World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:11, No:7, 2017

Model-Driven and Data-Driven Approaches for Crop Yield Prediction: Analysis and Comparison

Xiangtuo Chen, Paul-Henry Cournède,

***Abstract*—**Crop yield prediction is a paramount issue in agriculture. The main idea of this paper is to find out efficient way to predict the yield of corn based meteorological records. The prediction models used in this paper can be classified into model-driven approaches and data-driven approaches, according to the different modeling methodologies.

The model-driven approaches are based on crop mechanistic modeling. They describe crop growth in interaction with their environment as dynamical systems. But the calibration process of the dynamic system comes up with much difficulty, because it turns out to be a multidimensional non-convex optimization problem. An original contribution of this paper is to propose a statistical methodology, Multi-Scenarios Parameters Estimation (MSPE), for the parametrization of potentially complex mechanistic models from a new type of datasets (climatic data, final yield in many situations). It is tested with CORNFLO, a crop model for maize growth.

On the other hand, the data-driven approach for yield prediction is free of the complex biophysical process. But it has some strict requirements about the dataset.

A second contribution of the paper is the comparison of these model-driven methods with classical data-driven methods. For this purpose, we consider two classes of regression methods, methods derived from linear regression (Ridge and Lasso Regression, Principal Components Regression or Partial Least Squares Regression) and machine learning methods (Random Forest, k-Nearest Neighbor, Artificial Neural Network and SVM regression).

The dataset consists of 720 records of corn yield at county scale provided by the United States Department of Agriculture (USDA) and the associated climatic data. A 5-folds cross-validation process and two accuracy metrics: root mean square error of prediction(RMSEP), mean absolute error of prediction(MAEP) were used to evaluate the crop prediction capacity.

The results show that among the data-driven approaches, Random Forest is the most robust and generally achieves the best prediction error (MAEP 4.27%). It also outperforms our model-driven approach (MAEP 6.11%). However, the method to calibrate the mechanistic model from dataset easy to access offers several side-perspectives. The mechanistic model can potentially help to underline the stresses suffered by the crop or to identify the biological parameters of interest for breeding purposes. For this reason, an interesting perspective is to combine these two types of approaches.

***Keywords*—**Crop yield prediction, crop model, sensitivity analysis, paramater estimation, particle swarm optimization, random forest.

I. Introduction

IN agriculture research, crop yield prediction is a major topic of interest for farmers, decision makers and agricultural organizations. It is made very difficult by the variety of agricultural systems, the diversity of biophysical processes implied in plant growth, the complexity of crop responses to stress, etc. What's more, farmers decisions, such as land preparation, irrigation, sowing date or fertilizer applications, have also a great influence on crop yield. Normally, this prediction is carried out according to the farmers' long-term experience for specific fields, crops and climate conditions [1]. Nevertheless, a non-linear behavior of plant reaction to the environment introduces large deviations from year to year and makes the traditional method inaccurate [2]. Thus, more efficient methods have been developed, which can be generally classified as crop growth models and data-driven models.

Agronomic models, just as CORNFLO [3], are generally based on the mechanistic description of biophysical processes. In most cases, they are considered as a discrete dynamic system. That is to say, they can be represented in the following form:

$$X_{t+1} = F_t(X_t, U_t, \theta) \tag{1}$$

where X_t and U_t represent the state variables and environmental variables of the system at time t, θ is the parameter vector, and F_t denotes the eco-physiological processes involved in the model. It is assumed here that the process is deterministic, but it is also possible to consider modeling noises [4]. This kind of dynamic system has opened interesting perspectives for a better understanding of plant growth as well as for potential applications in breeding or decision aid in farm management. But the parameterization of such models is however a difficult issue due to the complexity of the involved biological processes and the interactions between these processes [5].

A typical parameterization method for discrete dynamic models is described in [6], and an application in the case of mechanistic plant growth model is given in [5]. Let $(t_k)_{1\leq k\leq n}$ denote the sequence of times at which the plant was observed, and y_k the observation vector at time t_k. The vector of observations is thus implicitly a function of the vector of parameters θ:

$$y = f(\theta) + \epsilon \tag{2}$$

with $\epsilon \sim \mathcal{N}(0, \Sigma)$ and f represents the model used. Then, generalized least squares estimator or maximum likelihood estimator can be implemented.

However, this methodology has several disadvantages: firstly, the experiment is so expensive in terms of time and money that the sample size is usually quite small; secondly, the experiments are conducted in the same environment, which makes the genericness of the calibrated model questionable.

X. Chen is with the Laboratory of Mathematics and Computer Science (MICS), CentraleSupélec, Gif sur Yvette, 91190, France (e-mail: xiangtuo.chen@ecp.fr).

P. H. Cournède is with the Laboratory of Mathematics and Computer Science (MICS), CentraleSupélec, Gif sur Yvette, 91190, France.

In this work, another type of plant observation data is used to calibrate crop models: $\{U_i, y_i\}_{1\leq i\leq N}$, where U_i is the records of the environmental variable in scenario i, while y_i is the yield (the final state) of plant in scenario i. The lost information of the plant at different stages is hedged by the diversity of crop performance in different scenarios. This methodology is named as multi-scenarios parameter estimation and the original idea can be found in [3]. After parameterization, the calibrated model will be compared with the data-driven methods for its predictive capacity.

