new phenomena like climate change and declining water resources. The confluence of these manifold trends makes finding ways to feed nine billion people by 2050 one of the most pressing issues of our time (Stone, 2011). However, the annual percentage increase rates for crop yields are only half those required to meet that goal (Godfray et al., 2010).

Beginning some 20 years ago, a paradigm has emerged offering the promise of dramatically accelerating breeding programs via improved phenotype prediction of prospective crop genotypes in novel, time-varying environments subject to sophisticated management practices (Cooper et al., 2016; Hammer et al., 2006, 2002; Technow et al., 2015; Welch et al., 2005a; White and Hoogenboom, 1996; Yin et al., 2003, 1999). The basic notion has two parts. The first is to exploit ecophysiological crop models (ECM's) to describe the intricate, dynamic, and environmentally responsive biological mechanisms that determine crop growth and development on daily or even hourly time scales. The aim is to use highly detailed, nonlinear crop simulation models to predict the phenotypes of interest within a subsample of possible environments and in-field management options. ECM's, whose origin is often credited to Wit. (1965), encode intra-species behavioral differences in terms of constant parameters that are presumed to summarize genotypic properties. On the strength of that presumption, the constants are termed *genotype-specific parameters* (GSP's).

The second part of the paradigm is to use quantitative genetic methods such as genomic prediction (Meuwissen et al., 2001) to relate the GSP's to genotypic markers (Cooper et al., 2016; Technow et al., 2015). Next, the outcomes of crosses are estimated by (1) calculating the GSP values that would arise from possible offspring genotypes. These values are then (2) used in ecophysiological model runs to predict the phenotypes in the target population of environments (for which detailed descriptive data must be available). In simplified instances, this approach has seen remarkable success (e.g., Reymond et al., 2003).

Composed of large coupled sets of continuous-time differential equations, ecophysiological models simulate many interacting processes (Jones et al., 2003; White and Hoogenboom, 2010) operating in the soil-plant-atmosphere continuum. These processes include physiology (e.g., photosynthesis, respiration, resource partitioning to various plant parts, and growth), phenology (leaf emergent timing, the date of vegetative-to-reproductive development, etc.), as well as chemistry and physics (soil water flows, chemical transformations, energy fluxes, gas exchange, etc.). During simulation runs, model formulas compute instantaneous process rates based on plant status and environmental conditions at each time point. These rates are integrated (*sensu* calculus) to output time series of dozens of plant variables. The models typically have 10 to 20 GSP's whose estimates are read from input files at the start of model execution. Numerous other inputs (e.g. soil water holding capacities by layer; measured daily solar radiation, rainfall, maximum and minimum temperatures; etc.) further quantify the physical environment.

The lynchpin of the entire two-step paradigm is the accurate estimation of the GSP's so that they can be related to allelic states in the genotype. Unfortunately, the direct measurement of GSP's is so time- and resource-demanding as to be infeasible for large numbers of lines. Indirect GSP estimation via model inversion is also challenging because easily-measured plant phenotypes

exhibit strong interactions with the environment (Chenu et al., 2009) thus increasing data requirements by necessitating trait measurement in multiple settings (Hammer et al., 1987). Even so, ecophysiological crop models enjoy extensive global use in areas ranging from global climate change, policy analysis, crop management, etc. Indeed, a Google search on the abbreviations of just two major model systems [namely "DSSAT" (Hoogenboom et al., 2015) and "APSIM" (Keating et al., 2003)] returned 134,000 hits. Not surprisingly, there is an extensive literature (reviewed briefly below) on ecophysiological model parameter estimation.

Initially, the authors' intent was to apply the two-step method to anthesis date using data from a very large panel of maize nested association mapping (NAM; McMullen et al., 2009) lines developed specifically to enable high-resolution studies of trait genetic architectures. Not only is anthesis date a phenotype of major biological significance, but it was also studied in this same panel using conventional statistical genetic methods (Buckler et al., 2009; Hung et al., 2012). Our hypothesis was that applying the proposed 2-step paradigm would demonstrate its merit in the specific context of the large data sets increasingly used in crop breeding programs to interrelate genotypes and phenotypes. We believed that contrasting the results of the standard and ecophysiological approaches would be interesting and informative. Granted, the model fitting methods to be used were not novel, but we expected that a further demonstration of their value with data sets much larger than ever used before would have utility.

However, something quite different happened. We discovered modeling issues and estimation artifacts that are of sufficient severity and generality that, if not addressed, are likely to imperil the breeding acceleration paradigm. Therefore, the objectives of this paper were 1) to describe these problems and the methods that revealed them (which can be applied as detection

tools in studies of other traits) and 2) to discuss research directions that might ameliorate the problems.

4.2 Background

Numerous optimization methods have been used to estimate parameters for ECM's. Oddly enough, a frequently used approach seems to have been that of trial and error (Wallach et al., 2001), wherein different parameters values are manually tested until an acceptable match between simulated and observed data is found. This approach, of course, becomes highly inefficient as the number of model parameter increases. Thus, numerous off-the-shelf, automated optimization techniques have been used. Examples include the simplex method (Grimm et al., 1993), simulated annealing (Mavromatis et al., 2002; Thorp et al., 2008), sequential search software (GENCALC) (Hunt et al., 2001), Uniform Covering by Probabilistic Region (UCPR) (Klepper and Hendrix, 1994; Román-Paoli et al., 2000), particle swarm optimization (PSO) (Koduru et al., 2007), and generalized likelihood uncertainty estimation (GLUE) He et al., 2010, 2009). While these traditional optimization techniques have advantages, they can be inefficient in terms of runtime and are highly dependent on optimization settings when thousands of combinations of line × planting site-years are involved – a situation that is becoming common in the era of massive genetic mapping populations. The fundamental issue is that, as the number of lines and environments increases, estimating GSP's for each line independently can require highly redundant simulation. To this end, we adapted an algorithm pioneered by Welch et al. (2000) and Irmak et al. (2000), as described in methods section. The approach exhibits particular efficiencies when individual plantings incorporate large numbers of lines and, serendipitously, supports a close examination of the estimation process itself.

The vast majority of prior ECM parameter estimation studies have been conducted in non-genetic contexts. Against these backgrounds, the sole merit criterion has been the predictive skill demonstrated by the GSP estimates obtained. However, the current setting is markedly different – GSP's are not just inputs to ecophysiological crop models; GSP's simultaneously function as the outputs (i.e. dependent) variables of genetic prediction models. As such, GSP's are at least as closely related to tangible biochemical processes at the molecular level as they are summative of physiological properties (e.g. maximum photosynthetic rates) in higher organizational realms. Therefore, a deeper inspection of their estimation is warranted and two concepts are helpful in achieving the enhanced discernment now required.

This report uses the word "expressivity" (and the adjective "expressive") to describe a model's innate ability to reproduce a set of observations independent of particular parameter values. An expressive model may fail to replicate data because an unskilled optimizer cannot find a meritorious combination of parameter values. In contrast, a model with low expressivity will not fully mimic actual data irrespective of what (biologically or physically reasonable) values are assigned to its parameters. In cases where the latter behavior is detected, remedies will be vigorously sought. However, as shown below, however, systematic gaps in expressivity can coexist even within an overall framework of predictively skilled model performance.

Another model property that has received sparse attention in prior estimation studies is equifinality. Equifinality describes a situation in which multiple sets of parameter values generate exactly the same model predictions. In statistics, a synonym for "equifinality" is "parameter non-identifiability" (Franks et al., 1997; Medlyn et al., 2005). When the only concern is prediction quality and that seems "good enough", it is easy to consider equifinality a non-problem. However, when parameters are intermediaries rather than just inputs and equifinality exists, it begs the

question as to what relationship, if any, putative GSP estimates might bear to allelic states across the genotype? A moment's reflection shows that equifinality and expressivity are different model properties. The former relates to how many different estimates yield identical predictions; the latter refers to the possible existence of systematic failures of those predictions to mimic observed data.

In this paper, we explore these issues in modeling and estimation using the anthesis phenology component of the CERES-Maize ECM (Jones et al., 1986; Kiniry and Bonhomme, 1991; Major and Kiniry, 1991) and observed dates from multiple plantings of three maize genetics panels totaling nearly 5300 lines. Anthesis initiates the period of grain development and is therefore a critical milestone toward grain yield. As such, it mediates the adaptation of the crop to its environment by customizing vegetative and reproductive growth phases and is a key target of breeding programs (Buckler et al., 2009). (Although at the apical meristem, floral initiation precedes the visible morphological change of anthesis, the linkage between the two is tight enough that we follow common modeling practice and consider them as effectively synonymous.) The genetics of flowering time has been intensively studied in the model plant *Arabidopsis thaliana* where well over 100 influential genes are now known (Andrés and Coupland, 2012; Bratzel and Turck, 2015). Indeed, gene expression models of flowering time of *A. thaliana* based on differential equations have been developed (Valentim et al., 2015), and genetically-informed approaches have established the relationships between network-level function and common ecophysiological time formulations (Wilczek et al., 2009). In maize, our understanding of the genetic control on flowering time is more limited but has been advancing in recent years. More than 30 genes have been described and conservation of key features from *A. thaliana* seems apparent (Table 1 in Dong et al., 2012). A quantitative gene network model based on a number of these loci has been published (Dong et al., 2012).

The general desire within applied quantitative genetics to probe genetic architectures has led to the construction of ever-larger and/or special purpose mapping populations (Buckler et al., 2009). The maize NAM panel (McMullen et al., 2009) was constructed by making bi-parental crosses between one common parent, B73, and each of a set of 25 other inbreds that collectively encompassed a wide range of maize diversity. Approximately 200 offspring from each of these 25 crosses were then inbred for a number of generations to ensure, to the greatest degree feasible, that the influence of each locus on any trait of interest reflected the contribution of one parent only. Individual plant genotypes produced in this fashion are called "recombinant inbred lines" (RIL's). Buckler et al. (2009) reported a seminal study of maize anthesis dates using this NAM panel. Demonstrating the power of these lines to finely dissect genetic contributions to traits of interest, they identified 36-39 QTL, where the exact number depended on the analysis method used. Most of the QTL had small effects but, collectively, explained 89% of total anthesis date variation.

For the reasons outlined above, accurate prediction of anthesis date is a major target for ecophysiological crop models (Román-Paoli et al., 2000). However, few studies exist in the literature that have used large data sets for ECM calibration. Mavromatis et al. (2002) reported 5,109 site-year-line-parameter combinations and Welch et al. (2002) estimated 4,620 site-year-line-parameters. In contrast, the effort presented herein, which required supercomputing capabilities, encompassed 197,964 site-year-line-parameter combinations – to our knowledge, the largest such study ever reported. As the following sections document, it was the sheer scale of this data set and the resulting scatterplots depicting thousands of lines that brought to light worrisome issues of equifinality and expressivity failures (described in detail next), that might well have been overlooked in studies of smaller scale.

4.3 Materials and Methods

4.3.1 Experimental data

Observations collected on anthesis date for a total of 5266 maize lines were obtained from the Panzea data repository (http://www.panzea.org) The lines used were members of three genetic panels. In particular, 4785 lines were from the 25 RIL panels comprising the maize NAM set described above. Also included were an additional 200 RIL lines commonly referred to as the IBM panel because they originated by **I**ntermating **B**73 × **M**o17 (Lee et al., 2002). Finally, a maize diversity panel (Flint-Garcia et al., 2005) contributed data on 281 additional lines. Various combinations of these lines were grown at six US sites: New York (NY), North Carolina (NC), Illinois (IL), Missouri (MO), Florida (FL) and Puerto Rico (PR), during 2006 and 2007 for a total of eleven site-years. In what follows "NY6" denotes the 2006 planting in New York, respectively by state abbreviation and year for other site-years. Table 4.1 gives the exact locations of the experimental sites, and the respective sowing dates. The "Total Lines" row of the table gives the number of lines from the three panels that were present in each study. The "Lines with data" row lists the number of lines with available observations on anthesis date. Data on daily maximum and minimum temperatures for each site were provided by the maize NAM collaborators (H. Hung, personal communication, 2010) and did not include metadata on position of the weather stations to the field plots, types and calibration of sensors or types of radiation shields used.

4.3.2 CERES-Maize model

The Crop Estimation through Resource and Environment Synthesis (CERES)-Maize model is one of the oldest, most widely used ecophysiological crop models for maize (Quiring and Legates, 2008). We used the CERES-Maize version incorporated in CSM 4.5 (Cropping System Model; (Hoogenboom et al., 2015; Jones et al., 2003). The CERES-Maize simulation of

development toward anthesis is controlled by a set of GSP's and environmental inputs (Kiniry and Bonhomme, 1991; Major and Kiniry, 1991). Specifically, the GSP's studied herein were thermal time from emergence to juvenile phase (P1), critical photoperiod (P2O), sensitivity to photoperiods longer than P2O (P2), and the phyllochron interval (PHINT) as measured in thermal time. The duration of Stage 1, the interval from emergence through the end of the juvenile phase, is calculated by accumulating daily thermal time until P1 is reached. Stage 2 follows immediately and lasts until tassel initiation. Stage 2 lasts a minimum of four days when the photoperiod (including civil twilight) is less than P2O. P2 specifies the number of extra days required for every hour by which the photoperiod exceeds P2O. The model continues to accumulate thermal time through Stage 2. The model assumes that (1) there are five embryonic leaves; (2) two new leaves initiate during each phyllochron interval; and (3) that anthesis date, which terminates Stage 3, occurs when all leaves present at the end of Stage 2 (i.e., total leaf number, TOLN) are fully expanded. The date on which this happens is when the ongoing thermal time accumulation reaches TOLN × PHINT.

Thermal time is calculated from inputs of daily maximum and minimum temperatures. Sowing dates (Table 4.1) determined the time series of weather data that control simulated plant growth and development. The model calculated daily photoperiods from geographic position. Other required model inputs did not affect predicted anthesis dates and thus were not required here. For example, the soil water and nutrient balance components of the model do not affect simulated anthesis date in the CERES-Maize model and therefore were not used in this study. The model also requires row spacing and planting depth, which were set to 0.5 m and 2.5 cm, respectively. No tillage, pest, or disease effects were simulated.

4.3.3 Parameter estimation

4.3.3.1 Search strategy

In the conventional approach to parameter estimation (Fig. 4.1a), an optimizer iterates through a series of trial solutions for which model predictions are generated in each environment. The entire process is repeated for each line. This approach becomes inefficient when many lines are planted together in large experiments and are therefore exposed to identical environments. This is because estimates approaching optimal goodness-of-fit will only emerge in the latter stages of an iterative optimization run. Therefore, the majority of early iterations for each line entail the repeated evaluation of estimates with mediocre predictive ability in the same environment.

To overcome this problem, we adapted an approach described by Irmak et al. (2000) and Welch et al. (2002, 2000). In their scheme (Fig. 4.1b), model simulations were conducted for each planting across a multidimensional grid of parameter value combinations. The resulting predictions were stored in a database. As a second step, for each line the root mean square error objective function (RMSE; Gill et al., 1981) between observed and predicted anthesis day of year was evaluated with respect to all combinations of parameter values across all site-years. That is, for line l,

$$RMSE_l = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_p - Y_o)^2} \quad (1)$$

where, n is the number of observations for that line (consisting of one observation per site-year combination), and Y_p (Y_o) is the predicted (observed) anthesis date. The optimizer goal was to minimize the RMSE for each line. If a unique minimum existed, it defined the combination of GSP values that best fit each line. Total computational time was reduced because time-consuming model simulations for each combination of GSP parameter values were only performed once, but their outputs were used many times in the much faster RMSE calculations. Another benefit is that a combination of GSP values that yielded poor predictability for one variety might perform better for a different line. Additionally, this process ensures that identical parameter combinations are

tested for each line, which can aid in comparing the results achieved. Finally, simply by retabulating the database, any number of different optimizations can be performed using different observations, alternative subsets of site-years, plantings or combinations of parameter values. The use of alternative objective functions is also possible without requiring additional simulations. Because of the central role played by the database of simulation outputs, we will refer to this scheme as the *database method*.

