Computers and Electronics in Agriculture 154 (2018) 400–410

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

# A dynamic agricultural prediction system for large-scale drought assessment on the Sunway TaihuLight supercomputer

Xiao Huang[a,b], Chaoqing Yu[a,c,*], Jiarui Fang[d], Guorui Huang[a,e], Shaoqiang Ni[a], Jim Hall[f], Conrad Zorn[f], Xiaomeng Huang[a], Wenyuan Zhang[a,g]

[a] Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing, China
[b] Norwegian Institute of Bioeconomy Research, Saerheim, Klepp st., Norway
[c] AI for Earth Laboratory, Cross-Strait Tsinghua Research Institute, Beijing, China
[d] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[e] Twente Water Centre, University of Twente, Twente, Netherlands
[f] Environmental Change Institute, University of Oxford, Oxford, UK
[g] Department of Zoology, University of Oxford, Oxford, UK

ARTICLE INFO

*Keywords:*
Drought
Dynamic prediction
Supercomputer
Accuracy
Risk analysis

ABSTRACT

Crop models are widely used to evaluate the response of crop growth to drought. However, over large geographic regions, the most advanced models are often restricted by available computing resource. This limits capacity to undertake uncertainty analysis and prohibits the use of models in real-time ensemble forecasting systems. This study addresses these concerns by presenting an integrated system for the dynamic prediction and assessment of agricultural yield using the top-ranked Sunway TaihuLight supercomputer platform. This system enables parallelization and acceleration for the existing AquaCrop, DNDC (DeNitrification and DeComposition) and SWAP (Soil Water Atmosphere Plant) models, thus facilitating multi-model ensemble and parameter optimization and subsequent drought risk analysis in multiple regions and at multiple scales. The high computing capability also opens up the possibility of real-time simulation during droughts, providing the basis for more effective drought management. Initial testing with varying core group numbers shows that computation time can be reduced by between 2.6 and 3.6 times. Based on the powerful computing capacity, a county-level model parameter optimization (2043 counties for 1996–2007) by Bayesian inference and multi-model ensemble using BMA (Bayesian Model Average) method were performed, demonstrating the enhancements in predictive accuracy that can be achieved. An application of this system is presented predicting the impacts of the drought of May–July 2017 on maize yield in North and Northeast China. The spatial variability in yield losses is presented demonstrating new capability to provide high resolution information with associated uncertainty estimates.

## 1. Introduction

The growing pressure on food security is expected to continue in the coming decades due to global climate change and population growth (Adams et al., 1998; Bloom, 2011; Field, 2012). The risk of agricultural production, posed by the increasing frequency and severity of extreme events (Howden et al., 2007; Lobell et al., 2008), calls for more accurate and effective predictions of crop growth dynamic on large scales (Rosenzweig and Tubiello, 2007; Yu et al., 2018). Process-based crop models are increasingly popular tools for climate change impact and adaptation studies (Rosenzweig et al., 2013, 2014). A real-time decision support system based on crop model simulations and predictions could provide useful information for users with different goals (e.g. farmers, decision makers), and hence improve our ability for disaster mitigation (Van Ittersum et al., 2008; Jakku and Thorburn, 2010; Yu et al., 2014). However, such applications still face a number of key challenges (Holzworth et al., 2015). We identify at least three of these as: (i) reducing predictive uncertainty, (ii) the computing requirement to enable simulations at sufficiently high spatial and temporal resolutions, and (iii) the evaluation of yield risk.

The first challenge is how to reduce the uncertainty of model simulations and provide more accurate predictions. As most crop models are primarily developed based on site experiments, their scaling and application to larger areas, such as county, state or nation, are likely to introduce predictive uncertainties (Ewert et al., 2015), arising from assumptions regarding model structures, model parameters, and

* Corresponding author.
*E-mail address:* chaoqingyu@yahoo.com (C. Yu).

https://doi.org/10.1016/j.compag.2018.07.027
Received 21 March 2018; Received in revised form 19 June 2018; Accepted 18 July 2018
Available online 27 September 2018

calibration/validation data, amongst others. Several methods have been explored to address this issue, including Bayesian inference (Iizumi et al., 2009; Dumont et al., 2014) and parameter optimization (Guo et al., 2006) for parameter uncertainty, multi-model ensemble predictions for model structure uncertainty (Martre et al., 2015; Huang et al., 2017), remote sensing data assimilation to correct the state variables (De Wit and Van Diepen, 2007) and the ensemble of climatic force for input uncertainty (Baigorria et al., 2008; Tao et al., 2009). However, most of these studies only address an individual source of uncertainty with fewer integrating multiple sources of uncertainty systematically. Therefore crop predictions in large region still remain highly uncertain.

The second challenge relates to computing efficiency and the need to design an advanced tool for large-scale crop simulation with high performance computing (HPC) technology. There are a number of major obstacles to achieve this: (i) The computational requirement for crop modeling is extremely large, both in the real-time simulation and scenario prediction at high spatial and temporal resolutions and for robust uncertainty analyses such as Markov Chain Monte Carlo (MCMC) methods. (ii) Unlike global climate models (Neale et al., 2010) or land models (Oleson et al., 2010) which consider the possibility of parallel optimization across a computer cluster at the beginning of their design, most traditional crop models are developed targeting to deploy on one single node. The much needed redesigning of workflows in these traditional models will require significant effort to exploit parallelism with large-size computing resources (Holzworth et al., 2015). (iii) The computing kernels inside crop models work in serial mode on general proposed processors. Complex dependencies exist between different kernels, which make it more difficult to get accelerated in parallel by multiple-threading method on powerful accelerators like Graphic Processor Units (GPUs) and Intel Xeon Phis. Several studies have made attempt to use HPC technologies in large clusters for regional simulation (Vital et al., 2013; Zhao et al., 2013; Elliott et al., 2014). These researches mainly focus on the parallel computation in grid level or between different models, as well as the high-speed transformation of model input/output files into standard format. However, we find no evidence in the literature attempting to further accelerate the scientific algorithms of crop models, which account for the majority of simulation time in crop modeling processes.