As for data-driven approaches, they refer to statistical learning methods using historical data to calibrate a non-mechanistic prediction system, that is to say a model not based on specific domain knowledge. In this work, since the available data-set is in the form $\{U_i, y_i\}$, the data-driven approaches will be used to build a regression model of the form:

$$y = g(U) + \epsilon \tag{3}$$

where $y = (y_1, y_2, \ldots, y_n)^t \in \mathbb{R}^n$, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, y_n)^t \in \mathbb{R}^n$ and g is a "black box" trained by available data and the specific methods called in modeling process. Solutions can be divided into two parts: the statistical methods like Ridge and Lasso regressions, principal component regression and partial least squares regression; the machine learning methods like regression trees, random forest, k-nearest neighbors (KNN), Artificial neural network (ANN), SVM regression, etc. In this work, a systematic analysis will be conducted to compare the predictive capacity of these data-driven methods in terms of crop yield prediction.

These methodologies are tested on a database provided by the statistical service of the United States Department of Agriculture (USDA), on corn yield at county scale. The data set is presented in Section II. Then crop model analysis, the MSPE methodology and its predictive capacity evaluation will be performed in Section III. In Section IV, some typical statistical methods and machine learning methods are briefly recalled and evaluated. Finally, a discussion of these two methodologies is detailed in Section V.

II. Data Description and Criteria

A. Data Description

The data used in this work is obtained for one specific corn genotype over 10 years from 2001 to 2010 in diverse counties of the USA. The data for a given site of a given year (site-year data) is called a scenario. There are 720 scenarios available in this work and each scenario is composed by a set of climate data and a final crop yield value (average yield at country scale).

1) Climate Data: Each scenario is composed by daily records of five important climate variables: daily maximum and minimum temperatures, radiation, precipitation and potential evapo-transpiration. Thus, each climate data set contains 365 measurements and each scenario instance is represented by a set of 1825 (365 * 5) numeric value. $X_i = (X_{i1}, X_{i2}, \ldots, X_{id})$ is a vector of feature measurement for the ith scenario's climate data.

2) Crop Yield Data: The crop yield of the ith scenario, noted y_i, is a single numeric value representing the total mass of maize seed collected on the harvest day (in g/m^2).

B. Criteria

1) Mean Squared Error Prediction (MSEP): The mean squared error of prediction, or MSEP, is a standard criterion for assessing the predictive capabilities of a model in ecological and agronomic studies [7] . It measures the difference between the observations y and the predictions of the model $f(\theta)$, and is defined as follows:

$$\mathrm{MSEP}(\hat{\theta}) = \mathbb{E}_{\hat{\theta}}[(y - f(\hat{\theta}))^2] \tag{4}$$

Since the dimension of MSEP is the squared dimension of the observation, it is more convenient to use the root of the quadratic error:

$$\mathrm{RMSEP}(\hat{\theta}) = \sqrt{\mathrm{MSEP}(\hat{\theta})} \tag{5}$$

If another independent data-set is available, that is different from the one on which the parameter estimation was conducted, an unbiased estimator of RMSEP is given by [8]:

$$\hat{\mathrm{RMSEP}}(\hat{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

where y_i and $\hat{y}_i$ are the observed and predicted crop yields in scenario i at harvest time.

2) Mean Absolute Percentage Error (MAPE): The mean absolute percentage error, or MAPE, is another measure of prediction accuracy for a forecasting method in statistics [9]. It usually expresses accuracy as a percentage, and is defined by the following formula:

$$\mathrm{MAPE}(\hat{\theta}) = \frac{100}{n}\sum_{i=1}^{n} \mid \frac{y_i - \hat{y}_i}{y_i} \mid \tag{7}$$

where y_i and $\hat{y}_i$ are the observed and predicted crop yield in scenario i at harvest.

III. Plant Model Analysis and Evaluation

A. CORNFLO: A Plant Model of Maize

The CORNFLO model is a plant growth model that simulates the growth and yield of maize [3]. It is inspired from the SUNFLO model for sunflower [10]. A detailed model description can be found in Appendix A.

B. Model Calibration with MSPE

Model calibration is based on the MSPE methodology. The main steps are the following ones:

- First, we take advantage of the prior knowledge on the parameters to derive prior distributions and perform a global sensitivity analysis of the model parameters to screen the most important ones that will be estimated in priority;
- Then, we implement an efficient nonconvex optimization method, the parallel particle swarm optimization, to

TABLE I
GENOTYPE PARAMETERS IN CORNFLO MODEL

Parameter	Unit	Meaning
$A2$		Rank of the leaf which has the largest leaf surface in the entire plant growth period
$A3$	cm^2	Potential surface of the largest leaf of the simulated plant
$phyllodeini$	°C·days	Phyllochrone parameters for plant leaves
$Ratio\_phyllodefe$		Ratio of phyllode and phyllofe
$F1$	°C·days	Thermal time needed the beginning of flowering
$M3$	°C·days	Thermal time needed for the physiological maturity
k_coeff		Leaf extinction coefficient
$phyllofeini$	°C·days	Phyllochrone parameters for plant leaves
NFF		Number of leaves
HI		Harvest Index: a constant biomass proportion of the wet grain
RUE		Maximum radiation use efficiency
$M0$	°C·days	Thermal time needed for the early maturation

search for the maximum of the distribution of the estimated parameters;
- Finally, we choose the best configuration regarding the number of estimated parameters by model selection criteria.

1) Reduction of the Variability: Model calibration is a critical issue for models with a large number of parameters and restricted data [11]. So the first task is to select the most important parameters according to some criteria.

The sensitivity analysis (SA) technique makes it possible to evaluate the sensitivity of the response variable to the disturbances of model inputs [12]. This kind of analysis makes it possible to identify the parameters having the most influence on the results of the model, specifically chosen to be the experimental variables (here the final yield).

Sobol's method is a popular method for sensitivity analysis based on the variance decomposition of the model's output, which allows a clear interpretation of the SA results. In the framework of plant growth models, [14] proposes an improvement of the Homma-Saltelli method for the calculation of the Sobol sensitivity indices. Parameters are ranked according to their global SI, and then, the models will be calibrated with the important parameters selected by SA.