4.3.3.2 Sampling the model parameter space with sobol sequences

Unlike Irmak et al. (2000) and Welch et al. (2002, 2000) who sampled the parameter space with a rectilinear grid, we employed Sobol sequences so as to avoid the combinatorial explosion in computational requirements that accompany increasing dimensionality. Sobol sequences belong to a family of quasi-random processes designed to generate samples of multiple parameters dispersed as uniformly as possible over the multi-dimensional parameter space (Press et al., 1992; Sobol, 1998). Sobol sequences are specifically designed to generate samples with low discrepancy – that is, a minimal deviation from equal spacing. Unlike random numbers, quasi-random algorithms can effectively identify the position of previously sampled points and fill the gaps between them (Saltelli et al., 2010), thus avoiding the formation of clusters. Further, Sobol sequences offer reduced spatial variation compared to other sampling methods (e.g., random, stratified, Latin hypercube; see Fig. 4.2a vs. 4.2b), make this method more robust (Burhenne et al., 2011). We used a Python-based algorithm to generate a Sobol sequence of quasi-random numbers for calculating 32,400,070 sets of the four CERES-Maize GSP's, leading to a uniformly-sampled four-dimensional parameter space for P1, P2, P2O, and PHINT. To construct the database, CERES-Maize calculated anthesis date for each GSP combination in each of the 11 site-years – a

total of 356,400,770 model runs. Table 4.2 describes the upper and lower bounds and the number of distinct values obtained for each parameter.

4.3.3.3 High performance computing

The large number of model runs could not be performed by lab-scale computing facilities. Instead, we used the "Stampede" supercomputer at the Texas Advanced Computing Center (TACC; Burhenne et al., 2011)). *In toto*, the CERES-Maize runs required 63,372 CPU-hours, which equates to ca. 176 simulations per second distributed across 112 processors. The predicted anthesis dates were collated and transferred to the "BeoCat" computing cluster at Kansas State University (https://support.beocat.ksu.edu/BeocatDocs/index.php/Compute_Nodes). There, RMSE values were tabulated for each line × parameter value combination across all site-years in which anthesis date was observed. As combinations of GSP values were found that had progressively lower RMSE values, they were recorded by the computer. This process required ca. 15 minutes of wall clock time per line so the total estimation process was completed in ca. 7 hrs on 200 Xeon E5-2690 cores.

4.3.4 Assessing estimate properties

4.3.4.1 Equifinality

Equifinality occurs when multiple combinations of parameter estimates generate the same minimal RMSE value, often because they generate identical model predictions (Beven, 2006; Luo et al., 2009), in this case identical integer DOY values for anthesis dates. In what follows, we concisely quantify "equifinality" in any specific context by defining "number of ties" as the number of Sobol sets of parameter combinations that produced the same optimal RMSE values, minus one. No equifinality is present in a line if there is only one combination of parameter values that minimizes the RMSE. That is, there are zero ties among its estimates. To illustrate the

magnitude of the problem and our subsequent desire to study it more closely, we note that 2254 (43%) of the 5266 lines available in the data exhibited equifinality. The worst case was represented by a line that had 1,043,933 distinct combinations of GSP values that produced identical anthesis date predictions, and thus the same RMSE, thereby yielding 1,043,932 ties.

During the database tabulation phase, the values of the "best combination of parameter estimates seen so far" were updated only if its RMSE value was strictly better than all previously evaluated ones. So, when equifinality was present, the final GSP estimate was the first combination of parameter values encountered that had a minimal RMSE value. As a result, some of the analyses described below are sensitive to equifinality, illustrating the fact that subtle optimizer algorithm idiosyncrasies can have marked impacts on the overall results. Such cases are noted explicitly along with the procedures used to mitigate the effects.

4.3.4.2 Interrelationships between parameter estimates

Correlations and other relationships between parameter estimates are highly important to breeding programs and related simulation studies. When correlations between parameter estimates are known to be present, opportunities exist to select on one plant trait by selecting on a related phenotype instead. Additionally, there have been a number of *in silico* studies where CERES models were used to design crop ideotypes (Laurila et al., 2012; Semenov and Stratonovitch, 2013). Such efforts find combinations of model parameter values that predict phenotypes well suited to the target population of environments. Once identified, lines with those values become breeding targets. However, a potential pitfall arises if realizing the desired genotype involves changing parameter values in directions contrary to the correlations that exist between them.

For this reason, we explored the pairwise correlation structure of the GSP parameter estimates and generated pairwise scatter plots of their line-specific values. However, the latter

revealed a bizarre pattern, the diagnosis of which ultimately led us to the second problem alluded to in the introduction – the inability of the model to reproduce certain observational combinations – and to the methods presented next.

4.3.4.3 Model expressivity

A common graphical method to assess the quality of model fit is to plot the predicted vs. observed values (e.g., Fig. 4.3). Such scatterplots can be informative in detecting areas of mismatch between observed and predicted values, thus providing specific characterization of the model's lack of fit. By definition, each point in the scatterplot corresponds to a prediction that a model is able to make given an optimized set of parameter values. However, an entirely different question is whether there are observations that a given model cannot reproduce using *any* reasonable combination of parameter values? That is, one might seek to assess whether a given model has the requisite expressivity to reproduce the data.

The database approach allows such a question to be addressed using what we term *phenotype space* scatter plots. In such plots, each axis corresponds to a different site-year. The coordinates along the axes represent the observed or predicted anthesis dates for each site-year. Model expressivity is then assessed by comparing the scatter of predicted anthesis date generated from a wide range of GSP value combinations to the scatter of observed values in large data sets. Because equifinality does not affect predictions, this method of evaluating model expressivity is independent of the order in which an optimizer locates points that minimize RMSE values (see the second paragraph in section 4.3.4.1).

4.3.4.4 Testing for parameter stability across environments

In order for the two-step paradigm outlined in the Introduction to work, the estimates of GSP's should not vary across the set of environments used to estimate them, a property called

"stability" (Hammer et al., 2006). If GSP estimates did vary across environments, there would be no way to tell what GSP values to input to the ecophysiological model to predict traits whenever daily weather time series or soils differed from those used in the paradigm's first step. This might seem an insuperable barrier to readers for whom G×E interactions are virtually ubiquitous among quantitative plant phenotypes, but it is not. This is because the *raison d'etre* of models like CERES-Maize is to explain crop variety × environment interactions mechanistically based on physiological (often first) principles.

Many GSP's, including all the ones in this study, explicitly relate plant behaviors (e.g., development toward anthesis) to environmental variables (e.g., temperature and photoperiod in the current case). Modelers assert that GSP's are properties of the individual lines (i.e., stable) and, therefore, by implication, have a genetic basis because genotypes do not change with the environment. Over time, it is thus expected that research will mechanistically link at least some GSP's to molecular genetic processes. For example, it is known that both short (P2O) and long day critical photoperiods are determined by the dynamics of the CONSTANS protein in a range of plants including *Arabidopsis* (Andrés and Coupland, 2012) and a number of grasses (Colasanti and Coneva, 2009; Hammer et al., 2006), albeit not maize (Coles et al., 2010; Mascheretti et al., 2015). In rice (*Oryza sativa*), critical short day length has even been successfully predicted from a differential equation model of the diurnal expression patterns of the *CONSTANS* ortholog (Welch et al., 2005b).

Because stability is both important and reasonable to expect given the goals of ecophysiological modeling, it has been argued (Welch et al., 2005a) that finding a putative GSP to be unstable is *prima facia* evidence of some problem. Possible causes of instability include: (1) the model incompletely or incorrectly disentangles G × E; (2) a stable answer exists but the

optimizer is insufficiently skilled to find it; (3) undiscovered equifinality is present, and the solutions found depend on low-level algorithmic idiosyncrasies of the optimizer (e.g. section 4.3.4.1); and (4) unique best GSP estimates exist that the optimizer can find, but because the model is over parameterized, the values obtained reflect noise signals that differ between environments.

All sources of instability, whether these or others, are detrimental to the two-step ecophysiological genetic approach to phenotype prediction. Thus, it is critical to know when parameter instability is present, so herein we developed a statistical approach to detect and test for it. The specific question asked was "Do the GSP estimates depend on the particular set of environments used to construct them?" A conceptually simple way to answer this might be to (1) obtain a combination of parameter estimates from one subset of site-years, (2) repeat the estimation with a different subset, and (3) test whether the two sets of parameter estimates differ according to an appropriate statistical test.

A more general and robust approach, however, might be to obtain parameter estimates from many site-year subsets chosen according to a principled method. Preliminary tabulations of the Sobol database revealed that equifinality increased dramatically when fewer than seven site-years were used for estimation (see Results). Therefore, the subset size was set to seven site-years. One method for selection of site-year subsets might be to resample site-years with replacement. However, as shown by analogy in the Fig. 4.2b, randomization adds a source of variability to the results that could be of concern given that sampling by replacement would have $P_7^{11} = 39{,}916{,}800$ possible site-year subsets. Therefore, analogous to the Fig. 4.2a, we used a combinatorics-based sampling pattern leading to more uniformly-distributed site-year subsets by taking all combinations of 11 site-years 7 at a time, of which there are $C_7^{11} = 330$ possibilities. To maximize

the amount of data available for each line in any subset, we focused on the 539 lines for which observation were available in all 11 site-years.

We then conducted 177,870 (= 539 × 330) four-dimensional optimizations to obtain estimates for the four GSP parameters for each line × site-year set combination. These optimizations involved only Sobol database retabulations rather than new model runs, again illustrating the computational utility of this approach. When forced to generate a single estimate, the database search returned the combination of GSP estimates yielding a minimal RMSE that it happened to encounter first. To focus on the subset that lacked this element of optimizer arbitrariness we first dropped the 114,314 line × site-year combinations that had ties. Because our primary interest was in the variability that different site-year combinations might contribute to GSP estimates, we restricted our attention to the 297 site-year subsets that had at least 100 lines remaining after ties were removed. Each of the 539 lines was present in at least 28 site-year subsets, which was deemed adequate for GSP estimation. These actions left a total of 60,834 estimates for each of the four GSP's in the study. This became our base group for analysis. However, dropping estimates that have a common property (i.e., ties) is a systematic procedure that might, itself, influence the results. So we also examined the set of (1) all estimates and (2) those for which ties existed. In both cases we used the optimizer-selected values

We then specified a statistical model to test for stability in parameter estimates across environmental subsets, as follows:

$$\rho_{l,e} = \mu_\rho + \alpha_l + \beta_e + \varepsilon_{l,e} \tag{2}$$

where $\rho_{l,\mathrm{e}}$ represents an estimate of the GSP ρ (i.e. either P1, P2, P2O, or PHINT) for the l^{th} line (l = 1,2,… 539) obtained from the e^{th} site-year set (e = 1,2,… 297), μ is the intercept parameter, acting as an overall mean of GSP ρ across all lines and site-year subsets; α_l is the

differential random effect of line *l*, assumed to be distributed $\alpha_l \sim N\left(0, \sigma_l^2\right)$; β_e is the differential random effect of the e^{th} set of site-years, assumed to be distributed $\beta_e \sim N\left(0, \sigma_e^2\right)$; and $\varepsilon_{l,e}$ is the left-over residual unique to the l, e^{th} observed GSP estimate and assumed $\varepsilon_{l,e} \sim NIID\left(0, \sigma_\varepsilon^2\right)$. The differential line effects α_l are considered to be random as is common in field studies of plant population biology. Further, the differential effects of site-year sets, β_e, were treated as random because the corresponding environmental sets are combinations of 7 out of 11 plantings considered to be a representative, if not random, sample of the population of possible site-years to which we are interested in inferring.

If the estimation of any GSP parameter ρ were stable across the site-year subsets, one would expect the variance of β_e, namely σ_e^2, to be zero; alternatively, if estimation is unstable, one would expect $\sigma_e^2 > 0$. To test this hypothesis set, we fit two competing versions of the statistical model in equation (1), one with and one without the random effect of site-year subsets β_e for each of the GSP's ρ = P1, P2, P2O, and PHINT. For each GSP, we then compared the two competing models using a likelihood ratio test statistic against a central chi-square distribution with half a degree of freedom to account for the fact that the test is being conducted on the boundary of the parameter space. Statistical models were fitted using the liner mixed-effects model package lmer in R (Bates et al., 2014) with optimization based on the log-likelihood option. The lmer package also calculated the Akaike and Bayesian Information Criteria [AIC (Akaike, 1973) and BIC (Schwarz, 1978), respectively], which allow for an additional assessment of fit for statistical models that include or exclude the random effects of site-year subsets.

4.4 Results

4.4.1 Observations vs. Predictions

Fig. 4.3 shows a color-coded scatterplot of observed vs. predicted days to anthesis for 49,491 line × site-year combinations; the cloud of points is concentrated along the identity line, therefore suggesting accurate prediction; the overall estimated RMSE is 2.39 days. Also, there seem to be considerable differences between sites on anthesis days, whereby Florida and Puerto-Rico show very short vegetative durations (ca. 50 d), which are more than doubled in New York (120 d). Empirical correlation coefficients ($\hat{r}$) were high across site-years and ranged from 0.86 to 0.95, thus indicating an overall responsiveness across lines to the range of site-year conditions on anthesis dates. The standard deviations of the predicted values and their corresponding observations are 10.336 and 10.639, respectively, which, with the overall empirical correlation coefficient of 0.974, account for a close to 1-to-1 estimated regression slope of observations vs. predictions [i.e. 1.002 = (10.639 /10.336) *0.974], as per the established statistical identity between these four sample quantities (Harrison and Tamaschke, 1984).

4.4.2 Equifinality

A more complex picture emerges when the prevalence of equifinality is considered. As noted in 4.3.4.1, for the 2,254 lines exhibiting equifinality, the number of ties can exceed 1M. The histogram in Fig. 4.4a tabulates the frequency of ties across lines. There are 2,153 lines with fewer than or equal to 40 ties. The line trace along the upper portion of the top and bottom panels shows the average number of site-years in each bin.

From Fig. 4.4a, it is apparent that the empirical distribution of ties is right skewed, thereby indicating that a relatively large number of maize lines have few ties and thus low levels of equifinality. This is particularly true when parameter estimates were computed using data from 7

– 11 site-years (right axis of Fig. 4.4b). Further, the distribution of ties appears to have a very long tail to the right, whereby the number of lines with increasing amounts of equifinality declines very slowly while the number of site-year combinations used for estimation seems to plateau (Fig. 4.4a). This pattern continues into Fig. 4.4b, which shows the 101 lines with more than 40 ties. Note that no bars are shown in Fig. 4.4b due to scale of the y-axis, as each bin generally contains one to three lines. Interestingly, the number of ties, and thus equifinality, seems to increase precipitously for the 56 out of 5,266 lines that have fewer than seven site-years of data (Fig. 4.4b).

As the number of ties increases, one can expect that the range of indistinguishable estimates for any GSP will widen. To illustrate this phenomenon, a set of GSP estimates were obtained using just two illustrative site-years (NY6 and NY7) so as to artificially inflate equifinality. Fig. 4.5 shows scatterplots of coordinate pairs of either predicted (a) or observed (b) values for anthesis days from NY6 (horizontal axes) and NY7 (vertical axes). Points in each scatterplot are color-coded to represent the number (on a $\log_{10}$ scale) of tied GSP combinations. Each tied GSP combination, when simulated using the weather data for NY6 and NY7, predicts the same anthesis dates that form the point's coordinates. Dark red indicates 235,976 ties and blue indicates 1 tie. It is reasonable to expect that as the number of ties increases, the range (max minus min) of the equifinal estimates will increase. The size of each circle indicates the range of tied P1 estimates expressed as a percentage of the mean. These percentages extend from 0.36% to 65.68%. The association of redder colors with larger circles indicates that estimate ranges do, indeed, increase with the level of equifinality.

This is an example of a phenotype space plot that can be used to show how properties of interest (e.g. number of ties and estimate ranges in this case) are distributed across the range of predictions made by the model given the weather in a pair of site-years. Notice that (1) the cloud

of observed points (Fig. 4.5b) is more dispersed than that of the predicted points (Fig. 4.5a) suggesting that model responses to the environment are less plastic than those of real plants and (2), as made clear by the red lines, the lowest numbers of ties in Fig. 4.5b (blue points) appear to fall in empty regions of Fig. 4.5a where predictions are lacking. This pattern has important consequences to be explained later in section 4.4.4.

4.4.3 Interrelationships between parameter estimates

Fig. 4.6 presents a combined plot depicting histograms of GSP parameter estimates based on all 5,266 lines along the main diagonal and corresponding pairwise GSP scatterplots in the upper right panels. The GSP estimates were obtained using all site-years. The lower left panels in Fig. 6 show the estimated Pearson correlation coefficients ($\hat{r}$), estimated regression slopes ($\hat{b}$), and corresponding *p*-values for each mirrored scatterplot. Two immediately apparent features on the scatterplots are to be noted, which might readily escape notice in data sets with fewer lines. The first is the pronounced banding pattern appearing in all plots except, perhaps, P2O vs. PHINT. Most bands seem to be linear except for those on the scatterplot of P2O and P2 plot, which exhibits curvilinearity. The second is the pronounced vertical gap in all P2O scatterplots. In an attempt to understand the reasons for such patterns, the authors explored multiple seemingly plausible hypotheses, ranging from genetics to input file coding quirks (e.g., unintended rounding of parameter values) and many more, all of which were tested and discarded. Ultimately, the results presented in the following sections provided the explanations.