The final challenge identified relates to the post-processing of simulation results into useful information for users – in particular the spatial distribution of production losses during periods of drought and the severity of yield losses for given return periods. The spatial distribution of production losses provided by the crop model simulation across large scale is essential for the monitoring of drought evolution (Báez-González et al., 2002; Launay and Guerif, 2005). To reveal the severity of yield losses in the local history, return periods seem to be a more useful index because it clearly demonstrates the comparable features in longer climatic background and is widely accepted by the public (Fernández and Salas, 1999; Bonaccorso et al., 2003). Besides the marginal distribution of loss in each region (Yu et al., 2014; Skakun et al., 2016), the joint distribution in multiple areas (Bárdossy and Pegram, 2009; AghaKouchak et al., 2010; Gaupp et al., 2017) can offer more critical information of the spatial correlation of agricultural production, and it will lead the decision makers to a comprehensive assessment of the overall agricultural risk.

In this paper, we seek to address these three challenges through the development of a novel integrated decision support system. The crop models AquaCrop (Steduto et al., 2009), SWAP (Kroes et al., 2000) and the biogeochemical model DNDC (Li et al., 1992), which are all validated from field scale to region scale for yield prediction (De Wit and Van Diepen, 2007; Heng et al., 2009; Yu et al., 2014; Huang et al., 2015, 2017), are mapped into a supercomputer platform, entitled Sunway TaihuLight (Fu et al., 2016; Zhao et al., 2018). The crop simulation in each model is modified to fit the hardware architecture of this supercomputer for model acceleration. Bayesian inference and BMA method (Huang et al., 2017) are used in model prediction to reduce the parameter uncertainty and model structure uncertainty respectively. We apply a scenario analysis method (Yu et al., 2014) to assess the evolution of crop growth under uncertain drought development. Finally, the system estimates both the marginal distribution of yield loss in different scales and the joint distribution of different regions using a Copula method.

Section 2 briefly introduces the architecture of the Sunway TaihuLight supercomputer. Section 3 describes the integrated framework for agricultural risk analysis, including the methodology and technological process. In Section 4, the case studies about model acceleration, parameter optimization, dynamic drought impact assessment and probability analysis of using this system for real practice are illustrated. Finally the discussion and summary is presented in Section 5.

## 2. The platform: Sunway TaihuLight supercomputer

The Sunway TaihuLight supercomputer based at the National Supercomputing Center in Wuxi, China (Fu et al., 2016) has held the 47–50th (most recent) TOP500 rankings of global supercomputers (https://www.top500.org/lists/). It achieves a Linpack performance (Dongarra et al., 2003) of 93 PetaFlops from a theoretical peak of 125 PetaFlops, with computing power originating from 40,000 SW26010 processors arranged in a unique heterogeneous many-core architecture and memory hierarchy (Fig. 1). This processor contains four Core Groups (CGs) connected via a network on chip (NoC). Each CG consists of a management processing element (MPE) and 64 computing processing elements (CPEs) organized into an 8X8 CPE cluster (Fang et al., 2017). Both MPEs and CPEs are complete 64-bit Reduced Instruction Set Computing (RISC) cores working at a frequency of 1.45 GHz, but they adopt different memory hierarchies. MPE has a 32 KB L1 data cache, a 32 KB L1 instruction cache and a 256 KB L2 cache. CPE has a 16 KB L1 instruction cache and a 64 KB local directive memory (LDM) as the user-controlled fast buffer. Users need to explicitly control data placement in the LDM. CPE clusters account for most of computing power and have access to the main memory of CGs indirectly through LDM, while MPE can directly access main memory and system interface for communication. In addition, a 128-bit memory controller is equipped with each CG to access the 8 GB double-Data-Rate Three Synchronous Dynamic Random Access Memory (DDR3) local memory with a theoretical peak bandwidth of 34 GB/s. Such design of processor enables relatively good power efficiency while maintaining high computing capacity.

The 40,000 SW26010 processors are connected using Peripheral Component Interconnect Express (PCI-E 3.0) connections in a customized Sunway Network. Peer-to-peer bidirectional communication between two processors via Message Passing Interface (MPI) is at 12 GB/s and a latency of about 1 μs. Further technical detail is provided by (Fu et al., 2016). The supercomputer is used for a wide range of earth system modelling, including global atmosphere dynamics (Fu et al., 2017), earthquake simulation (https://awards.acm.org/bell), sea ice modelling (Li et al., 2017) and ocean cycle (Qiao et al., 2016). We choose this supercomputer for our large-scale agricultural prediction because of its unique architecture of hardware for earth system modelling and the abundant computing resources available for real-time simulation, as well as the potential expandability of our system coupling with other earth system modules (e.g. atmosphere model) on this platform.

## 3. Methodology

### 3.1. The main structure of the system

We develop an agricultural drought monitoring and yield prediction system using HPC technology on the Sunway TaihuLight supercomputer for the dynamic prediction of drought-induced agricultural losses. The