As stated in Appendix A, the CORNFLO model has 12 genotype parameters. Their definitions are listed in Table I. A summary of the variation intervals adopted for each parameter and their recommended value settings can be found in Table II. The fact that the model describes mechanistic processes of crop growth allows the determination of reasonable ranges of variations for the parameters and thus prior distributions for SA.

TABLE II
VARIATION INTERVALS AND RECOMMENDED VALUES FOR MODEL PARAMETERS

Parameter	Interval	Recommended Value
$A2$	$[7, 19]$	14.07
$A3$	$[400, 720]$	645
$phyllodeini$	$[22, 42]$	32
$Ratio\_phyllodefe$	$[0.5, 0.9]$	0.7
$F1$	$[410, 890]$	723
$M3$	$[950, 1750]$	1477
k_coeff	$[0.4, 0.75]$	0.53
$phyllofeini$	$[30, 50]$	40
NFF	$[12, 26]$	21
HI	$[0.3, 0.8]$	0.5
RUE	$[0.5, 0.9]$	3.5
$M0$	$[550, 1060]$	884

The Sobol sensitivity analysis is carried out, the ordered parameters and their SI can be found in Table III. The combination: $A2$, NFF, RUE, $M3$, $F1$ and $M0$, is considered to be the "important parameters" to be calibrated first because the sum of their first order SI $> 90\%$, which means that the variability of these parameters alone contributes to more than 90% of the output's uncertainty. But the parameter, NFF, which represents the number of potential leafs, can be estimated accurately with direct experimental measurements. Finally the chosen parameter vector contains five elements: $A2$, RUE, $M3$, $F1$ and $M0$.

2) Least-Squares Minimization: As underlined above, model calibration with MSPE methodology does not use the commonly used data frame $(y_{t1}, y_{t2}, \ldots, y_{ti})_{t1 \leq ti \leq tN}$ for plant growth models as in [10], which contains several observation at different moments but in a single scenario (environment). Instead, another data frame $(U_i, y_i)_{1 \leq i \leq N}$, which contains the observation of the final state but in multiple scenarios (environments) is adopted.

For the least squares criterion, the estimation of the parameter vector $\hat{\theta}$ can be classically expressed as follows:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - f(\theta, U_i))^2. \tag{8}$$

where θ is the parameter vector containing the five elements selected beforehand, $\theta = (A2, RUE, M_3, F_1, M_0)$, U_i contains the environmental information of scenario i, y_i is the final observation of the crop yield in scenario i.

Note that since the prior on the parameters, finding the minimum of the least-squares criterion within the given interval, correspond to finding the maximum a posteriori in a Bayesian context, if errors in the different scenarios are supposed homoscedastic.

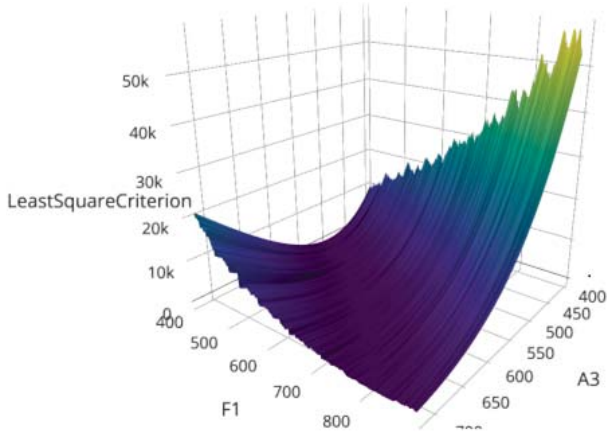
a) Nonconvex optimization: As illustrated in (8), the parameter estimation process is an optimization problem. However, a strong characteristic induced by the non-linearity of the model is the nonconvexity of the function to optimize. For illustration, two 3-D optimization surfaces of the objective function are drawn in Fig.1.

As a result, the particle swarm optimization, which is a global optimization method, will be applied to resolve the optimization problem.

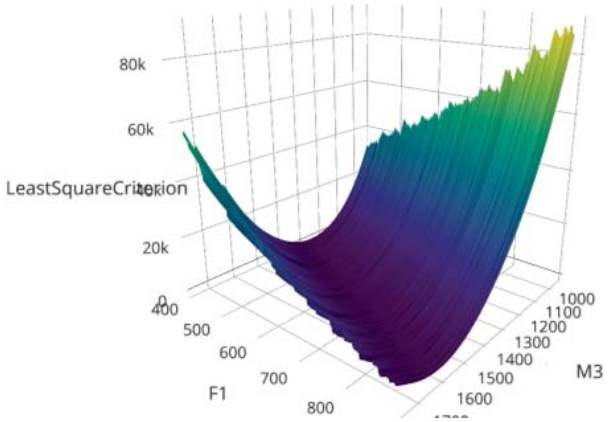
b) Particle swarm optimization method: The particle swarm optimization (PSO) is a global stochastic optimization

TABLE III
FIRST ORDER AND TOTAL SOBOL SENSITIVITY INDEX OF CONRNFLO MODEL

Index	$A2$	NFF	RUE	$M3$	$F1$	$M0$	$A3$	$\cdots$
S_1	0.256	0.213	0.173	0.161	0.075	0.051	0.011	$\cdots$
ST	0.293	0.247	0.224	0.173	0.081	0.059	0.015	$\cdots$