4.4.4 Model expressivity

The first clue to the cause of the banding pattern emerges from the phenotype space plots in Fig. 4.7. Each plot corresponds to an independent fit to just one particular pair of site years. The blue regions in each panel of Fig. 4.7 outline predicted anthesis date pairs for two consecutive

years in a given site, where model predictions are constrained by the bounds imposed on the range of values allowed for each of the four GSP's (Table 4.2). Also, for each panel in Fig. 4.7, a dot depicts an observed anthesis date pair for a line present in a given site in both 2006 and 2007. Yellow (red) dots represent observed anthesis date pairs that the model was able (unable) to reproduce. We characterize each observation corresponding to a yellow (red) dot as "expressible" ("inexpressible"). Except for the two North Carolina site-years, there were many lines (Table 4.3) for which observations on anthesis date could not be predicted despite (1) the seeming breadth of GSP values allowed by Table 4.2 and (2) the fact that the model was only being asked to match two data points, which would seem to greatly relax the constraints on GSP estimates.

This begs the question as to what would happen to model expressivity if an even broader range of GSP values were allowed. In an attempt to investigate in a computationally efficient way how the outputs of a more conventional optimizer might appear when viewed in phenotype space, the CERES-Maize anthesis date routine was ported to Python and fit to NY6/NY7 via Differential Evolution (DE; Das and Suganthan, 2011). DE is a well-established (63K Google Scholar hits on "Differential Evolution" as of October 21, 2016) and highly effective evolutionary algorithm that embodies mechanisms reminiscent of techniques ranging from the Nelder-Mead Simplex (Nelder and Mead, 1965) method to Particle Swarm Optimization (Kennedy, 2011; Koduru et al., 2007). Among the algorithm's initiating inputs is the range of parameter values within which to search, which were set as shown in Table 4.4. These ranges are greatly broadened from that used in the database search (Table 4.2); in fact, the values in Table 4.4 are intentionally broader than biological experience would suggest as reasonable.

Fig. 4.8 shows overlapping predictions based on the database search under the range of parameters in Table 4.2 and on the DE search under the extended range of parameter values (Table

4.4). Specifically, the light blue area represents the anthesis date region that was reachable through predictions based on the database search. In contrast, the dark blue area is the predicted anthesis date region within which the DE algorithm converged. Note the almost perfect overlap of the lower edges of the light blue (i.e. database search) and dark blue (i.e. DE search) areas, indicating that, despite its much larger starting parameter search space, DE did not extend model predictions. This suggests limitations in model expressivity that go beyond the method of parameter estimation or the initial parameter space used for the search.

As a corollary, it is worth noting that more site-years of data of similar quality are unlikely to improve model expressivity, as illustrated by the following thought experiment. Suppose a community has developed the univariate deterministic model $y = \arctan(\theta)$, where θ is a parameter, with $0 \le \theta \le 10$ by solid prior knowledge and *y* is some dependent variable of interest. Assume that this is viewed as a very complex model requiring simulation to solve. The community understands that no model is perfect but no specific flaws of this one are known. Extant data for *y* ranges from 1.31 to 1.61 and yields the point estimate $\hat{\theta} = 5.79$ (RMSE = 0.12). Due to its complexity, no one has noticed that the model cannot reproduce any $y > \arctan(10) = 1.47$ or, for any θ, a $y > \pi/2 \approx 1.57$. Now suppose that: a very large set of new *y* data is collected. Depending on the distribution of the new data either: (1) a new $\hat{\theta} < 10$ will be found or (2) $\hat{\theta}$ will rise significantly above 10, leading to a rejection of the model. However, what will *not* happen is that the increase in data will enable observations >1.57 to be reproduced. The model simply lacks the expressivity to do so. Analogously, increasing the amount of anthesis date data may narrow GSP estimate confidence limits, but the reachable region of predicted phenotype space is unlikely to

extend beyond the edges of the light blue regions. Therefore, any improvement in the ability to predict the large numbers of red points in Fig. 4.7 and 4.8 is unlikely.

Given these issues, a sensible follow-up question might be about what specific GSP estimates were reported for the red points? Here we report answers only for P1.

Fig. 4.9 shows scatterplots of P1 and P2O estimates generated using data from NY6 and NY7 via the database search and DE. The color coding is consistent with that in Fig. 4.7a. The pronounced bands at ca. P1=250 in both panels are immediately striking – although the scale is small, a corresponding band is quite evident at the same position in Fig. 4.6. A tabulation reveals that, of all 4,731 lines represented in the Fig. 4.9a, 3,227 (68.2%) have estimates of P1 ranging from 245 to 260. Of these, 1,493 are expressible (yellow) and 1,734 (red) are not expressible. Out of the total 4,731 points in the graph 2,189 (46.2%) are expressible and 2542 (53.8%) not. The Fig. 4.9b has similar proportions of expressible and inexpressible points (2327, 49.1%; and 2404, 50.9%; respectively), reinforcing the similarity of results for parameter estimates from DE and database searches. The differences are likely due to the ability of DE to explore the parameter space continuously whereas the database search is restricted to the predefined discrete Sobol points. Still, one may wonder why so many P1 estimates are near the 250 degree-days? Fig. 4.10 reveals the answer. The numbers in black are the "first-best-found" P1 estimated values that generate the corresponding row × column anthesis date combinations. A comparison with the corresponding dot colors and sizes in Fig. 4.5b indicates that, on the frontier (red borders Fig. 4.5a,b and 4.10) between expressible and inexpressible observations, there was essentially no equifinality and, concomitantly, narrow ranges of P1 values. Fig. 4.10 shows that of the P1 values along the frontier were all quite close to 250. For lines with observations falling outside the frontier, the RMSE was minimized by assigning GSP values associated with the closest achievable

dates, i.e. those directly on the frontier. Therefore, all the lines counted by the red numbers were assigned P1 values that are very close to 250 and have essentially no equifinality. The green arrow in Fig. 4.10 illustrates this phenomenon for one line. The nearest P1 estimate is 254 and the length of the arrow (ca. 5.8 days) is proportional to that line's RMSE. Specifically, in this case the length is $1/\sqrt{2}$ times the RMSE because there are $n = 2$ site-years.

Recall that the upper limit placed on P1 was 450 (and 600 in the DE search), therefore this outcome is likely not an artifact of constraints in the GSP search space but, rather, a result of poor model expressivity, that is the model inability to predict anthesis date pairs beyond those on the frontier. This mechanism accounts for the P1 band at 250 in Fig. 4.9a. Furthermore, as previously presented, more data cannot improve the prediction of inexpressible lines, the banding in Fig. 4.6 is not surprising.

4.4.5 P2O gap

We now investigate the vertical gap in scatterplots involving P2O estimates (Fig. 4.6), which documents the intricacy of the interactions that can occur between model mechanisms, parameter ranges searched, optimization algorithms used, and environments included. Exploratory re-tabulations of the Sobol-based parameter database revealed that the P2O gap was clearly present in the three site-years having shorter day lengths (FL6, FL7, and PR6) but absent in fits obtained by only including the remaining eight site-years with longer days (Fig. 4.11). Fig. 4.12 shows that a substantial number of observations for short-day site-years are outside the predicted phenotype ranges expressible by the model under either database or DE optimization. As described in section 4.3.2, the model operated by calculating the number of leaves initiated by the end of Stage 2 and predicts anthesis only after leaves are fully emerged. For any line, leaf number was a constant across all site-years, namely P1/(2×PHINT)+5. The variation of anthesis dates across plantings

was such that there were few, if any, combinations of P1 and PHINT that were compatible with the data from all site-years. Therefore, the optimizer relied more heavily on the P2 and P2O parameters.

Specifically, the optimizer settled on very small P2O estimates, much smaller than the short southern photoperiods. Instead, the optimizer relied on P2 estimates to generate anthesis date predictions that were delayed to the greatest extent possible by lengthening Stage 2. Recall that P2O values above the day length make Stage 2 only four days long, which is not enough time for temperature differences to accumulate the needed variation. The abundance of low P2O estimates thus created the gap observed in scatterplots of P2O with other GSPs (Fig. 4.11a). In contrast, the photoperiods in the remaining longer-day site-years exceeded the maximum allowed P2O values in the P2O database search during (and long after) the juvenile period. Therefore, there was no empty band in the scatter plot (Fig. 4.11b) because the optimizer was able to exploit delays for any value of P2O.

With the broader range of parameter values available to the DE runs and the increased flexibility available between P1 and PHINT, other options became available. In particular, in many cases DE found GSP combinations wherein P2O exceeded the southern day lengths so photoperiod had no influence on anthesis date and no gap artifact was generated (Fig. 4.11d,i). P1 and PHINT thus became the major explanatory parameters. This is shown in Fig. 4.13, whereby for each line, the parameter differences are plotted against the RMSE differences that result from changing the estimation methods from database to DE optimization. The DE estimate of P2O were larger in 4,507 out of 5,240 lines (87%; Fig. 4.13d), almost always by enough to put it above the local day lengths. In tandem, P1 values fell in 3,559 lines (Fig. 4.13a), whereas PHINT rose in 4,102 lines (Fig. 4.13c).

Note, however, that for *any* (P1, PHINT) combination, *any* P2O that exceeds the local day length will give the same RMSE – a clear source of equifinality. Thus, the changes in P2O will not, in all likelihood, lead to values that can be more closely related to genetics. Moreover, because of the limits on model expressivity, none of the DE solutions gave significantly better fits than the database estimates. This is why virtually all points in Fig. 4.13 had DE RMSE's within 0.5 days (horizontal axes) of the database-based parameter estimates. This, too, is an illustration of equifinality because the two optimizers were finding different GSP estimates although the RMSE were of similar magnitude.

4.4.6 Tests for stability of GSP estimates

Table 4.5 shows the effect of including or excluding the effect of different subsets of site-years on the modeling of estimates (Eq. 1) for each GSP when all 177,870 parameter estimates are used (ties+ no ties). For all GSP parameters, AIC and BIC values were considerably smaller for models that included the random effect of site-year subsets, β_e, therefore suggesting non-negligible variability across site-year subsets on the GSP estimates. The table provides indicators to illustrate the size of the effects. For scaling purposes, the grand mean column contains the average parameter value across all lines and site-year subsets. The Index of Variability (expressed as a percent) is the standard deviation of the β_e effect normalized by the grand mean. The percentage of the total GSP variance attributable to site-year subsets is also shown. Both of these numbers are substantial with variability indexes ranging from 5.9% for P2O to 33.6% for P2 and variance fractions all in excess of 20%.

The Chi square values from the likelihood ratio test and the associated *p*-values are in the last two columns of Table 4.5. For each GSP, the estimated values differed depending on the subset of site-years used to estimate them and, therefore, are not, in fact, genotype specific despite the

goodness-of-fit displayed in Figure 4.3. This result is completely understandable given the range of artifacts due to equifinality and model expressivity issues identified above.

Table 4.6 shows the results when estimates having (lacking) ties are tested separately. These two groups correspond to the parameter subsets that, respectively, fall inside (outside) the expressivity frontier. It is clear, however, that the grand means, index of variability, and percentages of GSP variance are highly similar between all three groupings in Table 4.5&4.6. Also, all p-values are extremely significant and increase with the amount of data used (from right to left in 4.6 and from 4.6 to 4.5).

4.5 Discussion

Since their inception, ecophysiological models have been evaluated in terms of predictive ability, which are superb in many circumstances (Batchelor et al., 2002). The parameters that drove the models were considered to be *inputs* whose genesis was of secondary importance as long as the model outputs proved useful. However, as often happens in science, perceived needs, desiderata, and requirements escalate as technologies evolves. In particular, we are now demanding that the model inputs themselves be the accurate outputs of processes at the genetic level that can be modeled by genomic prediction. It is not surprising, therefore, that modeling technologies (ranging from data collection to estimation) that were adequate for past applications now require improvement.

From a fundamental but traditional perspective, there are several issues of perennial concern in crop modeling. The first is model functional structure including both its degree of expressivity and its behavior under optimization. For example, estimation procedures like DE, that primarily yield point estimates, are limited in their ability to assess equifinality. At best, one can query the flatness of the goodness-of-fit function in the neighborhood of the estimate, but this does

not tell anything about the ubiquity of equifinality across the parameter space. Nor do these procedures allow one to detect observations that fall outside of the model's scope of expressivity unless the discrepancies are quite large. Doing so requires methods like the Sobol database scheme used here that can make broader assessments in both parameter and phenotype space. It may well be that the rarity with which database methods have been used has led to an underappreciation as to the prevalence of these adverse situations.

When expressivity issues are identified, results like those above are not likely to be solved merely by acquiring more data of the same type. In such situations, better models will often be needed and modern genetic studies can help. A great many plant component subsystems are currently under study at the molecular level. Indeed, some of these (e.g., Chew et al., 2014) are even being combined into multi-scale organ and whole plant models. Even without modeling directly at the genetic level one can use the derived insights to make informed choices between alternative representations of individual ecophysiological processes. Tardieu (2003) refers to such representations as "meta-mechanisms". It would seem plausible that building models from component parts of increased biological realism should increase the ability to reproduce field variation – at the very least, it is hard to see how it can hurt. As a concrete example, the B73 parent is photoperiod insensitive. In CERES-Maize, however, the only way to express this is by setting P2O in excess of the observed photoperiods, with the consequences we have seen.

This is not to say, however, that both more and better data are not needed. Indeed, data quality issues can impact both expressivity and GSP stability. For example, while the date seed that are physically sown in a field is usually known and not subject to error, researchers often report a subjective notion of "effective sowing date" based on their interpretation of whether low soil moisture delayed germination. If errors in sowing date push an anthesis observation across the

expressivity frontier erroneous GSP estimates will result. Such errors can also arise if different personnel are involved across locations or growing seasons, especially for visually evaluated phenotypes like most phenological traits. Providing the emergence date can provide a partial check for these problems and also for errors in simulating time from sowing to emergence. Unfortunately, emergence dates were not reported for the maize NAM dataset.

Another traditional modeling concern has always been the relationship between the observed environmental data and the immediate environmental conditions actually experienced by individual plants. Weather data can suffer from multiple sources of bias and error (Fall et al., 2011). For example, stations that are not located within or directly adjacent to experiments may have bias due to local variation in weather conditions. Additionally, although of limited concern for anthesis dates, the quality of soil and management data. In this study any systematic differences in protocols for collection of weather data between the sites as aggravated by small sample effects, might have contributed to some degree to the significance levels in Table 4.5. It would certainly be desirable to have a method by which this potential effect might be quantitatively assessed. Such a method could be instrumental in designing experimental procedures for reducing the problem. One potential example might be to eschew external measurements of some environmental variables (e.g., air temperature) and use sensors onboard UAV's or other automated vehicles to measure plant temperatures or other critical features directly at high temporal and spatial frequencies.

More involved data types and structures are also needed to resolve issues of equifinality when they arise. Equifinality is fundamentally a problem of discernment. In simple terms, given an equation $c = a + b$, if one only has data on *c*, then estimates of *a* and *b* are doomed to be equifinal. If one desires otherwise, one must find a way to measure either *a* or *b*. Current

technological efforts to develop high throughput phenotyping approaches might be quite helpful in this regard. For example, assuming that TOLN=P1/(PHINT×2)+5 is the correct way to model the number of leaves at anthesis, data on total leaf number would help constrain the parameter estimates. This leads toward a range of constrained and/or multiobjective estimation procedures on which there has been significant amounts of research (Rabotyagov et al., 2012; Tatsumi, 2016). Maximum entropy methods offer another opportunity wherein one identifies a probability distribution of values that is constrained but mathematically no more informed than is justified by a set of potentially diverse data types (Hess et al., 2002). Another alternative might be Bayesian methods with multivariate likelihood functions that combine several observational variables (Franks et al., 1999).