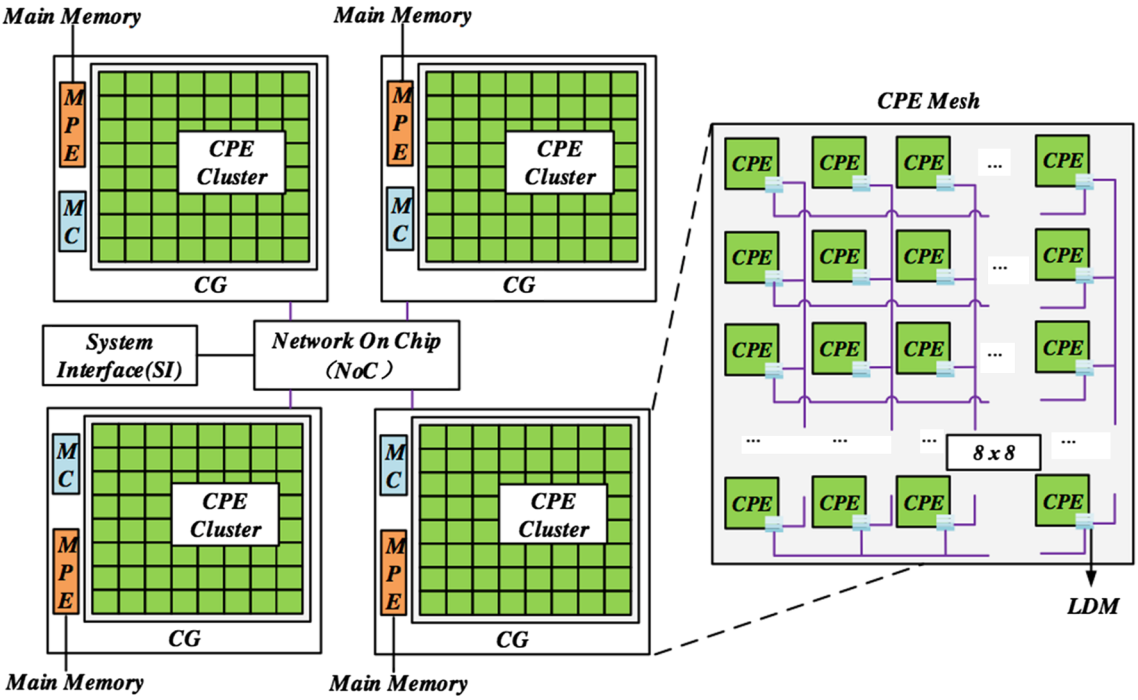


**Fig. 1.** The architecture of the SW26010 processor, modified from (Jiang et al., 2017).

framework of this system comprises three components: model optimization, dynamic prediction and result visualization (Fig. 2). The model optimization component provides the parameterization schemes for each model and the model weights for the ensemble. The dynamic prediction component collects climatic observations at daily time steps and monitors the development of precipitation deficits in areas with drought potential. Drought scenarios are used to evaluate the impacts of future drought development on yield formation. The climate data including the observations and future scenarios to the harvest time will be generated to drive the multi-model system for ensemble predictions. Finally, the model outputs are converted to the map of spatial distribution and risk assessment for drought mitigation. The following sections will explain each of these components in more detail.

### 3.2. Mapping multiple crop models onto the Sunway TaihuLight supercomputer

In this study, we use AquaCrop, SWAP and DNDC models for the prediction of crop growth. With water, radiation and nitrogen as the main driving element for each model, respectively (Huang et al., 2017), the multi-model ensemble prediction can be applicable in most cropland in China. For large-scale simulation, target regions are divided into a number of equal area rectangular grid cells or irregular grids based on the administrative boundary, assumed to be homogeneous in its environmental condition (e.g. soil, climate, farming practice) and independent from neighboring cells. Across the models, the crop growth modules and soil dynamic modules are usually the two main components with feedbacks between each other. In the growth modules, the

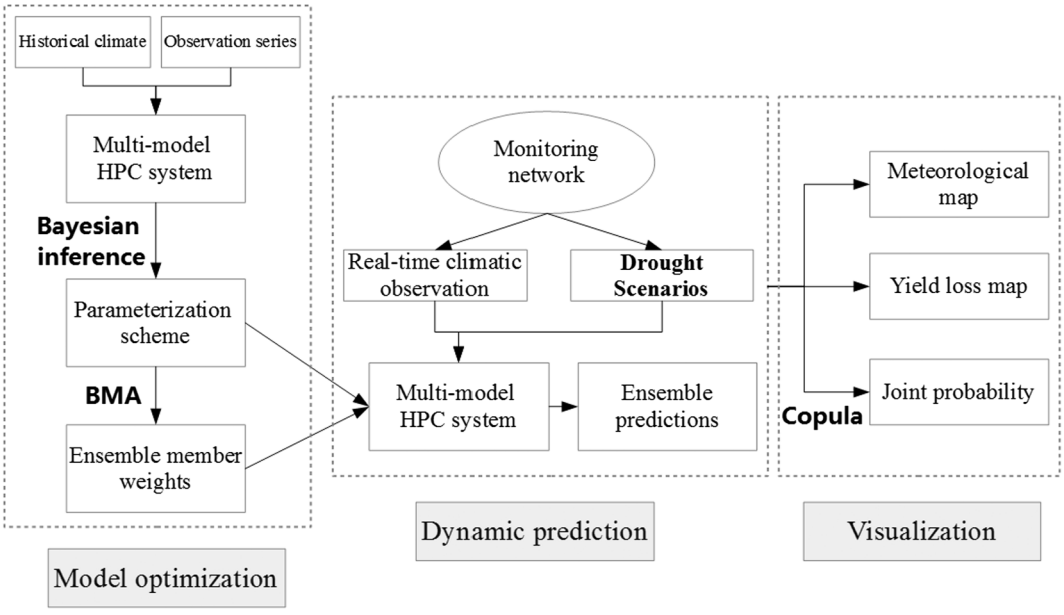


**Fig. 2.** The main structure of the system.

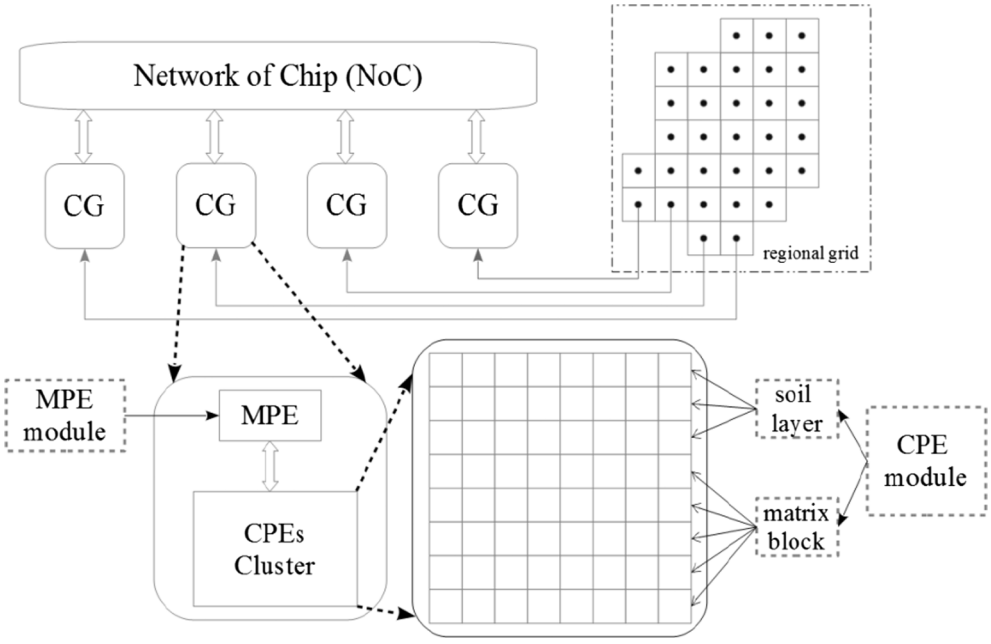


**Fig. 3.** Parallel strategy for CPE-based module and MPE-based module.

above-ground canopy of crop is depicted as a whole unit to model the processes of radiation interception, photosynthesis and biomass partition into different organs, which have a straightforward influence on soil processes (e.g. soil evaporation and nutrient uptake). In the soil modules, a soil column is divided into multiple layers to simulate the vertical dynamic of water and nutrition, and it in return reveals the water stress and root length conditions for crop growth. Simulation run times differ across each model due to additional complexities. In DNDC, the nitrification and denitrification processes account for over 70% of the running time, as such simulation is performed in hour step for each soil layers (usually greater than 20 layers) involving multiform chemical processes. For SWAP, the Richards equation for the soil water flow is the dominant part of calculation because of the iteration process for the convergence of matrix calculation (Kroes et al., 2009). These computing modules can be speed up by proper parallelization schemes.