(a) objective function surface with axes F1 and A3



(b) objective function surface with axes F1 and M3

Fig. 1 (a), (b) show the 3D objective function surface with different axes in the plant model calibration process

method proposed by Eberhart and Kennedy in 1995 [16]. It is based on the concepts of individual improvement, population cooperation and competition of social animals, such as bird flocking or fish schooling. In this algorithm, each individual of the swarm, called particle, keeps track of the best solution along its own discovery and that by the whole swarm. Then, the particles move in the search space according to Equ.9 and Equ.10:

$$\overrightarrow{v_i}^t = \overrightarrow{v_i}^{t-1} + c_1 \times r_1^t(\overrightarrow{p_i} - \overrightarrow{x_i}^{t-1}) + c_2 \times r_2^t(\overrightarrow{p_g} - \overrightarrow{x_i}^{t-1}) \quad (9)$$

$$\overrightarrow{x_i}^t = \overrightarrow{x_i}^{t-1} + \overrightarrow{v_i}^t \quad (10)$$

where i represents particle index, t denotes the iteration number, $\overrightarrow{v}$ is the velocity of the particle, while $\overrightarrow{x}$ represents the particle's position. The coefficients c_1, c_2 are search parameters that reflect the different influences of different resources on velocity. r_1 and r_2 are two random (time-dependent) variables with uniform distribution $\mathcal{U}[0,1]$. $\overrightarrow{p_i}$ and $\overrightarrow{p_g}$ denote respectively the best solution found by particle i and by the whole group.

However, as stated in [17] and [18], the basic algorithm faces some disadvantages: firstly, it is easy to fall into local optima; secondly, there will be a risk of particles' explosion; thirdly, it requires powerful computational capacity. In order to overcome these disadvantages, many different improvements have been proposed: [19] shows the importance of the number of particles, but the ideal number depends on the past experience; [20] introduces the notion of "neighbor topology" and improves significantly its global search ability; [21], [22] and [23] put forward the notions of "confidence coefficient", "maximum speed", "constriction factor" and "inertia factor" to limit the particles in the confined searching space; [24] comes up with a synchronized parallel PSO version and significantly improves computational efficiency.

TABLE IV
ESTIMATION RESULT WITH ALL 720 RECORDS

	$A2$	$F1$	$M0$	$M3$	RUE
recommended value	14.07	723	884	1477	3.5
estimated value	13.95	707.35	869.45	1453.70	3.33

In this work, the MSPE methodology is based on the local version with Von Neumann neighbour topology. The velocity upload equation is updated with the following equation:

$$\overrightarrow{v_i}^t = \omega \times \overrightarrow{v_i}^{t-1} + c_1 \times r_1^t(\overrightarrow{p_i} - \overrightarrow{x_i}^{t-1}) + c_2 \times r_2^t(\overrightarrow{p_k} - \overrightarrow{x_i}^{t-1}) \quad (11)$$

where the inertia factor $\omega = 0.7298$, the confidence coefficients $c_1 = c_2 = 1.948$, and $\overrightarrow{p_k}$ represents the best position (solution) of the kth subgroup. And the number of particles is set to be 1200 after several tests. The Improved algorithm as well as all the statistical methods are implemented in the PYGMALION modeling platform [15]

C. Results

1) Calibration Results: In a nonconvex optimization problem, there is no algorithm that can ensure the convergence to the global optimum, neither does the PSO [25]. In order to find out the optimal setting of the parameter vector $\hat{\theta}$, the optimization process of all the 720 records is repeated 100 times to check the stability of the computed optimal value and also tune the number of iterations for the algorithm. The estimated value in Table IV is considered to be the best setting of parameter vector θ.

2) Influence of the Number of Scenarios: It is interesting to study the influence of the number of scenarios to know in real-case situations the number of data that will be necessary estimate properly model parameters. For this purpose, the 720