Another approach to resolving equifinality might be to use simpler models. The fewer the number of processes and GSP's in a model, the smaller the opportunity for hard-to-spot tradeoffs to exist wherein adjustments to one parameter can be offset by tweaking another one. Of course, the tradeoff can be less expressivity leading to other problems, as we have seen. However, Welch et al. (2005) presented 12 dichotomies comparing gene network modeling and quantitative genetics approaches, where aspects of the former might also apply to ecophysiological modeling. They opined that an optimal modeling approach should entail a synthesis of both. The key features to be contributed from the network (i.e., ecophysiological) side would be (1) the ability to handle time-varying dynamics, (2) a far more parsimonious approach to expressing biological and biology × environmental interactions, and (3) a more mechanistic explanation of how traits originate. It is at least conceivable that some way station of moderate complexity exists between statistical genetics and full crop models that can achieve this.

At whatever level of complexity proves appropriate, one cannot accurately estimate the parameters controlling model components without collecting data on settings wherein the relevant processes operate differentially. This is clear from the P2O gap phenomenon, which was apparent when only short day data was used and absent under long days. Both settings distorted the results, in one case compressing estimates into a restricted range, leaving a gap, and, in the other, allowing them to spread out. Furthermore, this interacted with the range of values allowed, which caused shifts between (P1, PHINT) and (P2, P2O) as to which parameters appeared to be "explanatory". The debilitating influence of such behavior on linking parameter values to genes is terribly obvious.

However, it also should not escape notice that the gap was evident even in a mixture of environments, suggesting that good experimental design entails more than just making sure that a suitable range of environments is included. There is some notion of balance that needs to be established and applied globally to data selection. In this context, it is worth noting that despite the fact that thousands of lines were planted in each location, there were only 539 lines where data were reported from all 11 trials. However, given the expense of such large-scale trials and the multiple purposes each one will serve, "balance" cannot mean "orthogonality" where all lines are planted at all sites. Of course, an established benefit of ecophysiological models is to serve as guides to help prioritize experimentation over time. It seems likely that as their integration with statistical genetic models expands, they might also be able to assist in the rational planning and resource allocation for large, multi-site trials.

Another approach entirely would be to seek to move beyond a two-step "estimate and then map" paradigm. Conventional mapping methods essentially isolate genetic markers whose pattern of assignment to lines mirrors the pattern of phenotype values of interest. A general linear model

is assumed to mediate between marker states and realized phenotypes. There is no conceptual reason why that general linear model might not be replaceable by a crop model. In effect, one could conceive a hierarchical model in which a first-level model is specified on the data and higher order submodels are specified on the parameters that characterize the behavior of observed data, much like proposed by Bello et al. (2010).

One could conceptually implement this hierarchy in the context of crops by to fitting phenotypes with an ECM whose GSP's are then specified as functions of genetic markers at another level of the hierarchical model. Indeed, this is what the current paradigm attempts, except that the two-step estimation process curtails smooth borrowing of information across hierarchical levels of the model that could potentially help resolve the equifinality problem.

We acknowledge that one-step hierarchical model approach might not solve the sort of expressivity problems described in the thought experiment and documented in our results (both in 4.4). Yet, it would enable the genetic structure of the population to inform the GSP estimation process. The potential utility of this hierarchical modeling approach is currently under study in one of our labs. The approach would also enable more efficient use of data. Currently, the two-step approach requires data from multiple environments (Welch et al., 2002) for each line in order to estimate the GSP's before mapping can proceed. However, consider a line that was culled very early in the selection process, perhaps even after a single round. Because the parameters estimated in putative one-step hierarchical modeling schemes would include marker effects, even just one planting becomes a usable observation if the line is genotyped. This is a sufficiently inexpensive operation now that some programs (e.g. CIMMYT; Battenfield et al., 2016; Gaynor, 2015) are doing so routinely for the offspring of all crosses.

A one-step hierarchical modeling approach might also make it possible to utilize data taken on lines after they enter the market place. Analogously to high throughput phenotyping in breeding programs, precision agricultural management is also investing in sensor- and model-based approaches to improve productivity (Mohanty, 2013; Thompson et al., 2015; Thorp et al., 2015a; Thorp et al., 2015b) while collecting a wealth of multivariate data. Usually, of course, hybrids are released into areas where they show low G×E interactions. For example, a line with a particular P2O is not likely to be released across a sufficient range of latitudes to have great differences in day length. This would make it difficult to directly estimate P2O for the line using the methods described in this paper.

However, in a one-step hierarchical model approach, one would only be looking for markers that influenced P2O. In this case, data from many lines and geographical areas could be used together. This would also make such data usable for the sorts of hypothesis testing about genes discovered by other means, thus facilitating genetically-informed ecophysiological modeling. For such approaches to be workable, however, there are many policy issues to be resolved including information property rights and fair economic returns to data, not to mention the need to greatly harden cybersecurity protections (FBI, 2016). However, if this can be done then issues of environmental coverage would likely be ameliorated due to the extent of the data that would become available.

4.6 Conclusions

The original and seemingly simple goal of this study was to first fit the anthesis date component of the CERES-Maize model to data from over 5000 genotyped lines and then genetically map the resulting GSP values. However, we were unexpectedly detoured when we found that despite the high predictive quality of the values obtained, there were numerous artifacts

that emerged in the estimation process, thereby making our immediate goal unachievable. We find it interesting that the problems we encountered would likely be invisible, though present, in smaller data sets and, unless addressed by suitable research, these problems bode ill for understanding any genetic underpinnings of ecophysiological models. This is worrisome given the recent escalating attention that has been given to this method of melding ecophysiological and statistical genetic models as a way of accelerating the crop improvement process so as to help meet global food and fiber needs by 2050.

The constraining issues fall into two categories. The first arises in situations where the model is unable to express the observed data for some line even by a relatively few days. In this circumstance, the line is assigned the GSP associated with the nearest point on model's expression frontier – values which can, however, change only slowly along that boundary. The result is that many and in some cases a large majority of lines are assigned the same GSP values independent of their actual genetics.

The second symptom arises when the model can reproduce the data. In these instances, there can be many combinations of GSP values that predict equally well. When such equifinality exists, there is no principled way to assign the line a genetically relevant value. In short, when the model can express the data there is no unique combination of GSP values and, when unique combinations do exist they are often values being given to many lines because of a deficiency in model expressivity.

This finding is rather remarkable because in both breeding efforts and, indeed, genetic studies as a whole, anthesis date is considered, if not a simple trait, at least one that has proved much easier to elucidate than many others. In addition, it is generally, much more readily predicted by classical phenology models for reasons that, themselves, have become generally understood

(Wilczek et al., 2009). This cannot but make one wonder, what pitfalls might lie in wait for efforts to probe other, more involved traits.

Therefore, the next question to be asked by follow-on research is how prevalent are these phenomena. The best way to do that would seem to be to use Sobol database search methods. This is because, unlike optimizers that find single "best estimates", the database approach will reveal the both the extent of the expressible phenotype regions as well as a direct measure of the extent of any equifinality.

However, despite the ability to reuse results databases for many searches, undertaking such a program in any broadly based fashion will be highly demanding computationally. For this reason, strong consideration should be given to disaggregating comprehensive models into separate modules that can be studied independently at much lower computational cost. (This is what we did for the limited DE run, although Python certainly is not a high performance language.) A better long-term strategy would be to program future models in a manner that supports single-module testing at the source code level. Doing so will facilitate the whole-model verifications needed to ensure that fragmentation into modules for testing and improvement by different labs does not compromise integration at the level of the scientific community.

As module testing and innovation progress, it will be of strategic value to ground improvements in advancing genetic understanding at the molecular level. While this might seem daunting to those versed in purely physiological approaches, it need not be so. One of the most venerable concepts in all of the life sciences is that of the biological hierarchy that is, a series of many functional levels extending from molecules to the biosphere. One of the perspectives emerging from molecular science is that that hierarchy might, be operationally much flatter than commonly believed. That is, simple changes at lower levels can easily create tangible responses

multiple levels higher. To the extent that this is true, it greatly reduces the complexity of bridging across those levels. This is the philosophy behind the meta-mechanism approach mentioned earlier (Tardieu, 2003; Tardieu and Tuberosa, 2010).

That approach has a proven ability to account for environmental interactions with sufficient skill to eliminate observed G×E interactions from GSP's in the data sets used (Reymond et al., 2003). However, as shown by the *p*-values in Table 5, the very large data set used herein conveyed an extraordinary power to detect site-year dependencies in GSP estimation. Indeed, so powerful as to make one wonder if an insignificant result is scientifically achievable by any even remotely feasible research effort? A better number to use for practical evaluations might be the index of variability in Table 4.5. This would give a clear index of the size of the effect as a percentage of the parameter values. Also, means exist for comparing such indices to see if reductions in their values (i.e. by an improved model with lowered site-year set dependency) are statistically significant (Vangel, 1996).

A final message from our research is that one cannot fix problems that one does not know exist. Community interest in the fitting-and-mapping paradigm has been high as is shown by the heavy citation rates for the seminal papers in this area. For example, as of September, 2016, the Hammer et al. (2006) paper had been cited 257 times and those publications, *themselves*, had been cited by 6,370 others (Source: Google Scholar). There is also no doubt as to the importance of the ability to predict the behaviors of novel genotypes in novel environments while crosses are still in the planning stage. Indeed, this is precisely the genotype-to-phenotype problem, which has been declared by the National Research Council to be a top-priority goal for applied biology (NRC, 2008). So these impediments need to be overcome. However, with methods now in hand to detect adverse model behaviors under estimation, research that is probing ever more deeply into the

control mechanisms of plant growth and development, and concrete tests to document model improvements, there is no reason to believe that we cannot do so.

4.7 Acknowledgements

The plan to use the maize NAM data to first developed through discussions with iPlant (www.**iplant**collaborative.org) on novel applications of cyber infrastructure in plant science. The authors acknowledge the Texas Advanced Computing Center (TACC; http://www.tacc.utexas.edu) at The University of Texas at Austin and Beocat, Kansas State University for providing high performance computing resources that have contributed to the research results reported within this paper. Support for this effort was also supplied by the Department of Agronomy at Kansas State University. Additional support derived from a Higuchi-KU Endowment Research Achievement Award through the University of Kansas and the University of Kansas Endowment Association. This paper is contribution number 17-134-J of the Kansas Agricultural Experiment Station at Kansas State University, Manhattan, KS.

4.8 Reference

Akaike, H., 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 60, 255–265.

Andrés, F., Coupland, G., 2012. The genetic basis of flowering responses to seasonal cues. Nat. Rev. Genet. 13, 627–639.

Batchelor, W.D., Basso, B., Paz, J.O., 2002. Examples of strategies to analyze spatial and temporal yield variability using crop models. Eur. J. Agron. 18, 141–158.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2014. Fitting linear mixed-effects models using lme4. ArXiv Prepr. ArXiv14065823.

Battenfield, S.D., Guzmán, C., Gaynor, R.C., Singh, R.P., Peña, R.J., Dreisigacker, S., Fritz, A.K., Poland, J.A., 2016. Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. Plant Genome 9.

Bello, N.M., Steibel, J.P., Tempelman, R.J., 2010. Hierarchical Bayesian modeling of random and residual variance–covariance matrices in bivariate mixed effects models. Biom. J. 52, 297–313.

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320, 18–36.

Bratzel, F., Turck, F., 2015. Molecular memories in the regulation of seasonal flowering: from competence to cessation. Genome Biol. 16, 1.

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C., others, 2009. The genetic architecture of maize flowering time. Science 325, 714–718.

Burhenne, S., Jacob, D., Henze, G.P., 2011. Sampling based on Sobol'sequences for Monte Carlo techniques applied to building simulations, in: Proc. Int. Conf. Build. Simulat. pp. 1816–1823.

Chenu, K., Chapman, S.C., Tardieu, F., McLean, G., Welcker, C., Hammer, G.L., 2009. Simulating the Yield Impacts of Organ-Level Quantitative Trait Loci Associated With Drought Response in Maize: A "Gene-to-Phenotype" Modeling Approach. Genetics 183, 1507–1523. doi:10.1534/genetics.109.105429

Chew, Y.H., Wenden, B., Flis, A., Mengin, V., Taylor, J., Davey, C.L., Tindal, C., Thomas, H., Ougham, H.J., de Reffye, P., others, 2014. Multiscale digital Arabidopsis predicts individual organ and whole-organism growth. Proc. Natl. Acad. Sci. 111, E4127–E4136.

Colasanti, J., Coneva, V., 2009. Mechanisms of floral induction in grasses: something borrowed, something new. Plant Physiol. 149, 56–62.

Coles, N.D., McMullen, M.D., Balint-Kurti, P.J., Pratt, R.C., Holland, J.B., 2010. Genetic control of photoperiod sensitivity in maize revealed by joint multiple population analysis. Genetics 184, 799–812.

Cooper, M., Technow, F., Messina, C., Gho, C., Totir, L.R., 2016. Use of Crop Growth Models with Whole-Genome Prediction: Application to a Maize Multienvironment Trial. Crop Sci.

Das, S., Suganthan, P.N., 2011. Differential evolution: a survey of the state-of-the-art. IEEE Trans. Evol. Comput. 15, 4–31.

Dong, Z., Danilevskaya, O., Abadie, T., Messina, C., Coles, N., Cooper, M., 2012. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. PLoS One 7, e43450.

Fall, S., Watts, A., Nielsen-Gammon, J., Jones, E., Niyogi, D., Christy, J.R., Pielke, R.A., 2011. Analysis of the impacts of station exposure on the US Historical Climatology Network temperatures and temperature trends. J. Geophys. Res. Atmospheres 116.

FBI, 2016. Smart Farming May Increase Cyber Targeting Against US Food and Agriculture Sector (No. 160331-001). United States Federal Bureau of Investigation. Private Industry Notification.

Flint-Garcia, S.A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S.M., Mitchell, S.E., Doebley, J., Kresovich, S., Goodman, M.M., Buckler, E.S., 2005. Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J. 44, 1054–1064.

Franks, S., Beven, K.J., Quinn, P., Wright, I., 1997. On the sensitivity of soil-vegetation-atmosphere transfer (SVAT) schemes: equifinality and the problem of robust calibration. Agric. For. Meteorol. 86, 63–75.

Franks, S.W., Beven, K.J., Gash, J.H., 1999. Multi-objective conditioning of a simple SVAT model. Hydrol. Earth Syst. Sci. Discuss. 3, 477–488.

Gaynor, R.C., 2015. Genomic selection for Kansas wheat. Kansas State University.

Gill, P.E., Murray, W., Wright, M.H., 1981. Practical optimization. Academic Press.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C., 2010. Food security: the challenge of feeding 9 billion people. science 327, 812–818.

Grimm, S.S., Jones, J.W., Boote, K.J., Hesketh, J.D., 1993. Parameter estimation for predicting flowering date of soybean cultivars. Crop Sci. 33, 137–144.

Hammer, G., Cooper, M., Tardieu, F., Welch, S., Walsh, B., van Eeuwijk, F., Chapman, S., Podlich, D., 2006. Models for navigating biological complexity in breeding improved crop plants. Trends Plant Sci. 11, 587–593.

Hammer, G.L., Kropff, M.J., Sinclair, T.R., Porter, J.R., 2002. Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. Eur. J. Agron., Process Simulation and Application of Cropping System Models 18, 15–31. doi:10.1016/S1161-0301(02)00093-X

Hammer, G.L., Woodruff, D.R., Robinson, J.B., 1987. Effects of climatic variability and possible climatic change on reliability of wheat cropping—a modelling approach. Agric. For. Meteorol. 41, 123–142.

Harrison, S.R., Tamaschke, H.U., 1984. Applied statistical analysis. Prentice-Hall of Australia.

He, J., Dukes, M.D., Jones, J.W., Graham, W.D., Judge, J., 2009. Applying GLUE for estimating CERES-Maize genetic and soil parameters for sweet corn production. Trans. ASABE 52, 1907–1921.

He, J., Jones, J.W., Graham, W.D., Dukes, M.D., 2010. Influence of likelihood function choice for estimating crop model parameters using the generalized likelihood uncertainty estimation method. Agric. Syst. 103, 256–264. doi:10.1016/j.agsy.2010.01.006

Hess, C.P., Liang, Z.-P., Lauterbur, P.C., 2002. Maximum cross-entropy generalized series reconstruction, in: 5th IEEE EMBS International Summer School on Biomedical Imaging, 2002. Presented at the 5th IEEE EMBS International Summer School on Biomedical Imaging, 2002, p. 8 pp.–. doi:10.1109/SSBI.2002.1233979

Hoogenboom, G., Jones, J.W., Wilkens, P.W., Porter, C.H., Hunt, L.A., Singh, U., Lizaso, I., White, J., Uryasev, O., Ogoshi, R.M., Koo, J., Shelia, V., Tsuji, G.Y., 2015. Decision Support System for Agrotechnology Transfer (DSSAT) version 4.5 (htttp://dssat.net)., DSSAT Foundation. Prosser, Washington.