As shown in Fig. 3, we design our paralleled crop models based on the hardware characteristics of the Sunway TaihuLight system in 3 steps:

(1) Similar to land models (e.g. Community Land Model, (Lawrence et al., 2011)), we adopt a multiple-processing strategy to distribute computing tasks of a number of grids on multiple CGs. When a process is launched on each CG to perform simulation for each independent grid, the driving data (e.g. climate, management, parameter values) is shared among different grids by communication with MPI. In this way, computation for different grids inside the targeting region can be conducted simultaneously.
(2) To exploit parallelism capacity inside CGs, we classify the scientific algorithms of crop modeling into two categories: (i) MPE-based modules that feature light computation and heavy communication and (ii) CPE-based modules that feature heavy computation and light communication. CPE-based modules include the independent calculation of state variable in each soil layer and the matrix calculation that can be divided into small blocks. The detailed classifications of each model are presented in Table 1.
(3) We apply a multiple-threading strategy to accelerate the CPE-based modules on the CPE cluster. A light-weight thread can be launched on each CPE to process data stored in its LDM in parallel of other CPEs of this cluster. Once the computation of CPE-based module finished, we explicitly copy all the relevant variables from the main memory to LDM. For instance, the values of variables (including water content, temperature, etc.) in the *i*th soil layer at current step will get passed to the LDM of the *i*th CPE for the calculation of nitrification process in DNDC, as well as the small blocks of the water head matrix in water flow simulation in SWAP. Then different CPEs use these data in LDM to calculate the substantial processes in parallel, and return the final results to the main memory for the next round of computation. The multi-layer acceleration is achieved by the AThread Library for C programming while matrix calculation by the BLAS (Basic Linear Algebra Subprograms) package in the Sunway TaihuLight supercomputer (Jiang et al., 2017).

**Table 1**
The classification of MPE-module and CPE-module for each model.

| Model | MPE-based module | CPE-base module |
|---|---|---|
| SWAP | Crop growth<br>ET and water uptake | Soil water flow<br>Soil temperature conductance |
| DNDC | Crop growth<br>Soil water flow<br>Soil temperature conductance<br>ET and water uptake | Nitrification and denitrification<br>Decomposition |
| AquaCrop | All | None |

Through the three steps above, we realize the parallelization and acceleration of our models. The detailed acceleration steps using this MPE + CPEs mode are shown in Fig. 4. Our system can be scaled to 20,000 processes with MPI for global-level simulation. It should be noted that from Table 1, only MPE is used for AquaCrop model due to the proportion of CPE-based modules is relatively low. The use of CPEs may lead to extra run time (time for data communication between main memory and LDM) according to our tests.

### 3.3. Model uncertainty

Here, the optimization of model parameters for each individual model is based on the maximization of the posterior probability density function (PDF) of model parameters. According to Bayes' theorem, the posterior PDF can be expressed as:

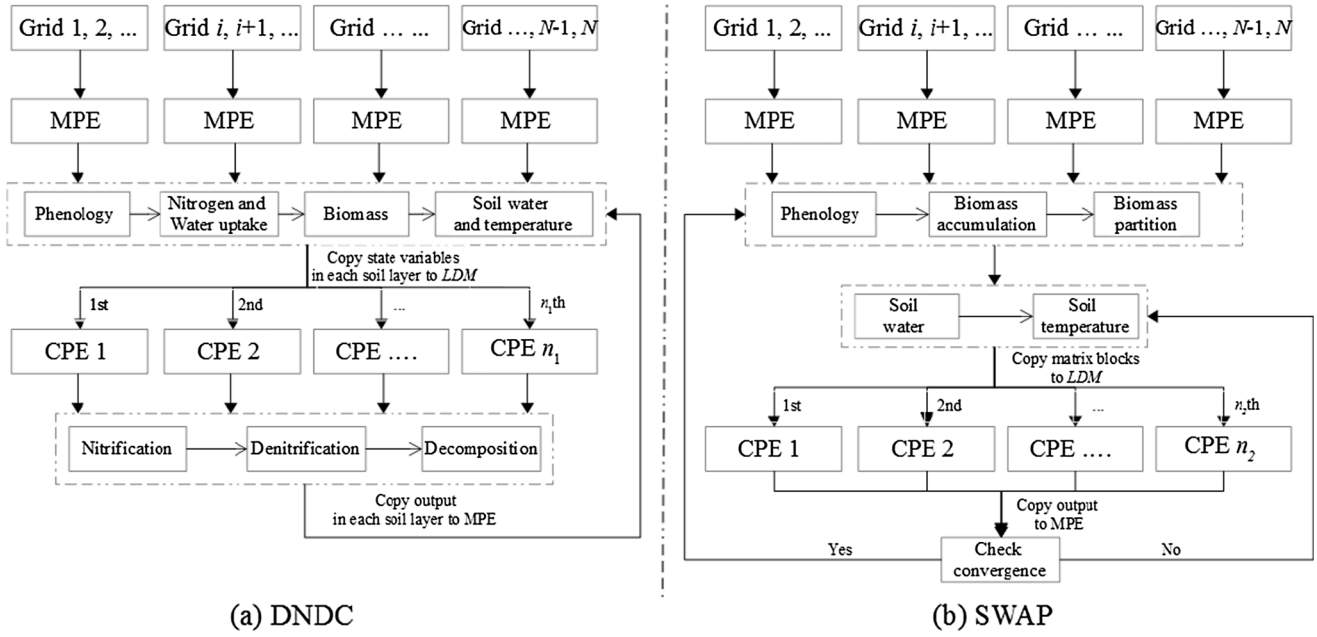


**Fig. 4.** The detailed acceleration steps using MPE + CPEs mode for (a) DNDC model and (b) SWAP model.

**Table 2**
The proportions of MPE-based module and CPE-base module to the total run time and the speedup ratio for CPE-module.