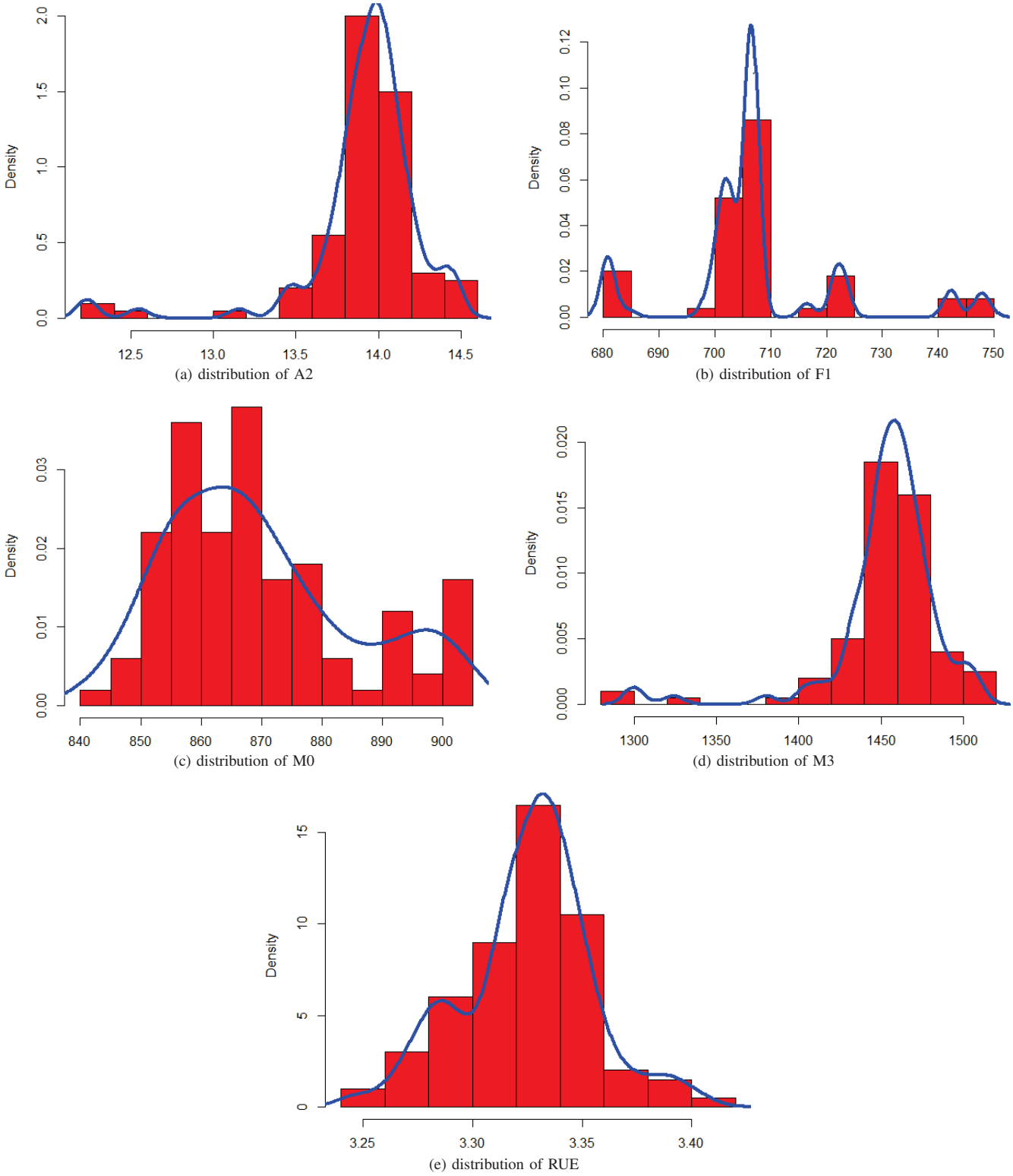


Fig. 2 (a), (b), (c), (d), (e) give the distributions with kernel interpolation lines of the five components in parameter vector θ: A2, F1, M0, M3, RUE, with sample size = 500

records are divided into two parts: the training set with 600 records and testing set with 120 records.

Samples of different sizes, like 300, 400, 500, are extracted from the training set, with which the optimization process is carried out. And these processes are repeated for 100 times from which we get the distributions (see Figure 2) and we can obtain the mean and standard error evaluating the estimation uncertainty for each parameter, see Tables V and

TABLE V
MAXIMUM POSTERIORS OF THE FIVE PARAMETERS WITH DIFFERENT SAMPLE SIZE

sample size	A2	F1	M0	M3	RUE
300	13.9157	708.1574	876.6731	1453.391	3.3143
400	14.0123	707.9132	868.7222	1462.214	3.3242
500	13.9189	707.3595	869.3464	1453.617	3.3253

TABLE VI
STANDARD ERRORS OF THE FIVE PARAMETERS WITH DIFFERENT SAMPLE SIZES

sample size	A2	F1	M0	M3	RUE
300	0.6172	20.4081	37.0013	57.7638	0.0556
400	0.3662	18.5054	19.084	35.0282	0.0346
500	0.3572	14.7435	15.3157	33.4155	0.0292

TABLE VII
FITNESS AND PREDICTION CAPACITY EVALUATION

sample size	RMSEP0	RMSEP	MAEP0	MAEP
300	70.34584	72.58696	6.01%	6.23%
400	70.29415	72.47566	5.93%	6.18%
500	70.21552	72.38343	5.87%	6.12%

VI respectively. The fitness and the prediction capacity are evaluated and compared according to their RMSEP0, RMSEP, MAEP0 and MAEP as shown in Table VII.

IV. DATA-DRIVEN METHODS

Data-driven regression is a family of statistical methods that do not impose domain-based knowledge in the regression process, and only consider data sets in the form $(U_i, y_i)_{1\leq i\leq N}$, where U_i represents the multidimensional input variables and y_i represents the response variable. The goal is to highlight the relationships that may exist between the different data and to derive statistical information which allows a more succinct description of the data [26].

In agriculture research, data-driven methods have known an increasing attention in the last years. Classically, the goal is to predict crop yield y_i from the climate series records U_i as in Equ.3. As can be easily understood, climatic data between different scenarios raise some colinearity issues (data the analysis of these data often comes across colinearity issues. A criteria named "condition number" is widely used to evaluate this property [28].

A. Condition Number

In numerical analysis, condition number measures the dependence of the solution of a numerical problem with respect to the data of the problem, in order to check the validity of a calculated solution with respect to these data. More generally, it can be said that the condition number associated with a problem is a measure of the difficulty of numerical computation of the problem, a problem with small condition number is said to be well conditioned [29].

For the multivariate linear model $Y = X\beta + \epsilon$ with standardized covariates, the associated condition number is calculated by the following expression:

$$\kappa(X) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{12}$$

where λ_{max}, λ_{min} represent maximal and minimal eigenvalues of $(X^{\mathrm{T}}X)$ respectively.

In our study, the condition number $\kappa(U) = 1.56 \times 10^4$ is high, which corresponds to the fact that the climate variable series are strongly correlated, as expected.

B. Solutions to High Correlation

In order to deal with the high correlation, devise methods have been propose. Generally, they can be classified into classical statistical and machine learning methods [30] according to their different principals.

Normally, the classical statistic methods try to reduce dimension by choosing independent components. Meanwhile the selected components should keep as much as possible the discrepancy of the initial data U. Ridge regression [31], [32], principal component regression (PCA) [33] or partial least squared regression (PLS)[34] are classical methods for this purpose, contrary to lasso regression which is known to handle poorly correlations.

We also test classical machine learning methods: Decision Tree (DT), Random Forest (RF), KNN, ANN and SVM regression are tested in this study [36].

C. Results

A 5-fold cross validation is applied, and the average RMSEP0, RMSEP, MAEP0 and MAEP are compared to choose the best data-driven method to do crop yield prediction. The results are listed in Table VIII.