Hung, H.-Y., Shannon, L.M., Tian, F., Bradbury, P.J., Chen, C., Flint-Garcia, S.A., McMullen, M.D., Ware, D., Buckler, E.S., Doebley, J.F., Holland, J.B., 2012. ZmCCT and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proc. Natl. Acad. Sci. 109, E1913–E1921. doi:10.1073/pnas.1203189109

Hunt, L., White, J., Hoogenboom, G., 2001. Agronomic data: advances in documentation and protocols for exchange and use. Agric. Syst. 70, 477–492.

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

Jones, C.A., Richie, J.T., Kiniry, J.R., Godwin, D.C., 1986. Subroutine structure. CERES-Maize Simul. Model Maize Growth Dev. CA Jones JR Kiniry Contrib. PT Dyke Al.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. Eur. J. Agron. 18, 235–265.

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N., Meinke, H., Hochman, Z., others, 2003. An overview of APSIM, a model designed for farming systems simulation. Eur. J. Agron. 18, 267–288.

Kennedy, J., 2011. Particle swarm optimization, in: Encyclopedia of Machine Learning. Springer, pp. 760–766.

Kiniry, J.R., Bonhomme, R., 1991. Predicting maize phenology. Predict. Crop Phenol. 11, 5–131.

Klepper, O., Hendrix, E.M., 1994. A method for robust calibration of ecological models under different types of uncertainty. Ecol. Model. 74, 161–182.

Koduru, P., Welch, S.M., Das, S., 2007. A particle swarm optimization approach for estimating parameter confidence regions, in: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation. ACM, pp. 70–77.

Laurila, H., Mäkelä, P., Kleemola, J., Peltonen, J., 2012. A comparative ideotype, yield component and cultivation value analysis for spring wheat adaptation in Finland. Agric. Food Sci. 21, 384–408.

Lee, M., Sharopova, N., Beavis, W.D., Grant, D., Katt, M., Blair, D., Hallauer, A., 2002. Expanding the genetic map of maize with the intermated B73$\times$ Mo17 (IBM) population. Plant Mol. Biol. 48, 453–461.

Luo, Y., Weng, E., Wu, X., Gao, C., Zhou, X., Zhang, L., 2009. Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. Ecol. Appl. 19, 571–574.

Major, D.J., Kiniry, J.R., 1991. Predicting daylength effects on phenological processes. Predict. Crop Phenol. 15–28.

Mascheretti, I., Turner, K., Brivio, R.S., Hand, A., Colasanti, J., Rossi, V., 2015. Florigen-encoding genes of day-neutral and photoperiod-sensitive maize are regulated by different chromatin modifications at the floral transition. Plant Physiol. 168, 1351–1363.

Matsumoto, M., Nishimura, T., 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. TOMACS 8, 3–30.

Mavromatis, T., Boote, K., Jones, J., Wilkerson, G., Hoogenboom, G., 2002. Repeatability of model genetic coefficients derived from soybean performance trials across different states. Crop Sci. 42, 76–89.

McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S.E., Peterson, B., Pressoir, G., Romero, S., Rosas, M.O., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J.C.,

Goodman, M., Ware, D., Holland, J.B., Buckler, E.S., 2009. Genetic Properties of the Maize Nested Association Mapping Population. Science 325, 737–740. doi:10.1126/science.1174320

Medlyn, B.E., Robinson, A.P., Clement, R., McMurtrie, R.E., 2005. On the validation of models of forest CO2 exchange using eddy covariance data: some perils and pitfalls. Tree Physiol. 25, 839–857.

Meuwissen, T.H., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Mohanty, B.P., 2013. Soil hydraulic property estimation using remote sensing: A review. Vadose Zone J. 12.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7, 308–313.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical recipes in FORTRAN: the art of scientific computing. Cambridge University Press Cambridge.

Quiring, S.M., Legates, D.R., 2008. Application of CERES-Maize for within-season prediction of rainfed corn yields in Delaware, USA. Agric. For. Meteorol. 148, 964–975.

Rabotyagov, S., Campbell, T., Valcu, A., Gassman, P., Jha, M., Schilling, K., Wolter, C., Kling, C., 2012. Spatial multiobjective optimization of agricultural conservation practices using a SWAT model and an evolutionary algorithm. J. Vis. Exp. JoVE e4009. doi:10.3791/4009

Reymond, M., Muller, B., Leonardi, A., Charcosset, A., Tardieu, F., 2003. Combining quantitative trait Loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiol. 131, 664–675. doi:10.1104/pp.013839

Román-Paoli, E., Welch, S.M., Vanderlip, R.L., 2000. Comparing genetic coefficient estimation methods using the CERES-Maize model. Agric. Syst. 65, 29–41.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S., 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput. Phys. Commun. 181, 259–270.

Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Stat. 6, 461–464. doi:10.1214/aos/1176344136

Semenov, M.A., Stratonovitch, P., 2013. Designing high-yielding wheat ideotypes for a changing climate. Food Energy Secur. 2, 185–196.

Sobol, I.M., 1998. On quasi-Monte Carlo integrations. Math. Comput. Simul. 47, 103–112. doi:10.1016/S0378-4754(98)00096-2

Stone, T., 2011. Sustainability and the needs of 2050 agriculture: Developed and developing world perspectives (NABC No. 23), Food Security: The Intersection of Sustainability, Safety and defense.

Tardieu, F., 2003. Virtual plants: modelling as a tool for the genomics of tolerance to water deficit. Trends Plant Sci. 8, 9–14.

Tardieu, F., Tuberosa, R., 2010. Dissection and modelling of abiotic stress tolerance in plants. Curr. Opin. Plant Biol. 13, 206–212. doi:10.1016/j.pbi.2009.12.012

Tatsumi, K., 2016. Effects of automatic multi-objective optimization of crop models on corn yield reproducibility in the U.S.A. Ecol. Model. 322, 124–137. doi:10.1016/j.ecolmodel.2015.11.006

Technow, F., Messina, C.D., Totir, L.R., Cooper, M., 2015. Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PloS One 10, e0130855.

Thompson, L.J., Ferguson, R.B., Kitchen, N., Frazen, D.W., Mamo, M., Yang, H., Schepers, J.S., 2015. Model and Sensor-Based Recommendation Approaches for In-Season Nitrogen Management in Corn. Agron. J. 107, 2020–2030.

Thorp, K.R., DeJonge, K.C., Kaleita, A.L., Batchelor, W.D., Paz, J.O., 2008. Methodology for the use of DSSAT models for precision agriculture decision support. Comput. Electron. Agric. 64, 276–285.

Thorp, K.R., Gore, M.A., Andrade-Sanchez, P., Carmo-Silva, A.E., Welch, S.M., White, J.W., French, A.N., 2015a. Proximal hyperspectral sensing and data analysis approaches for field-based plant phenomics. Comput. Electron. Agric. 118, 225–236.

Thorp, K. R., Hunsaker, D. J., French, A. N., Bautista, E., & Bronson, K. F. (2015b). Integrating geospatial data and cropping system simulation within a geographic information system to analyze spatial seed cotton yield, water use, and irrigation requirements. *Precision Agriculture*, *16*(5), 532-557.

Valentim, F.L., van Mourik, S., Posé, D., Kim, M.C., Schmid, M., van Ham, R.C., Busscher, M., Sanchez-Perez, G.F., Molenaar, J., Angenent, G.C., others, 2015. A quantitative and dynamic model of the Arabidopsis flowering time gene regulatory network. PloS One 10, e0116973.

Wallach, D., Goffinet, B., Bergez, J.-E., Debaeke, P., Leenhardt, D., Aubertot, J.-N., 2001. Parameter Estimation for Crop Models. Agron. J. 93, 757. doi:10.2134/agronj2001.934757x

Welch, S.M., Dong, Z., Roe, J.L., Das, S., 2005a. Flowering time control: gene network modelling and the link to quantitative genetics. Crop Pasture Sci. 56, 919–936.

Welch, S.M., Roe, J.L., Das, S., Dong, Z., He, R., Kirkham, M.B., 2005b. Merging genomic control networks and soil-plant-atmosphere-continuum models. Agric. Syst. 86, 243–274.

Welch, S.M., Wilkerson, G., Whiting, K., Sun, N., Vagts, T., Buol, G., Mavromatis, T., 2002. Estimating soybean model genetic coefficients from private–sector variety performance trial data. Trans. ASAE 45, 1163.

Welch, S.M., Zhang, J., Sun, N., Mak, T.Y., 2000. Efficient estimation of genetic coefficients of crop models, in: The Third International Symposium on System Approaches for Agricultural Development.

White, J.W., Hoogenboom, G., 2010. Crop response to climate: ecophysiological models, in: Climate Change and Food Security. Springer, pp. 59–83.

White, J.W., Hoogenboom, G., 1996. Simulating effects of genes for physiological traits in a process-oriented crop model. Agron. J. 88, 416–422.

Wilczek, A.M., Roe, J.L., Knapp, M.C., Cooper, M.D., Lopez-Gallego, C., Martin, L.J., Muir, C.D., Sim, S., Walker, A., Anderson, J., Egan, J.F., Moyers, B.T., Petipas, R., Giakountis, A., Charbit, E., Coupland, G., Welch, S.M., Schmitt, J., 2009. Effects of genetic perturbation on seasonal life history plasticity. Science 323, 930–934. doi:10.1126/science.1165826

Wit, C.T. de., 1965. Photosynthesis of leaf canopies, Agricultural Research Reports ; 663; Verslagen van Landbouwkundige Onderzoekingen ; 663. Centre for Agricultural Publications and Documentation, Wageningen.

Yin, X., Kropff, M.J., Stam, P., 1999. The role of ecophysiological models in QTL analysis: the example of specific leaf area in barley. Heredity 82, 415–421.

Yin, X., Stam, P., Kropff, M.J., Schapendonk, A.H., 2003. Crop modeling, QTL mapping, and their complementary role in plant breeding. Agron. J. 95, 90–98.

Table 4.1 Sowing dates, geographical coordinates, total number of lines planted and number of lines for which anthesis dates were observed for all site-year combinations used in this study.

	NY6	**NY7**	**NC6**	**NC7**	**MO6**	**MO7**	**IL6**	**IL7**	**FL6**	**FL7**	**PR6**
Sowing Date (DOY)	128	135	122	120	137	138	128	137	265	280	314
Latitude (deg)	42.73	42.73	35.67	35.67	38.89	38.89	40.08	40.08	25.51	25.51	18.00
Longitude (deg)	-76.66	-76.66	-78.49	-78.49	-92.23	-92.23	-88.2	-88.2	-80.49	-80.49	-66.51
Number of total lines sown	5478	5478	5478	5478	5478	5478	5478	5478	5026	3753	5131
Number of lines with data	4743	5236	5236	5160	3261	2555	5036	5178	4943	3742	4401

Table 4.2 Parameter ranges used in generating Sobol sequence.

Parameter	**Definition**	**Unit**	**Min**	**Max**	**No. of unique values**
P1	Thermal time from seedling emergence to end of juvenile phase	GDD (^{o}C)	150	450	30,001
P2O	Critical photoperiod hour	hrs.	10	14	401
P2	Days of anthesis date delay for each hour by which the day length exceeds P2O	rate	0	2	20,001
PHINT	Phylochron interval (Interval between successive leaf tip appearances)	GDD (^{o}C)	25	70	45001

Table 4.3 Numbers of model expressible and inexpressible observations for selected site-year pairs.

Lines that are[a]	NY6/NY7	NC6/NC7	IL6/IL7	MO6/MO7	FL6/FL7
Expressible	2189	4964	2024	146	193
Inexpressible	2542	168	2946	637	3339

[a]These numbers refer to lines with data in both years of each pair and therefore do not precisely align with Table 4.1.

Table 4.4 Extended range of parameter values used for DE search.

Parameter	Definition	Unit	Min	Max	Percent of Sobol Range
P1	Thermal time from seedling emergence to end of juvenile phase	GDD (°C)	75	600	175%
P2O	Critical photoperiod hour	hrs.	6	21	300%
P2	Days of anthesis date delay for each hour by which the day length exceeds P2O	rate	0	6	375%
PHINT	Phylochron interval (Interval between successive leaf tip appearances)	GDD (°C)	20	110	200%

Table 4.5 Estimated likelihood, fit statistics, summary statistics, and a likelihood ratio test for competing statistical models fitted on GSP estimates with and without the random effect of site-year subset from all 177,870 data points

GSP	Log likelihood w/o (top) and w/ (bot) a site-year set effect[a]	AIC w/o (top) and w/ (bot) a site-year set effect[b]	BIC w/o (top) and w/ (bot) a site-year set effect[b]	GSP Grand Mean μ	Index of Variablility[c] σ_e/μ	Variance contributed by site-year sets[c] σ_e^2/σ_{tot}	Chi-square test statistic	Chi-square *p*-value[d] (df =0.5)
P1	-952735 -912485	1905475 1824979	1905506 1825019	270	11.48	29.94	80499	10^{-34955}
P2	-123924 -63454	247855 126917	247885 126957	0.9593	35.5	33.8	120940	10^{-52518}
P2O	-291181 -268373	58236 536754	582398 536794	12.42	4.88	21.27	45616	10^{-19806}
PHINT	-730099 -702761	1460204 1405530	1460234 1405570	43.94	17.3	24.35	54676	10^{-23740}

[a]Larger is better; [b]Smaller is better; [c] σ_e is the site-year-set std; values are percents; [d]Chernoff upper bound on Chi-squared cum. dist.

Table 4.6 Estimated fit statistics, summary statistics, for competing statistical models fitted on GSP estimates with and without the random effect of site-year subset from all data with only ties and without ties.

GSP	GSP Grand Mean μ	Index of Variablility[c] σ_e/μ	Variance contributed by site-year sets[c] σ_e^2/σ_{tot}	Chi-square *p*-value[d] (df =0.5)	GSP Grand Mean μ	Index of Variablility[c] σ_e/μ	Variance contributed by site-year sets[c] σ_e^2/σ_{tot}	Chi-square *p*-value[d] (df =0.5)
	With Ties				Without Ties			
P1	273.5	11.37	29.77	10^{-23283}	264.625	12.30	34.38	10^{-13334}
P2	0.9137	36.33	35.23	10^{-34723}	1.037	33.55	33.92	10^{-18163}
P2O	12.49	4.43	19.70	10^{-11883}	12.2440	5.88	27.83	10^{-8635}
PHINT	43.57	18.65	26.31	10^{-17348}	44.167	15.44	22.62	10^{-6919}

[c] σ_e is the site-year-set std; values are percents; [d]Chernoff upper bound on Chi-squared cum. dist.

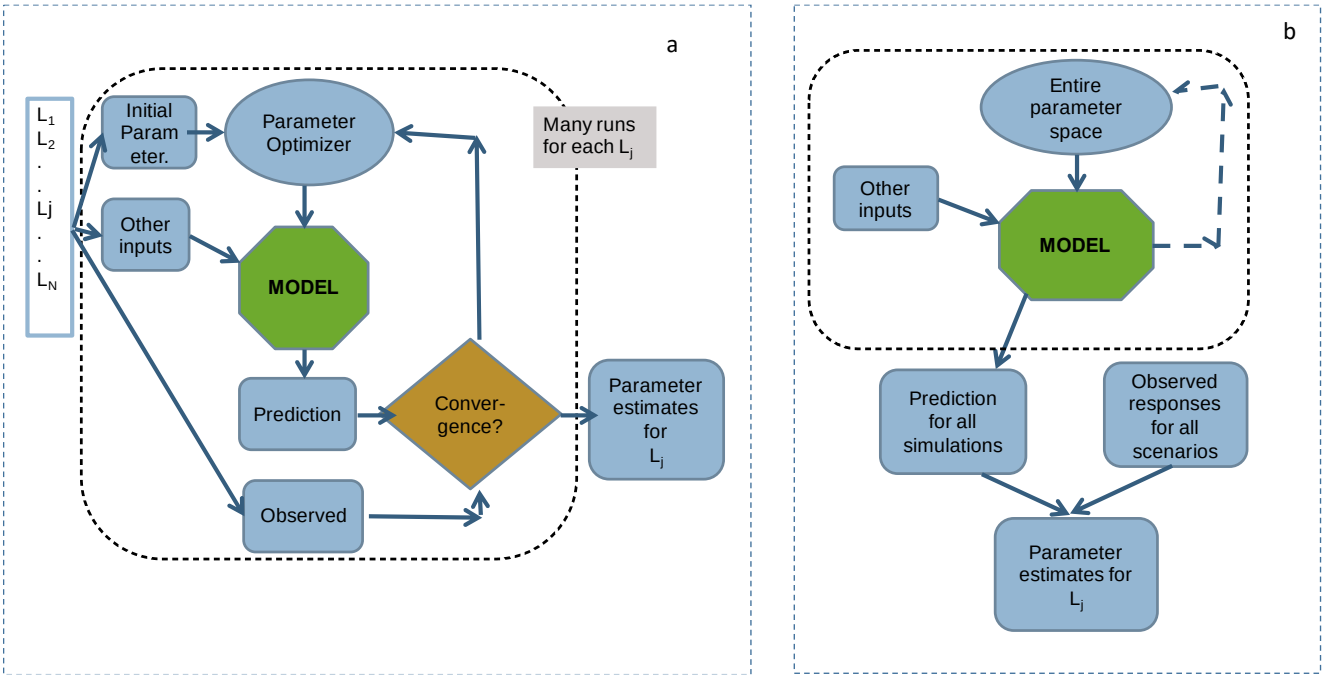

Fig. 4.1 Parameter search strategies a. Conventional method b. Database method. $L_{1...N}$ is the number of lines.