| Model | MPE-based module | CPE-base module in MPE mode | CPE-base module in MPE + CPEs mode | Speedup ratio |
|---|---|---|---|---|
| SWAP | 20% | 80% | 8% | 8 |
| DNDC | 15% | 85% | 10% | 8.5 |

$$p(\theta|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\theta)*p(\theta)}{p(\boldsymbol{Y})} \quad (1)$$

where $\theta$ is the vector of model parameters, $\boldsymbol{Y}$ is the series of observations and $p(\theta)$ is the parameter prior. To avoid further assumption of residual error in large-scale simulation, the Jeffreys' uninformative prior (Box and Tiao, 2011) is applied to Eq. (1) and it leads to:

$$p(\theta|\boldsymbol{Y}) \propto (\sim_{i=1}^{N}(Y_i - \widetilde{Y}_i(\theta))^2)^{-\frac{N}{2}} \quad (2)$$

where $N$ is the length of observations, and $\widetilde{Y}$ is the model simulation. For Eq. (2), we use the accelerated Markov Chain Monte Carlo (MCMC) algorithm 'Dream' (Vrugt et al., 2009) to obtain the posterior distribution of parameters and the optimal parameter sets for each crop model.

To address the uncertainty of model structure, we use BMA method (Huang et al., 2017) to generate the multi-model ensemble from the individual model predictions. For each ensemble member, it is assumed there is a conditional PDF $p(\boldsymbol{Y}|\widetilde{\boldsymbol{Y}}^j)$ of the yield $\boldsymbol{Y}$ with the $j$th model prediction $\widetilde{\boldsymbol{Y}}^j$considered. Following (Raftery et al., 2005) and the law of total probability, the BMA conditional PDF of the ensemble on all the individual models can be written as:

$$p(\boldsymbol{Y}|\widetilde{\boldsymbol{Y}}^1, \cdots, \widetilde{\boldsymbol{Y}}^M) = \sum_{j=1}^{M} w_j p(\boldsymbol{Y}|\widetilde{\boldsymbol{Y}}^j) \quad (3)$$

where $M$ is the total number of our crop models, and $w_j$ is the posterior probability (weight) for the $j$th model with $\sum w_j = 1$. Further assuming the conditional PDF $p(\boldsymbol{Y}|\widetilde{\boldsymbol{Y}}^j)$ follows a Gaussian distribution provides:

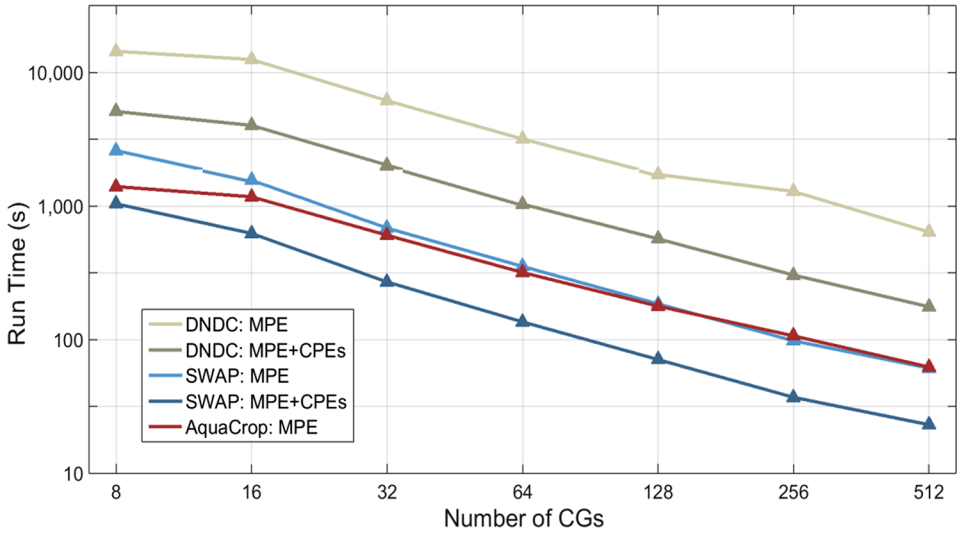


**Fig. 5.** The run time (10 iterations) of three models for 512 counties in MPE-only and MPE + CPEs modes.

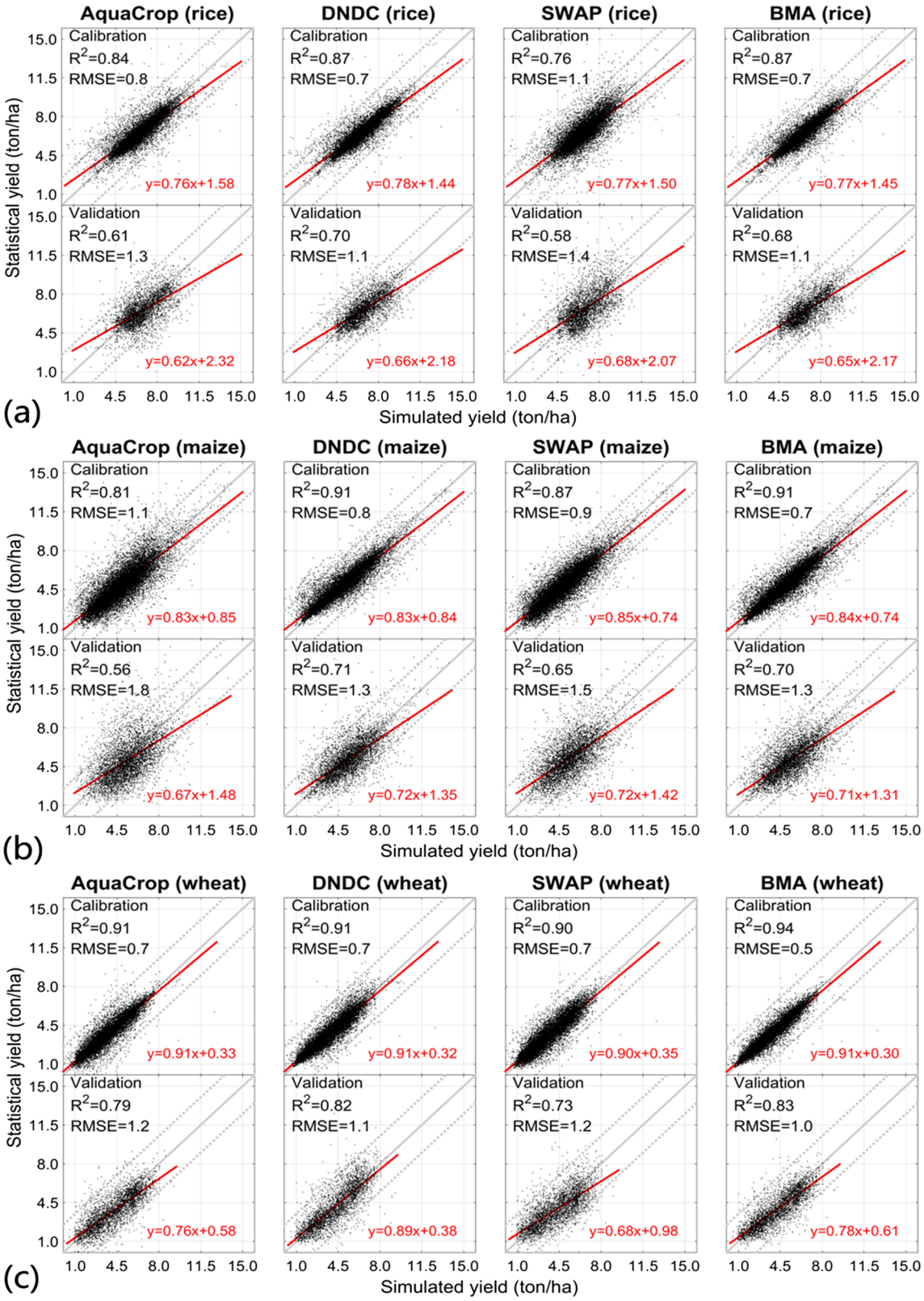


**Fig. 6.** (a)–(c) statistical yield observation (X axis, t/ha) VS model yield simulation (Y axis, t/ha); (d) The cumulative distribution of residual error;

$$\boldsymbol{Y}|\widetilde{\boldsymbol{Y}}^j \sum \boldsymbol{N}(a_j \widetilde{\boldsymbol{Y}}^j + b_j, \sigma_j^2) \tag{4}$$

where $a_j$ and $b_j$ are the coefficients of linear bias correction of $\widetilde{\boldsymbol{Y}}^j$. Then we can obtain the value of $w_j$ and $\sigma_j$ by maximizing Eq. (3). The ensemble member $\widetilde{\boldsymbol{Y}}^j$ used here is the optimal prediction with the maximum posterior PDF in Eq. (2).

### 3.4. Scenario based prediction of crop yields

The climate forecast is one of the most important factors when crop

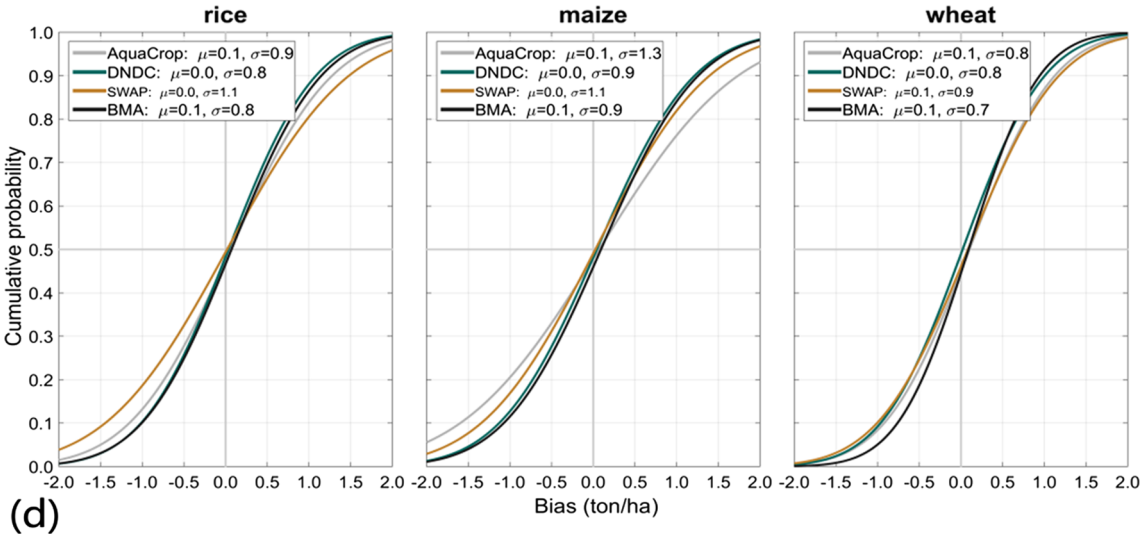


**Fig. 6.** (*continued*)

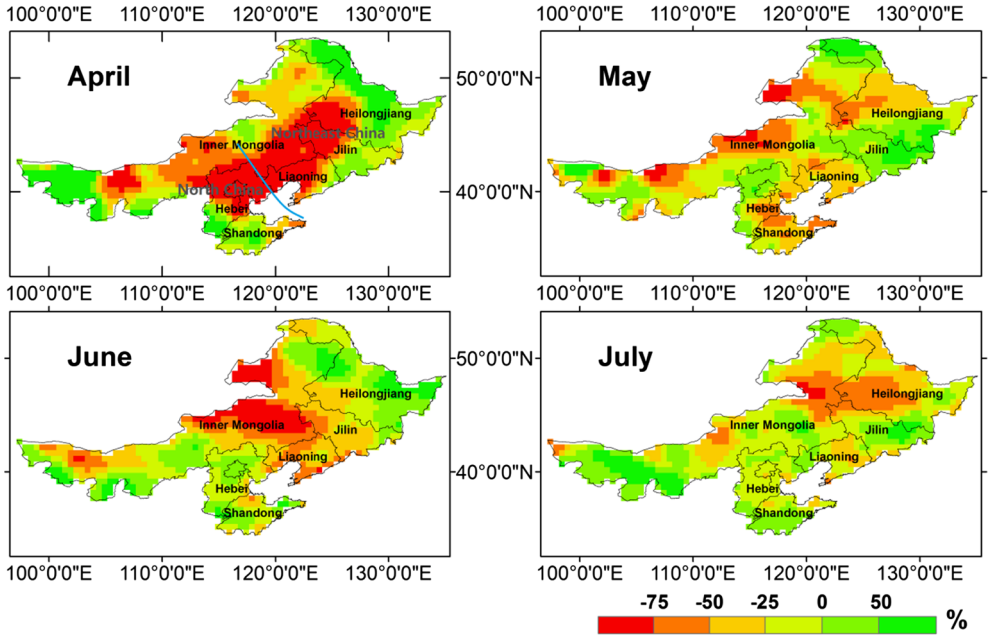


**Fig. 7.** The monthly precipitation anomaly in North and Northeast China from April to July in 2017 (departure from the average values of 2000–2016). Data source: CPC 0.5° × 0.5° Global Daily Gauge-Based Analysis of Precipitation, https://www.esrl.noaa.gov/psd/data/gridded/data.cpc.globalprecip.html.