The methods based on dimension reduction, such as ridge and lasso regression, PCA and PLS, demonstrated their approximate ability in fitting model (with RMSPE about 49 g/m^2). And their predictive capacities are also close to each other (with RMSPE around 58 g/m^2). As for the machine learning methods, the differences are stronger because of their different learning principles. The random forest and ANN, which are considered to be the typical nonlinear model, show their remarkable capacity in model fitness (with RMSE0 less than 30 g/m^2). It also implies on the contrary that the nonlinearity of the plant biological process. In brief, the random forest is chosen to best the most effective data-driven method for crop yield prediction in this research.

V. CONCLUSION AND DISCUSSION

In this work, we have studied a new methodology named multiple scenarios parameter estimation to calibrate the plant model with another data frame $\{U, y\}$. In order to improve the robustness of this method, a synchronous parallelised PSO optimization, which is a global optimization algorithm, has also been introduced. The CORNFLO model is well calibrated and its prediction capacity is improved with MAEP 6.11%. On the other hand, some data-driven approaches have also been applied to predict the crop yield. These methods are commonly used to deal with the collinearity, since there is always a very strong correlation among the climate variables series. Finally, Random Forest regression is demonstrated to be the most efficient in terms of crop yield prediction with MAEP 4.27%.

TABLE VIII
FITNESS AND PREDICTION CAPACITY EVALUATION OF DATA-DRIVEN METHODS

Method	RMSEP0	RMSEP	MAEP0	MAEP
Ridge	48.969	57.260	4.01%	4.63%
Lasso	48.004	57.345	3.92%	4.64%
PCA	49.536	58.043	4.14%	4.76%
PLS	49.356	58.112	4.21%	4.82%
DT	49.855	64.417	4.03%	5.05%
RF	**21.897**	**54.018**	**1.69**%	**4.27**%
KNN	43.600	57.459	3.53%	4.62%
ANN	29.247	69.786	2.17%	5.11%
SVM	53.793	58.931	4.17%	4.75%

From the above, the data-driven approaches outperforms our model-driven approach on the principle of crop yield prediction. Nonetheless, mechanistic models present some advantages: the fact that parameters and intermediate output variables have biological meanings can help understand crop growth, like for example stress periods or the identification of biological parameters of interest, while such analysis are not possible with data-based methods. Thus, some effort will be made in the future to combine these two methodologies.

APPENDIX

Description of the CORNFLO Model

The CORNFLO model is a plant growth model that simulates the growth and yield of maize [37]. It is inspired from SUNFLO model for sunflower [10]. It consists of the following three important modules: crop phenology module, morphogenesis and photosynthesis module, biomass production and distribution module.

A. Phenology Module

Normally, the initiation and the development of an organ depends on the cumulative time and also their environmental temperature. So does the development of plant from one stage to another. In order to combine the influence of these two factors and to simplify the model complexity, a new notion (variable) named "thermal time" (cumulative heat) has been introduced into the plant modeling [38]. It has been proved in [39] that the cumulative heat used in this way often has a significant advantage over the use of normal calendar time. In this model, the development of the plant is characterized by a succession of physiological stages according to four phases calculated by the thermal time: flowering bud appearance time ($E1$ °C·days), beginning of flowering ($F1$ °C·days), beginning of grain filling ($M0$ °C·days) and physiological maturity ($M3$ °C·days). The daily efficient temperature ($T_{eff}(d)$°C) at day d is calculated by (13):

$$T_{eff}(d) = T_{moy}(d) - T_{base}. \tag{13}$$

with $T_{moy}(d)$ the daily average temperature at day d and T_{base} the phenology base temperature. According to [40], it is generally equal to 10 for maize. The thermal time $TT(d)$(°C·days) at day d is calculated as the accumulation of $T_{eff}(d)$. Then it is used to determine at which stage the plant is.

B. Morphogenesis and Photosynthesis Module

$A3(cm^3)$ is the parameter giving the potential surface of the larger leaf of the simulated plant and $A2$ is the rank of the leaf which has the largest leaf surface in the entire plant growth period. $Ae(i)(cm^2)$ is the largest leaf surface for the leaf of rank i. It is calculated with $A2$ and $A3$ in (14):

$$Ae(i) = A3 * e^{-0.0344*(i-A2)^2+0.000731*(i-A2)^3} \tag{14}$$

The thermal time of appearance and death for leaf i are designated as $TT_{de\_pot}(i)$(°C·days) and $TT_{fe\_pot}(i)$(°C·days). The duration of leaf expansion denoted as $TT_{exp\_pot}(i)$(°C·days), expressed in thermal time is as following:

$$TT_{depot}(i) = \begin{cases} 1, & for\ 0 \le i \le 2 \\ TT_{depot}(i-2) * phyllodeini, & for\ 2 \le i \le 5 \\ TT_{depot}(4) + (i-5) * phyllodepot, & for\ 5 \le i \end{cases} \tag{15}$$

$$TT_{fepot}(i) = \begin{cases} TT_{fepot}(i-1) + (i-8) * phyllofepot, & for\ i > 8 \\ TT_{fepot}(i-1) * phyllofeini, & for\ i \le 8 \end{cases} \tag{16}$$

$$TT_{exp\_pot}(i) = TT_{fe\_pot}(i) - TT_{de\_pot}(i) \tag{17}$$

where $phyllodeini$(°C·days) and $phyllofeini$(°C·days) are parameters of phyllochrone for leaf of rank below 8. For the other leaf, the beginning and ending time are noted as $phyllodepot$(°C·days) and $phyllofepot$(°C·days).They are the variables that depend on the number of leaves NFF, the stage $F1$, and the parameter $Ratio\_phyllofede$ defined as (18) and (19):