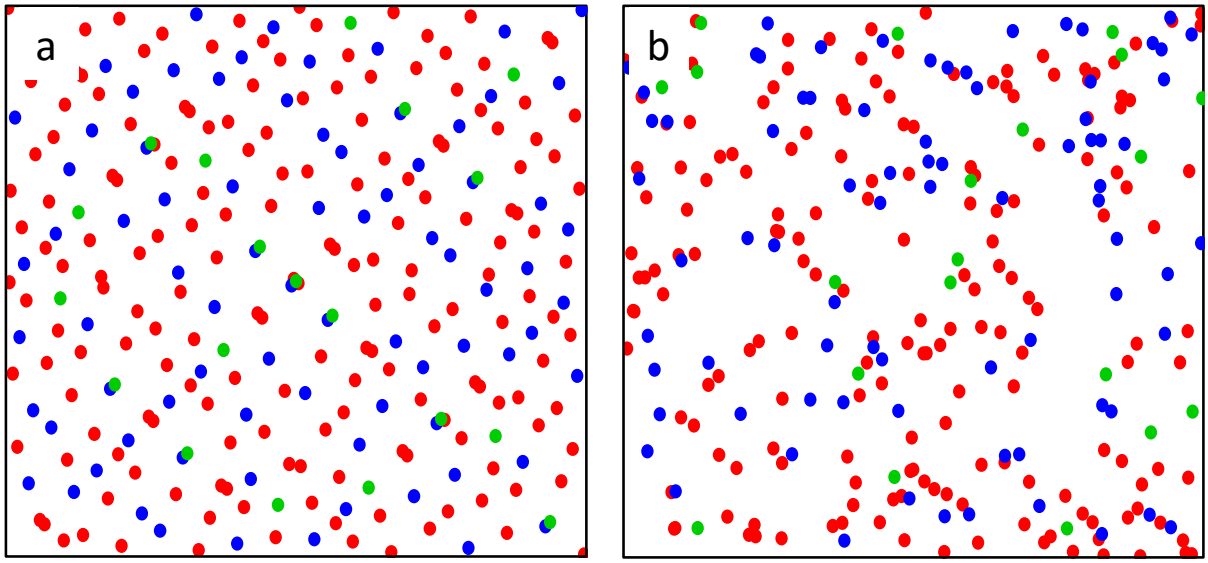

Fig. 4.2 (a) The first 275 quasi-random points from a two-dimensional Sobol sequence. (b) The first 275 points produced by the commonly used Mersenne twister pseudo-random number generator (Matsumoto and Nishimura, 1998). The Sobol sequence covers the space more evenly. The first 20 points are green, the next 80 are blue, and the final 175 are red, thus demonstrating Sobol gap filling.

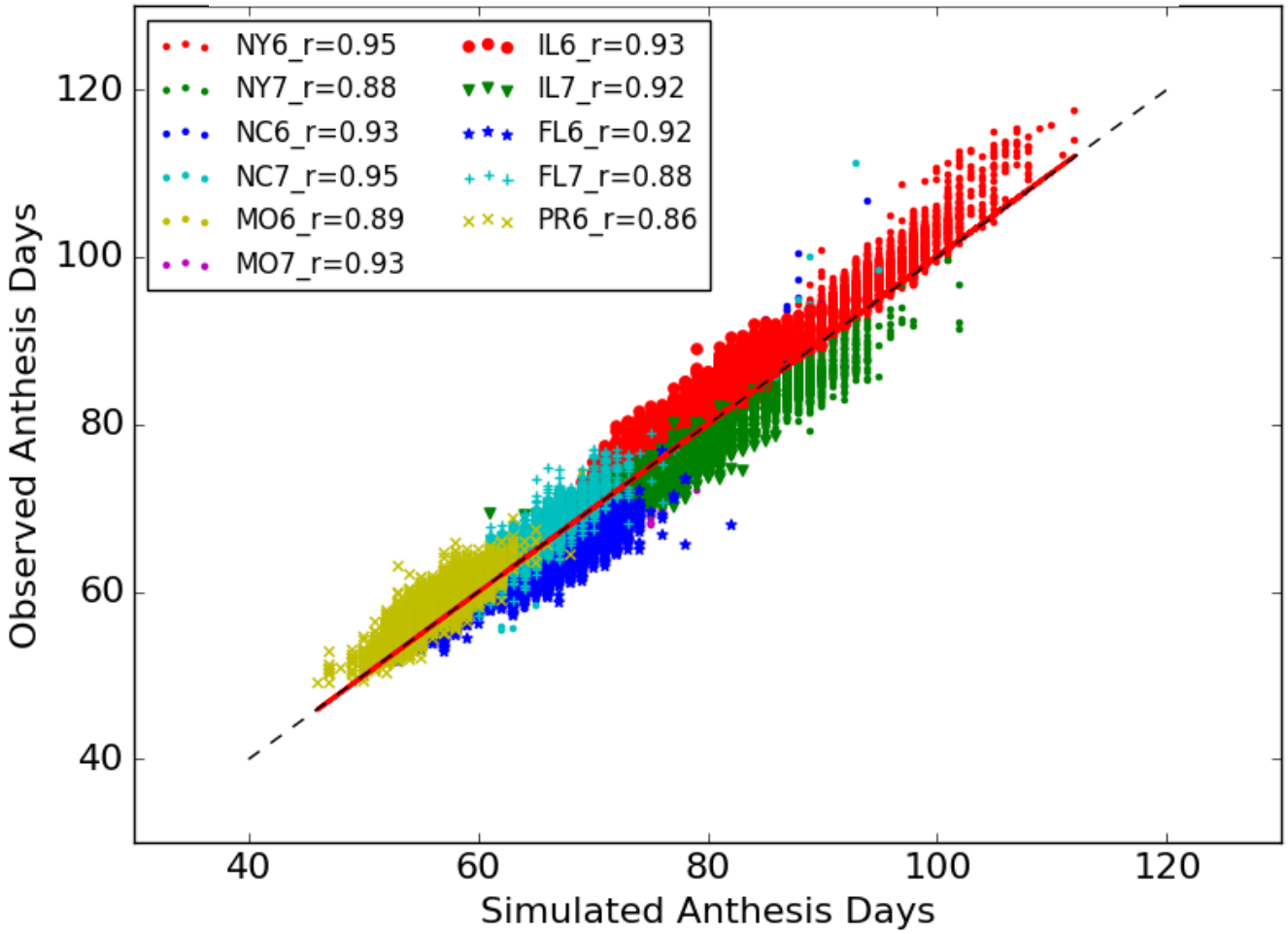

Fig. 4.3 Predicted and Observed anthesis days of all 5,266 lines from 11 site-year combinations. The graph has 49,491 points and an overall RMSE of 2.39 days.

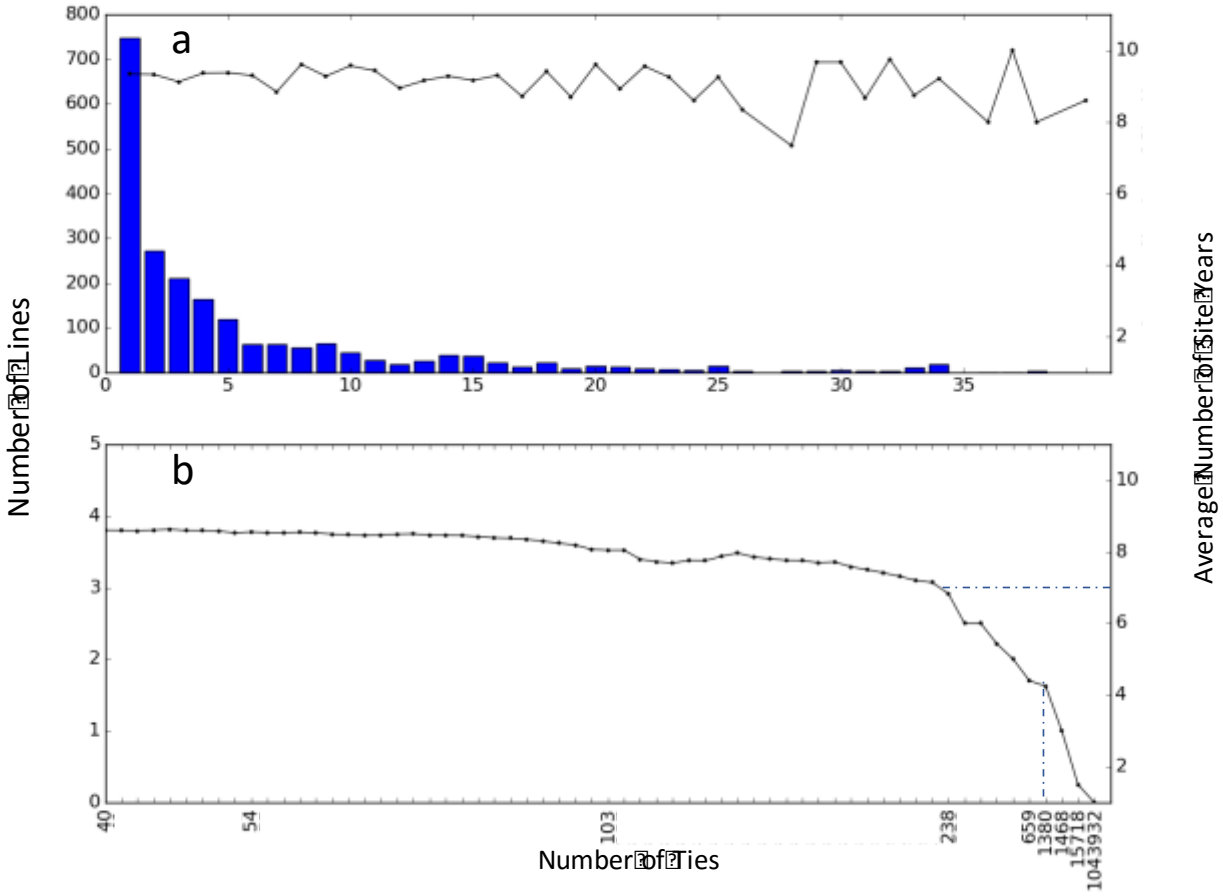

Fig. 4.4 Histogram depicting the frequency distribution of number of ties for 2,254 lines, used here to characterize equifinality. (a): Histogram of number of ties for 2153 lines with fewer than or equal to 40 ties. (b): Continuation of the histogram tail from figure a representing frequency of ties for the 101 lines with more than 40 ties. The trace at the top of each panel represents the average number of site-year combinations (right axis) used as data for parameter estimation.

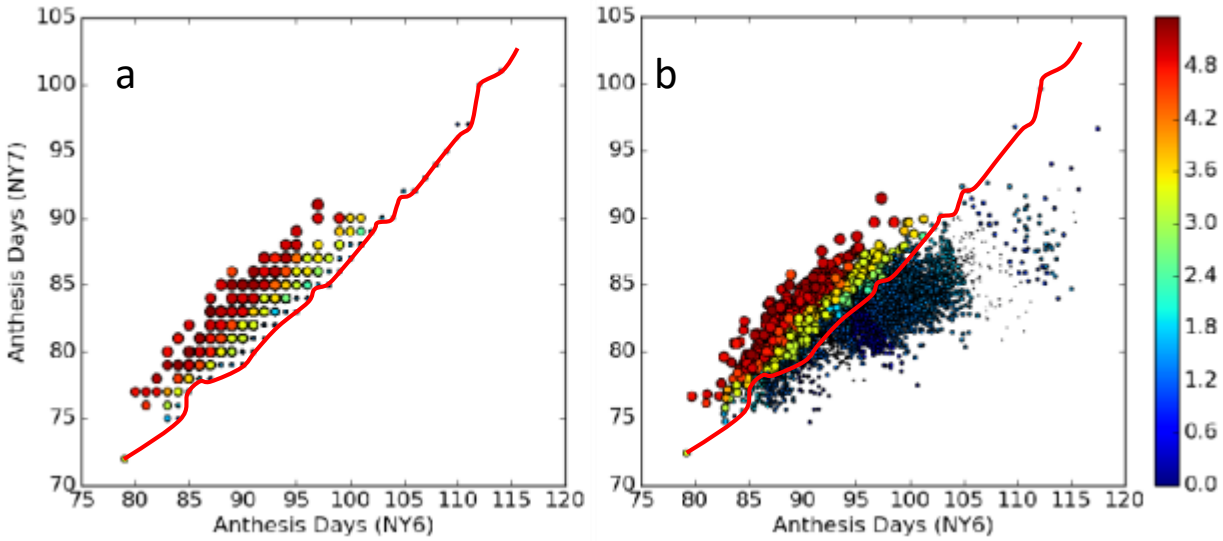

Fig. 4.5 Phenotype space plots of predicted (a) and observed (b) values of anthesis dates for site-years NY6 and NY7. The marker sizes and colors respectively express the levels of equifinality based on number of ties for P1 ($\log_{10}$ scale) and the relative ranges of its tied values. The red line is explained in the text.

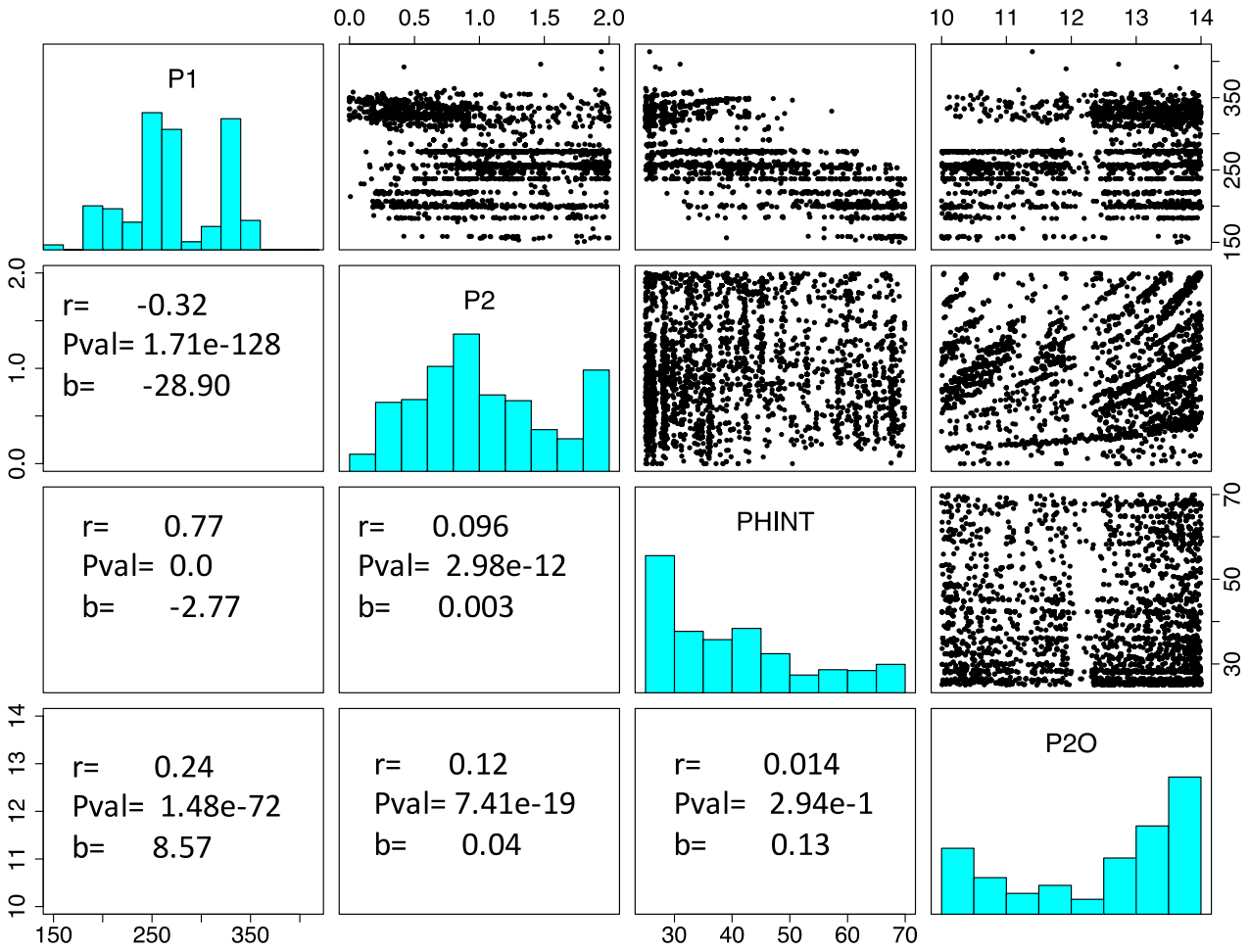

Fig. 4.6 Empirical distribution of selected GSP parameter estimates (main diagonal), pairwise scatterplots (upper right triangle) and empirical estimates of Pearson correlation coefficients, regression coefficients and p-values (Lower left triangle). Each dot in the scatter plots represents a pair of GSP estimates from a single line.