models are used for agricultural predictions. However, most seasonal weather forecasts still remain uncertain for crop growth prediction while short-term forecast are insufficient to meet this demands for its limited duration (Hansen et al., 2006). Addressing the uncertainty of drought development, we use scenario analysis approach (Yu et al., 2014) to estimate the potential interval of drought impacts on crop yields, so as to provide the best and the worst outcomes to decision makers. For future drought development, future climates are assumed to follow one of three scenarios:

(i) The drought ceases after the current day, and the water demand of crop growth is fully satisfied until harvest. The maximum yield to date (or the unrecoverable yield losses) can be derived in this scenario.
(ii) The drought continues in a user-specific period of time (e.g. 5, 10 or 20 days), but will return to the ideal condition afterwards. The extra yield loss in this given period of time can be derived from the modeling results.
(iii) The climate data after the current day will be replaced with representative climate in history (e.g. the historical wet-, medium- or dry-year data) so that users can compare the current drought impacts with that in historical series.

The crop models receive the latest real-time climatic observations at each time step and update the outputs for the scenario analysis. Though extreme events are likely to occur beyond the scope of historical record, the long historical series are still expected to cover most possible conditions. This scenario-based assessment provides comprehensive information about the potential trajectory of crop growth, based on which mitigation management can be made combined with the expert's judgment.

### 3.5. Risk analysis

We use a copula model of the dependence structure between spatial locations to compute the joint distribution of the relative production

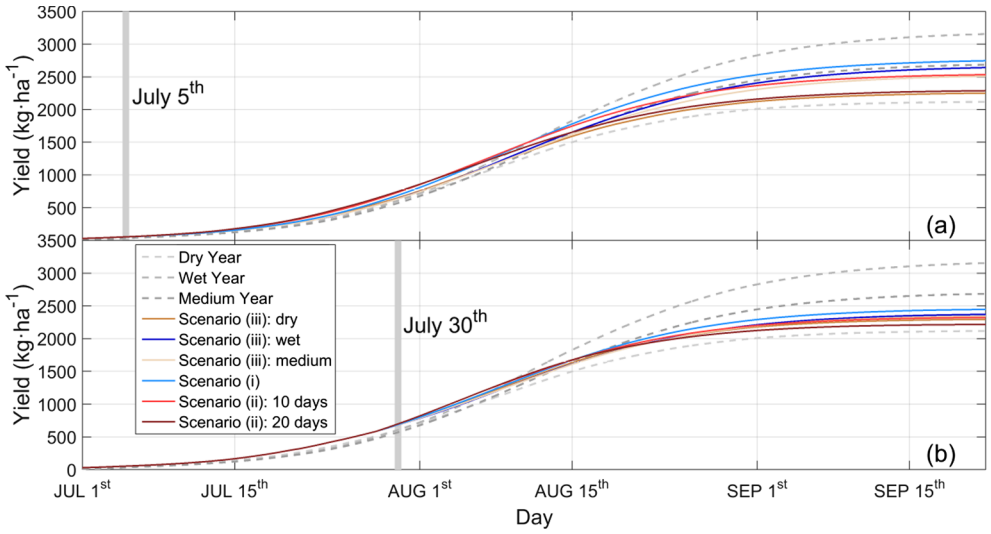


**Fig. 8.** The average yield predictions using different scenarios in this whole region: (a) real-time observation updated to July 5th; (b) real-time observation updated to July 30th.

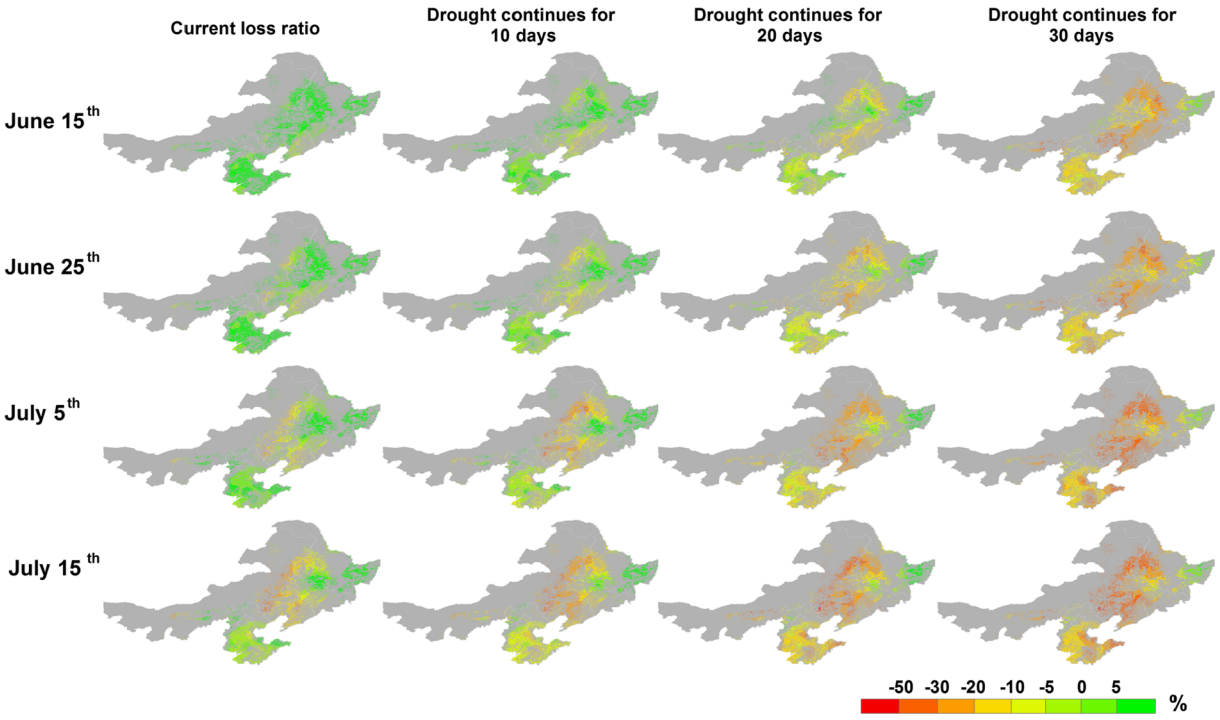


**Fig. 9.** Distributions of the expected maize yield losses (the modeled yield departure from the average county-level values of 2000–2016.