$$phyllofepot = \frac{F1 - 7 * phyllofeini}{NFF - 8} \tag{18}$$

$$phyllodepot = phyllofepot * Ratio\_phyllofede \tag{19}$$

The expansion speed of the leaf i is calculated by its potential surface of leaf $Ae(i)$ and its thermal expansion time $TT_{exp\_pot}(i)$:

$$V_{exp\_pot}(i) = \frac{Ae(i)}{TT_{exp\_pot}(i)} \tag{20}$$

Thus, the surface of leaf i on day d, $SF_{i\_pot}(d,i)(cm^2)$ is given by (21):

$$SF_{pot}(d,i) = SF_{pot}(d-1,i) + V_{exp\_pot}(d,i) * T_{eff}(d) \tag{21}$$

It is initialized by $SF_{pot}(0,i) = 0, \forall i$.Then, the total leaf area $SFP_{pot}(d)(cm^2)$ at day d is given by (22):

$$SFP_{pot}(d) = \sum_{i=1}^{n} SF_{pot}(d,i) \tag{22}$$

The ratio of the green portion of all the leaf surface is noted as $Frac\_verte$. This coefficient will be used to calculate the index of leaf area $LAI\_pot(cm^2/m^2)$ as in (24):

$$Frac\_verte(d) = 1 - \frac{TT(d) - F1}{M3 - F1} \tag{23}$$

$$LAI\_pot(d) = dens * SFP\_pot(d) * Frac\_verte(d)/10000 \quad (24)$$

where $dens(m^{-2})$ is the planting density of maize.

C. Biomass Production and Biomass Distribution Module

In order to calculate the biomass, another two parameters should be introduced: the radiation absorption efficiency $E_i(d)$ and the radiation use efficiency $E_b(d)(g.MJ^{-1})$. They are defined as in (25) and (26):

$$E_i(d) = 0.95 * (1 - e^{-k\_coeff * LAI\_pot(d)}) \quad (25)$$

$$E_b(d) = \begin{cases} RUE & for\ M0 \geq TT(d) \\ RUE * (1 - \frac{TT(d)-M0}{M3-M0}) & for\ M3 \geq TT(d) > M0 \\ 0, & for\ TT(d) > M3 \end{cases} \quad (26)$$

with the extinction coefficient k_coeff and the maximum radiation use efficiency $RUE(g.MJ^{-1})$. Both are genotype parameters.

According to the energetic approach of [41], the daily biomass production $dMS(g.m^{-2})$ should be calculate by the energy transferred from the solar energy. In this model, the solar energy is represented by the radiation $RG(d)(MJ.m^{-2})$. Finally, a climate efficiency coefficient which is relatively constant at 0.48 will be used to adjust this equation:

$$dMS(d) = 0.48 * RG(d) * E_b * E_i(d) \quad (27)$$

So the total biomass at day d,$MS_{tot}(d)(g.m^{-2})$ results from the accumulation of the daily biomass production given in (28):

$$MS_{tot}(d) = \sum_{t=1}^{d} dMS(t) \quad (28)$$

In order to determine the final crop yield, denoted as $MS_{grain}(d)$, a constant proportion of biomass (harvest index, HI) is assigned to the grain compartment:

$$y = MS_{grain}(d) = MS_{tot}(d) * HI \quad (29)$$

ACKNOWLEDGMENT

Xiangtuo Chen is supported by a fellowship from the China Scholarship Council.

REFERENCES

[1] Drummond S T, Sudduth K A, Joshi A, et al. *Statistical and neural methods for site-specific yield prediction*(J). Transactions-American Society of Agricultural Engineers, 2003, 46(1): 5-16.

[2] Liu J, Goering C E, Tian L. *A neural network for setting target corn yields*(J). Transactions-American Society of Agricultural Engineers, 2001, 44(3): 705-714.

[3] Kang F. *Modèles de croissance de plantes et méthodologies adaptees à leur parametrisation pour l'analyse des phénotypes*(D)0.5em minus 0.4emChatenay-Malabry, Ecole centrale de Paris, 2013.

[4] Cournede P H, Chen Y, Wu Q, Baey C, Bayol *Development and evaluation of plant growth models: Methodology and implementation in the PYGMALION platform*, 0.5em minus 0.4emMathematical Modelling of Natural Phenomena, 2013, 8(4): 112-130.

[5] Cournede P H, Letort V, Mathieu A, et al. *Some parameter estimation issues in functional-structural plant modelling*(J). Mathematical Modelling of Natural Phenomena, 2011, 6(2): 133-159.

[6] Goodwin G C, Payne R L. *Dynamic system identification: experiment design and data analysis*(J). 1977.

[7] Wallach D, Goffinet B. *Mean squared error of prediction in models for studying ecological and agronomic systems*(J). Biometrics, 1987: 561-573.

[8] Wallach D. *Evaluating crop models*(J). Working with Dynamic Crop Models Evaluation, Analysis, Parameterization, and Applications, Elsevier, Amsterdam, 2006: 11-54.

[9] Messéan A, Bernard H, de Turckheim É. *Concevoir et construire la décision: Démarches en agriculture, agroalimentaire et espace rural*(M). Editions Quae, 2009.

[10] Lecoeur J, Poiré-Lassus R, Christophe A, et al. *Quantifying physiological determinants of genetic variation for yield potential in sunflower. SUNFLO: a model-based analysis*(J). Functional plant biology, 2011, 38(3): 246-259.

[11] Brun F, Wallach D, Makowski D, et al. *Working with dynamic crop models: Evaluation, analysis, parameterization, and applications*(M). Elsevier, 2006.

[12] Saltelli A, Tarantola S, Campolongo F, et al. *Sensitivity analysis in practice: a guide to assessing scientific models*(M). John Wiley and Sons, 2004.

[13] Saltelli A, Chan K, and Scott EM, eds. *Sensitivity analysis*. Vol. 1. New York: Wiley, 2000.