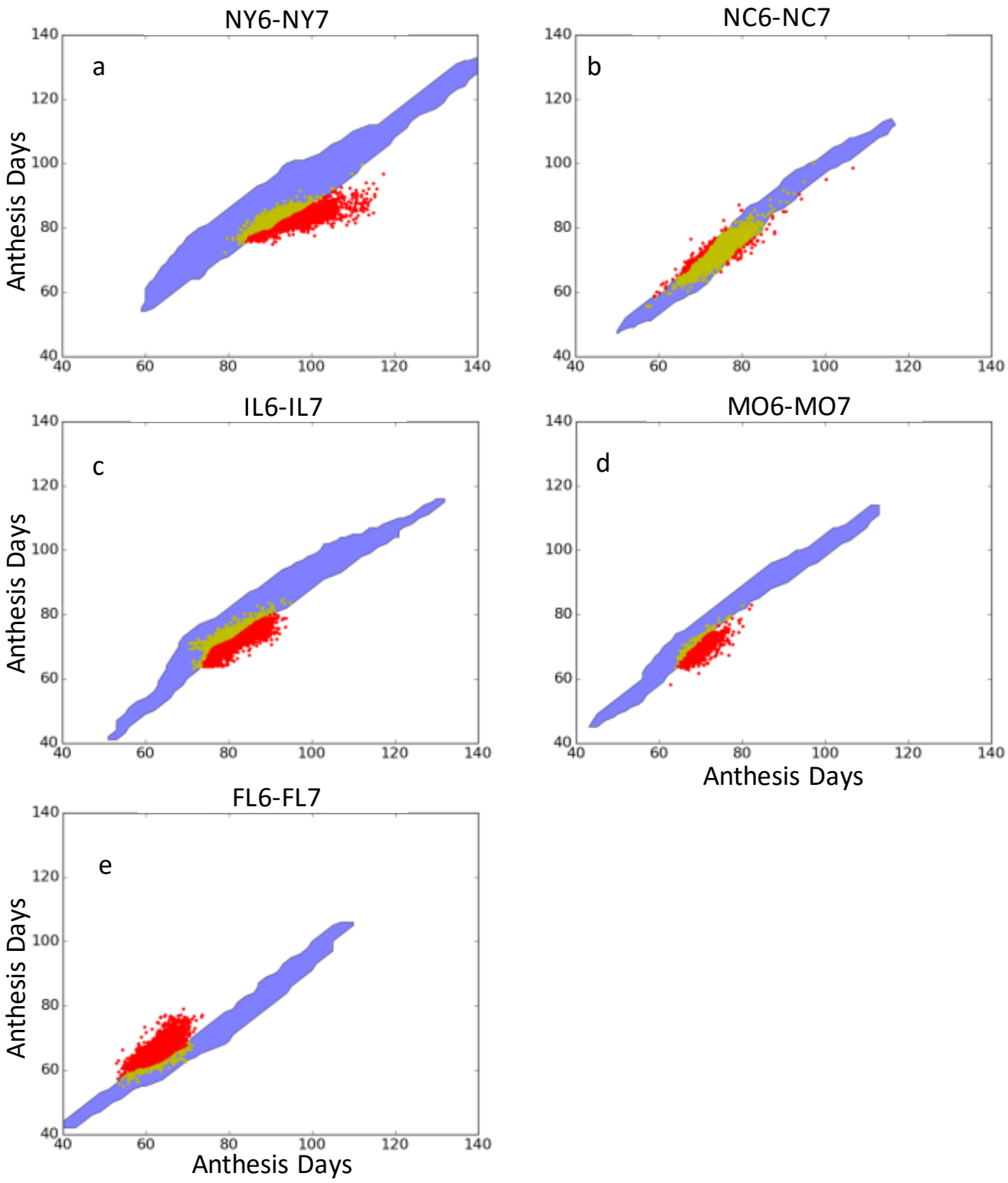

Fig. 4.7 Phenotype space plots for predicted and observed anthesis dates. Each panel corresponds to a pair of site-years for which fits were done. Regional color codes are described in the text.

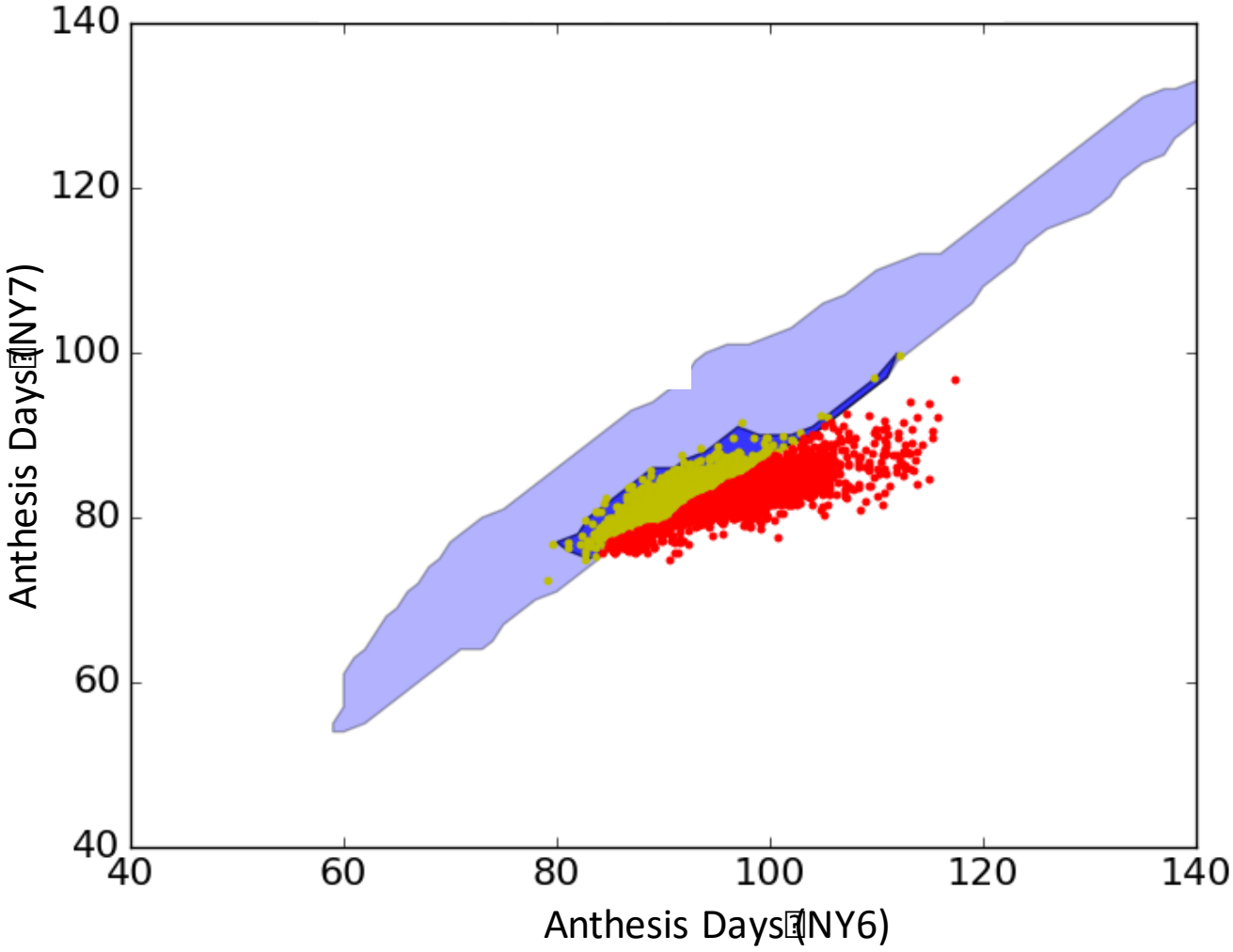

Fig. 4.8 Superimposed anthesis date results using NY6 and NY7 data illustrating that searches via database and DE optimization over a much larger parameter space are equally unable to reproduce the observations for lines shown as red dots.

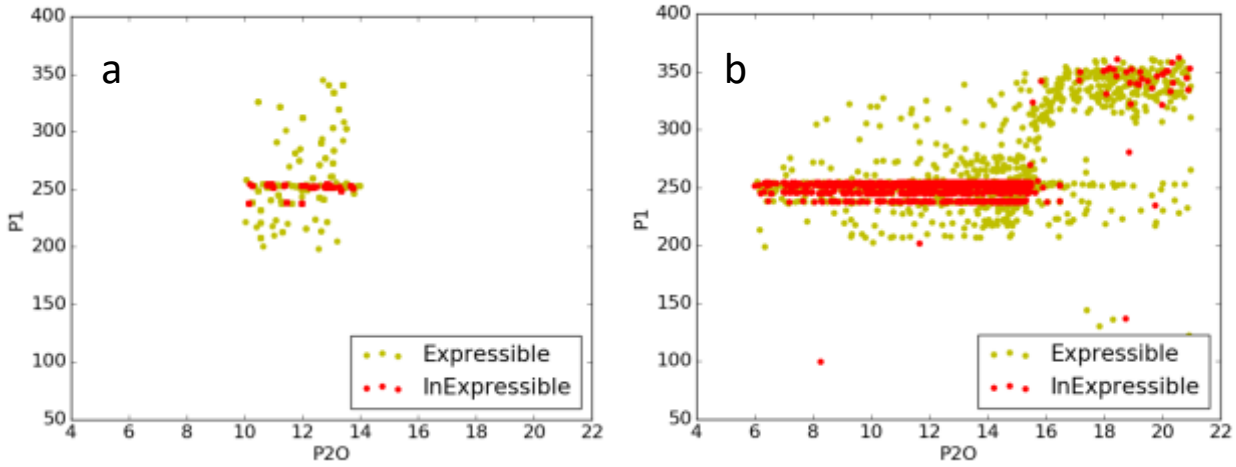

Fig. 4.9 Scatterplot of P1 vs. P2O estimates using data from NY6 and NY7 based on the database search (a) and Differential Evolution (b). Yellow and red dots are, respectively, observations characterized as expressible and inexpressible by model predictions.

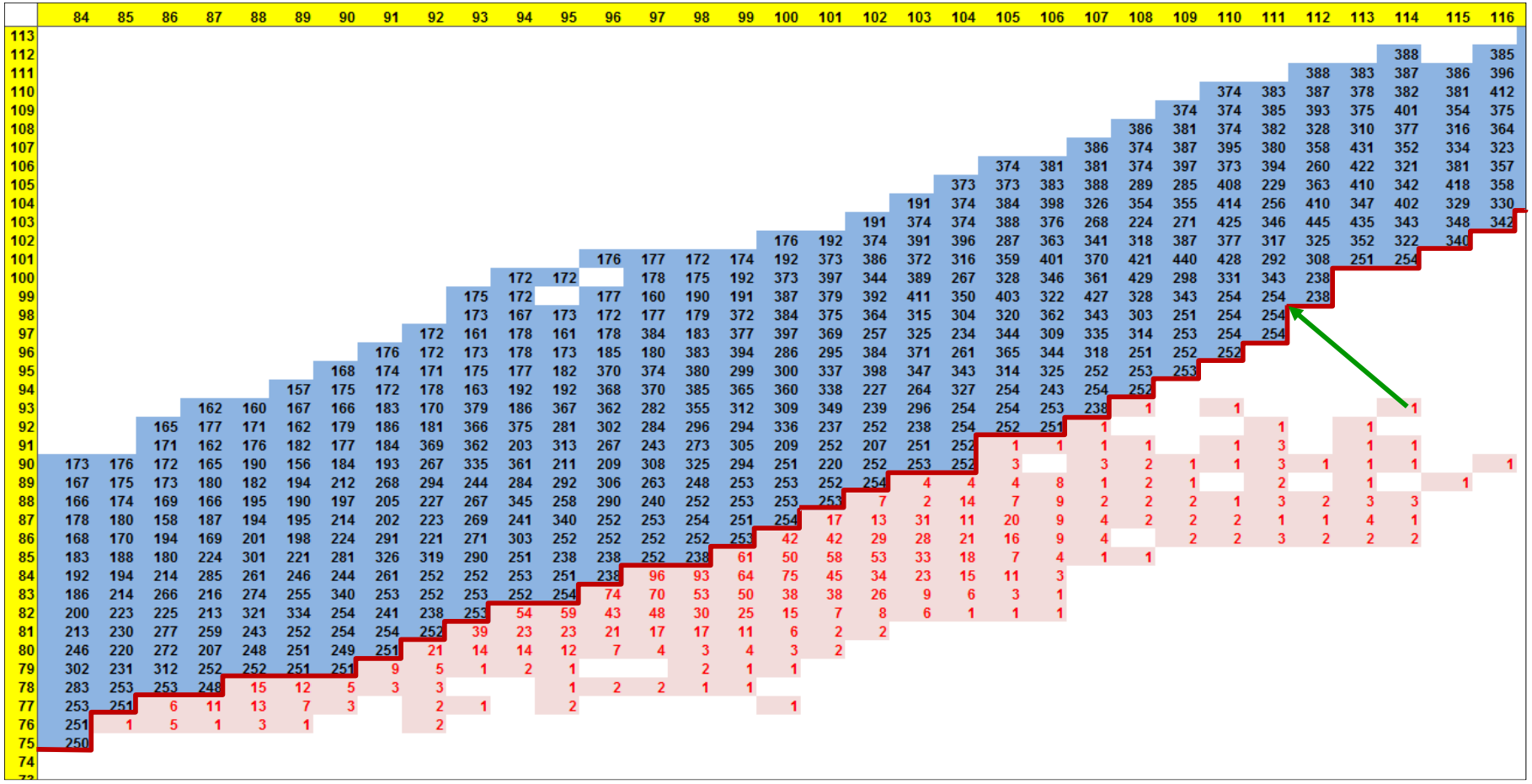

Fig. 4.10 P1 estimates from the database search (black) and the numbers of lines with inexpressible observations (red) arranged in a tableau organized as a phenotype space plot corresponding to the center portion of Fig. 8. The dark red line is the expressibility frontier and the green arrow shows the P1 value (254) from the GSP combination that minimizes the RMSE for one illustrative line. Horizontal and vertical yellow strips are the anthesis dates for NY6 and NY7

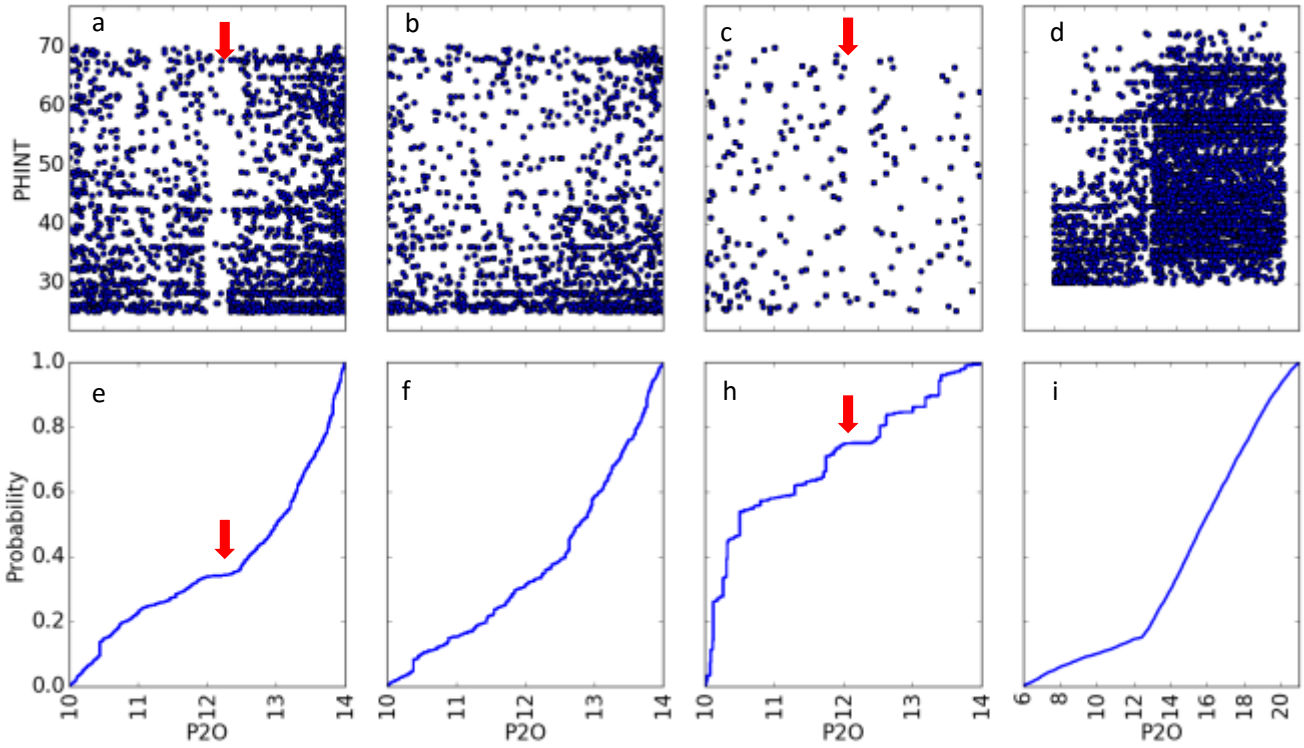

Fig. 4.11 P2O and PHINT scatter plots (top row) and P2O cumulative density functions (bottom row) using (a & e) all 11 site-years, (b & f) longer day site-years, (c & g) shorter day site-years based on the database approach, and (d & i) shorter day site-years using the DE approach. All horizontal axes in both rows have the same scale.

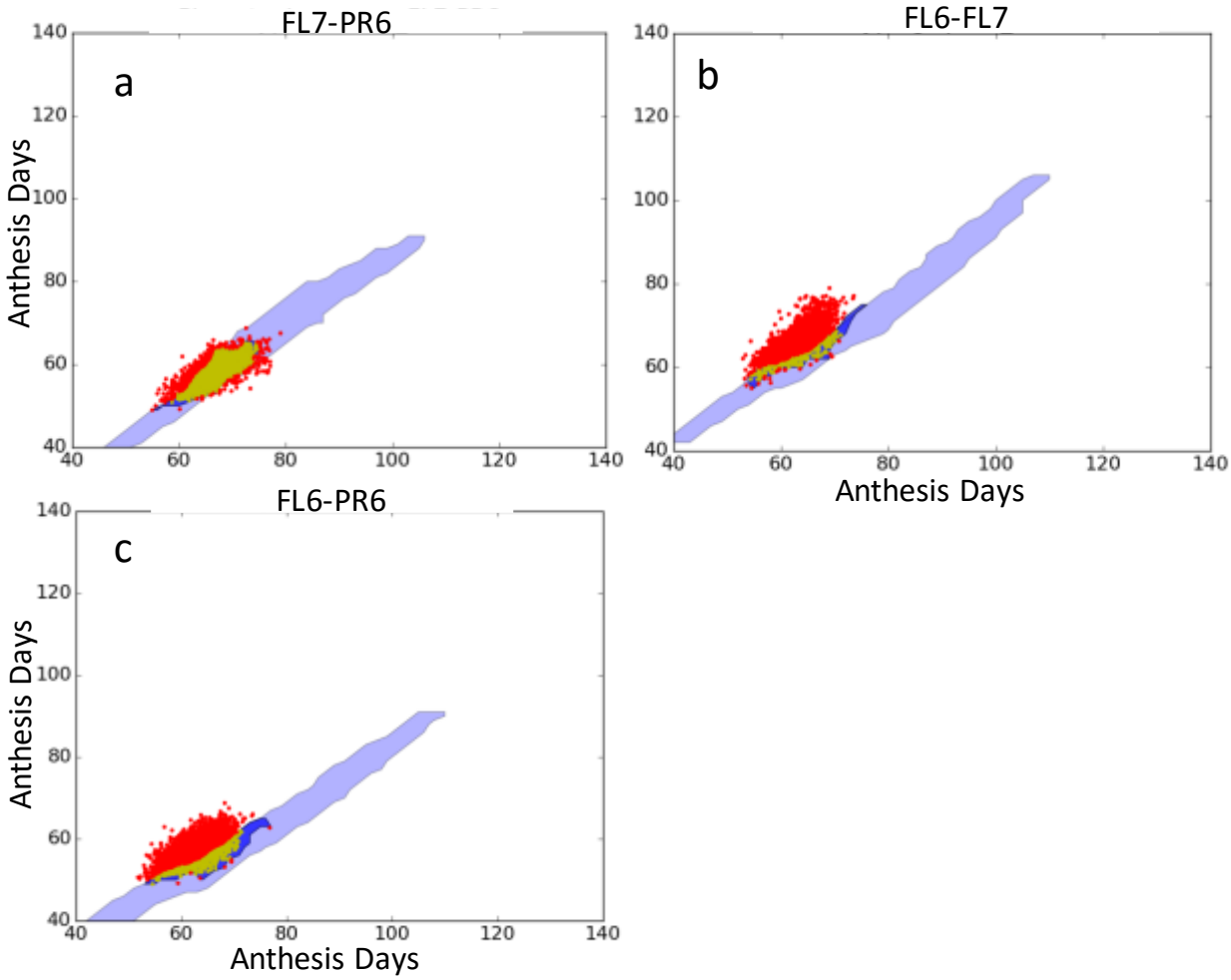

Fig. 4.12 Phenotype space plots of observed and predicted values based on the three site-years with shorter days. Note the large number of points in the FL6-PR6 and FL6-FL7 plots that lie above the dark blue prediction region based on DE.

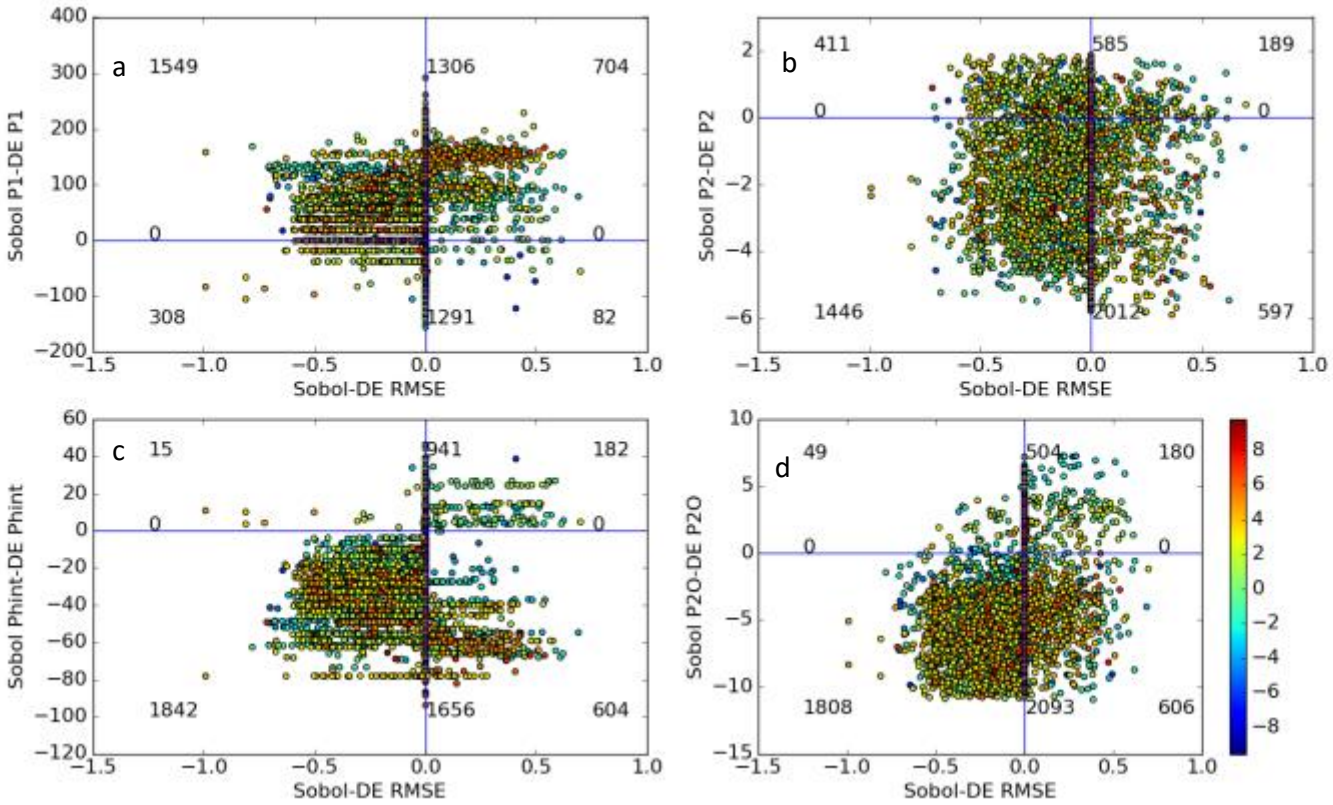

Fig. 4.13 The differences in parameter estimates from database search vs. DE (vertical axes) plotted against the corresponding difference in RMSE for 5240 lines in FL6, FL7, and/or PR6. The color encodes the sum of residual (observed minus mean) across site-years for each line.

CHAPTER 5 - CONCLUSION

Ecophysiological models have been used in agriculture for more than half a century. Since their inception, such models have been evaluated in terms of their predictive ability. As long as this proved to be useful, researcher were less concerned as to the biological accuracy of the parameters despite the fact that they were the *pro forma* drivers of the models. However, this priority is shifting with the advance of science and the need for these parameters to be correctly predicted from processes at the genetic level via statistical genetics models.

Having said that, due to use of very low amount of data, types of methods used to estimate the model parameters, and model evaluation in terms of overall prediction, the problems that we have outlined in previous chapters haven't been revealed before. It is, therefore, the modeling technologies (ranging from data collection to estimation) that were adequate for past applications that now require improvement.

The major problems encountered during this study can be categorized as 1) model expressibility, 2) parameter equifinality and 3) parameter instability across environment. These defects could not have been detected had we not used the database approach for estimation (Chapter 4) and global sensitivity analysis (Chapter 2).

Based on the results presented in this study, lack of expressibility is both a novel discovery and the most important issue as it entails an absolute inability of the model to closely predict the observations. A model with poor expressibility cannot be fixed only by acquiring large amount of data but only by developing a new model or changing the structure of the old one. Better models can be developed with the help of molecular gene studies and high throughput phenotyping. The former are beginning to look at connecting multiple gene systems together that control alternative

physiological systems. While not yet including the range of processes contained in full ECM's, it could be quite useful to evaluate the expressivity of such models using the techniques presented here.

Parameter equifinality is the second most important problem. Equifinality was not a big issue in the past, especially for researchers who just wished to use crop models for prediction purposes. However, current applications seek to link crop models to genetics to begin designing ideotypes and to guide the selection process in crop breeding programs, both of which will help to accelerate annual rates of gain. This cannot be achieved as long as crop models remain afflicted with issues of expressibility, equifinality, and GSP instability.

There are two approaches to equifinality that attack the problem from different angles. The first way is to try change the model structure to make it simpler – specifically, to have fewer parameters. In this way, the interactions between parameters that allow equifinality to occur can be eliminated. Additionally, parameters that have very limited impact on model outputs can be set to nominal values, thus excluding them from the estimation process. Global sensitivity analysis proved to be a possible approach to both of these ideas – identifying interacting parameters and/or ones of limited influence. The second way to address equifinality would be to collect data sets involving both large amounts and many types of data. There are many developmental trajectories, each with its own set of parameter values, that can result in the same anthesis date. However, the more measures one has of other ancillary traits, the greater one's ability to winnow through and discard those trajectories until just one remains. For example, total number of leaves is needed to simulate anthesis days in CERES-Maize. When not measured, leaf number is estimated as [P1/(PHINT*2) +5]. If, however, leaf number is available, fewer parameter combinations can agree with all the data. Recent advances in high throughput phenotyping approaches can be very

helpful in both widening the range of data types that can be obtained along with enlarging the total amounts.

Parameter instability is another issue that limits the use of crop model in multiple environments. Instability was detected in all three studies included in this dissertation. Parameter instability is important because if the estimates vary depending on the environments used to estimate them, it is impossible to know what values to use in an ECM under soils and weather different from those present in the original field studies. For example, the cultivar parameters (e.g. P1, P2, PHINT, P2O), explicitly relate to plant behaviors (e.g., development toward anthesis) to environmental variables (e.g., temperature and photoperiod in the current case). From the time the models were first created, it has been assumed that such parameters are properties of the individual lines (i.e., stable). The corollary to this assumption is that, by implication, the parameters have a genetic basis because genotypes do not change with the environment. Thus, the expectation is that research will be able to mechanistically link at least some GSP's to molecular genetic processes. But this cannot be done for unstable parameters.

It may be that part of the instability arises because of discrepancies in measurement errors between different plantings. Usually weather data are measured at single points either within; adjacent to; or, sometimes, at some distance from the experimental sites. Soil variables, to which plants are highly sensitive (chapter 2) may be quantified at multiple points within a field but seldom at a high level of horizontal or vertical resolution. Thus, there will always be a distribution of errors between the measured environments and those actually present at the plants. Furthermore, these distributions will vary from one planting to the next and those differences will be modulated by the highly nonlinear nature of the models into GSP estimates lacking in stability. Given the amount of data in used in chapter 4 the p-values in Tables 5 are, perhaps, not surprising. However, the

sizes of the Variability Indices and variance percentages severely compromise the abilities to link these GSP's to genetics.

The two remedies mentioned above, better models and more data, might also be, to a greater degree or less, helpful with parameter instability. However, a second strategy might also contribute – moving away from point estimates of environmental factors to ones that are taken directly on an area basis. One example might be using leaf temperature as a direct external input to the model in place of air temperature. This would overcome the considerable field-to-field divergence in weather station placement. Using leaf temperature in this way would have been impractical to the point of inconceivable in the early days of crop modeling but sensing capabilities have greatly advanced since then. In particular, high throughput phenotyping automated ground or airborne vehicles with thermal cameras make such measurements quite straight forward today. Perhaps, models are in need of revision not only to catch up with biological knowledge but to better align with modern instrument technology as well.

Whatever the steps will be taken in the future, it is obvious that the amounts of available data are going to explode. Thus, model simulations can be expected to generate a high computational demand. Using supercomputers (e.g. Stampede at TACC) will be a virtual necessity to accomplish the large number of simulations that some tasks such as parameter estimation will require. Even so, part of the solution will have to be better optimization methods to estimate model parameters in feasible times despite the large amounts of data. The Holographic Genetic Algorithm that we developed in chapter 3 is a good foundation on which to build. It was highly efficient at estimating large number of parameter using large number of cultivar-site-years data.

Finally, one cannot fix the problems that one does not know exist. Over the last 20 years, crop models have been considered as important tools that can help accelerate breeding programs

via improved phenotype prediction of prospective crop genotypes in novel, time-varying environments and sophisticated new management practices. However, based on the results presented here, it is very uncertain that crop models having current architectures and driven by current sensor systems can achieve this goal. Thus, future research should be directed towards solving these issues. Indeed, it is imperative these problems be fixed so that crop models can be used as tools to help breeders, farmers, and researchers address global food security issues. The tools, statistical tests to detect problems and monitor progress, and the algorithms developed herein are a foundation on which to build.