[14] Wu, QL, Cournède PH and Mathieu, A *An efficient computational method for global sensitivity analysis and its application to tree growth modelling*(J). Reliability Engineering & System Safety, 2012, 107: 35-43.

[15] Cournède PH, Chen Y, Wu QL, Baey C, Bayol B *Development and evaluation of plant growth models: Methodology and implementation in the pygmalion platform*(J). Mathematical Modelling of Natural Phenomena, 2013, 8: 112-130.

[16] Eberhart R, Kennedy J. *A new optimizer using particle swarm theory*(C) Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on. IEEE, 1995: 39-43.

[17] Shi Y. *Particle swarm optimization: developments, applications and resources*(C) Evolutionary computation, 2001. Proceedings of the 2001 Congress on. IEEE, 2001, 1: 81-86.

[18] Shi Y, Eberhart R. *Parameter selection in particle swarm optimization*(C) Evolutionary programming VII. Springer Berlin/Heidelberg, 1998: 591-600.

[19] Kennedy J. *Particle swarm optimization*(M) Encyclopedia of machine learning. Springer US, 2011: 760-766.

[20] Kennedy J, Mendes R. *Population structure and particle swarm performance*(C) Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on. IEEE, 2002, 2: 1671-1676.

[21] Clerc M. *The swarm and the queen: towards a deterministic and adaptive particle swarm optimization*(C) Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on. IEEE, 1999, 3: 1951-1957.

[22] Shi Y, Eberhart R. *A modified particle swarm optimizer*(C) Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on. IEEE, 1998: 69-73.

[23] Eberhart R C, Shi Y. *Comparing inertia weights and constriction factors in particle swarm optimization*(C) Evolutionary Computation, 2000. Proceedings of the 2000 Congress on. IEEE, 2000, 1: 84-88.

[24] Schutte J F, Reinbolt J A, Fregly B J, et al. *Parallel global optimization with the particle swarm algorithm*(J). International journal for numerical methods in engineering, 2004, 61(13): 2296.

[25] Clarke F H. *Optimization and nonsmooth analysis*(M). Society for Industrial and Applied Mathematics, 1990.

[26] Singh A, Ganapathysubramanian B, Singh A K, et al. *Machine learning for high-throughput stress phenotyping in plants*(J). Trends in plant science, 2016, 21(2): 110-124.

[27] Von Storch H. *Misuses of statistical analysis in climate research*(M) Analysis of Climate Variability. Springer Berlin Heidelberg, 1999: 11-26.

[28] Belsley D A. *Conditioning diagnostics*(M). John Wiley & Sons, Inc., 1991.

[29] Cline A K, Moler C B, Stewart G W, et al. *An estimate for the condition number of a matrix*(J). SIAM Journal on Numerical Analysis, 1979, 16(2): 368-375.

[30] Yin S, Ding S X, Haghani A, et al. *A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process*(J). Journal of Process Control, 2012, 22(9): 1567-1581.

[31] Shin M Y. *The use of ridge regression for yield prediction models with multicollinearity problems*(J).. Journal of Korean Forestry Society, 1990, 79(3): 260-268.

[32] Hassan S S, Farhan M, Mangayil R, et al. *Bioprocess data mining using regularized regression and random forests*(J). BMC systems biology, 2013, 7(1): S5.
[33] Chang J, Clay D E, Dalsted K, et al. Corn (L.) *yield prediction using multispectral and multidate reflectance*(J). Agronomy journal, 2003, 95(6): 1447-1453.
[34] Abdel-Rahman E M, Mutanga O, Odindi J, et al. *A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data*(J). Computers and Electronics in Agriculture, 2014, 106: 11-19.
[35] Hall M A. *Correlation-based feature selection of discrete and numeric class machine learning*(J). 2000.
[36] Ru G. *Data mining of agricultural yield data: A comparison of regression models*(C) Industrial Conference on Data Mining. Springer Berlin Heidelberg, 2009: 24-37.
[37] Albuquerque M C F, de Carvalho N M. *Effect of the type of environmental stress on the emergence of sunflower (Helianthus annus L.), soybean (Glycine max (L.) Merril) and maize (Zea mays L.) seeds with different levels of vigor*(J). Seed Science and Technology (Switzerland), 2003, 31(2): 465-479.
[38] Midmore E K, McCartan S A, Jinks R L, et al. *Using thermal time models to predict germination of five provenances of silver birch (Betula pendula Roth) in southern England*(J). Silva Fennica, 2015, 49(2).
[39] Atwell B J, Kriedemann P E, Turnbull C G N. *Plants in action: adaptation in nature, performance in cultivation*(M). Macmillan Education AU, 1999.
[40] Williams M M. *Agronomics and economics of plant population density on processing sweet corn*(J). Field Crops Research, 2012, 128: 55-61.
[41] Monteith J L, Moss C J. *Climate and the efficiency of crop production in Britain (and discussion)*(J). Philosophical Transactions of the Royal Society of London B: Biological Sciences, 1977, 281(980): 277-294.

Xiangtuo Chen is currently pursuing his Ph.D. applied mathematics, received his M.Sc. degree from Beihang university, Beijing, Chine, and his engineer degree from Ecole Centrale de Pekin in June 2015. His research areas are Mathematical modeling of plant growth, parameter estimation, Particle swarm optimization, and the machine learning methods in plant modeling.

Paul-Henry Cournède is a Professor in applied mathematics with the CentraleSupélec , the Head of Laboratory MICS and also head of the Research Team Biomathematics, Paris, France. His current research interests include Mathematical modeling of plant growth, functional-structural plant growth models, sensitivity analysis, parameter estimation and optimization. He is also very interested in data-driven approaches and their application in plant modeling.