loss in multiple regions (Vedenov, 2008; Gaupp et al., 2017). Sklar's theorem (Sklar, 1959) states the joint distribution $F(x_1, \cdots, x_n)$ of $n$-dimension random variables $\{X_1, \cdots, X_n\}$ and the marginal distribution $F_i(x_i)$ for each individual random variable $X_i$ can be connected by the copula function as (Nelsen, 1999):

$$F(x_1, \cdots, x_n) = C[F_1(x_1), \cdots, F_n(x_n)] \tag{5}$$

where $C(\cdot)$ is the copula function with $C = [0, 1]^n \rightarrow [0, 1]$. When $F(\cdot)$ and $C(\cdot)$ are both differentiable, the joint PDF can be written as:

$$f(x_1, \cdots, x_n) = c[F_1(x_1), \cdots, F_n(x_n)] \cdot f_1(x_1) \cdots f_n(x_n) \tag{6}$$

where $c(\cdot)$ is the PDF of the copula function as:

$$c(\mu_1, \cdots, \mu_n) = \frac{\partial^n}{\partial \mu_1 \cdots \partial \mu_n} C(\mu_1, \cdots, \mu_n) \tag{7}$$

The advantage of using copula method is that users can only consider the marginal distribution of each random variable with fixed copula type. In our system, we employ a Gumbel copula method and pairwise ordered coupling as the copula structure (Gaupp et al., 2017). Users are able to choose corresponding predictions of yield loss in the regions of interest (e.g. multiple counties or provinces) for the analysis of joint distribution.

## 4. System testing and applications

### 4.1. Computational performance

To demonstrate the performance of model acceleration with SW26010 on this platform, we perform a 10-year simulation of maize yield in Changtu county with MPE-only and MPE-CPEs hybrid modes.

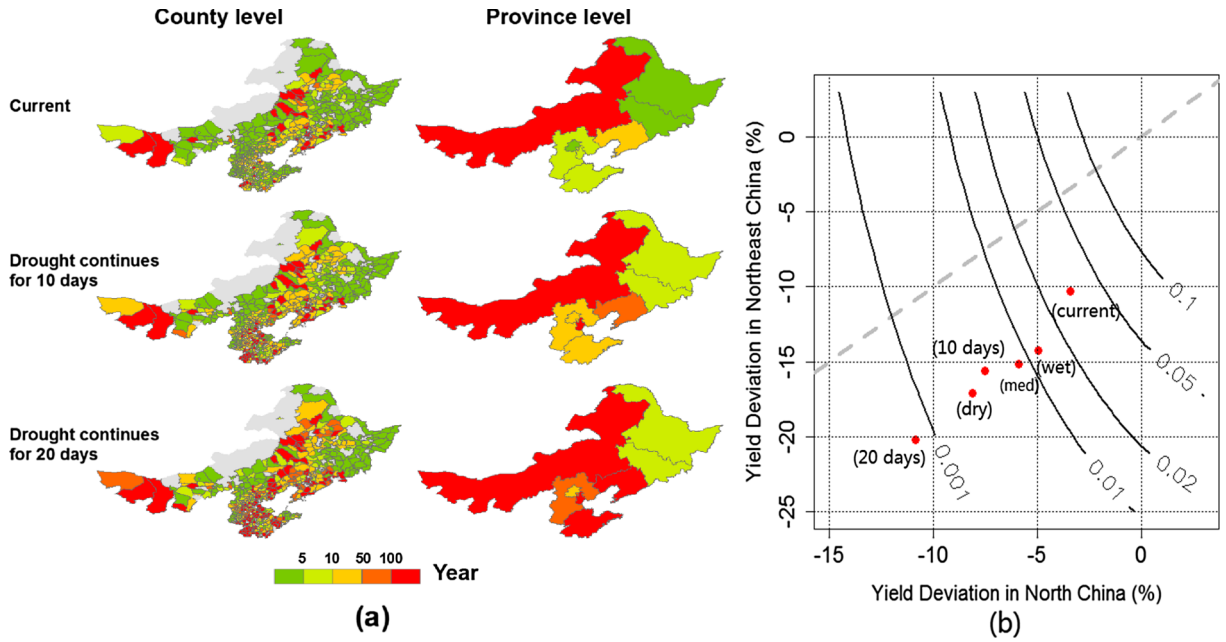


**Fig. 10.** (a) The return period of yield loss on July 30th in county and province levels; (b) The cumulative probability (contour line) of joint distribution of yield losses in both North and Northeast China.

The speedup ratio of CPE-based module is over 8, as shown in Table 2.

Then we randomly selected 512 counties in China for county-level modeling. For each model, the simulation period is 1996–2007, with corresponding run times for MPE-only and MPE-CPEs hybrid modes shown in Fig. 5, for CGs increasing from 8 to 512. In this experiment, we ran 10 iterations for each model and summated the run times to reduce the random error of run time measurement.

In Fig. 5, the run time for each trial decreases with the increase of CGs numbers. However, the actual accelerating efficiency is lower than the ideal speedup with the increasing CGs numbers, because the costs of both communication between different processes and I/O (input/output) processes will expanded rapidly when more CGs are used simultaneously. In general, by using CPEs, the total run time of DNDC model is accelerated by about 3.6 times while SWAP by 2.6 times. The overall performance is lower than the CPE-based part because of the investable part of MPE-based module, as well as the extra cost for the communication between MPE and CPEs (e.g. copying data from main memory to LDM). The difference in model algorithms (e.g. soil water simulation), the proportions of MPE-based module to the whole computing part and the detailed model setting (e.g. number of soil layers, convergence condition) is believed to contribute to the different efficiencies of speedup for these models. The overall reduction in computation time demonstrates the potential for near real-time predictions across large scales using the Sunway TaihuLight supercomputer resource.

### 4.2. Parameter optimization and multi-model ensemble for China

We simulate the annual yields of wheat, rice and maize for all 2403 counties in China. The soil properties (including clay content, hydraulic conductivity, maximum soil depth, etc.) are from the database "The Soil Database of China for Land Surface Modeling" (Shangguan et al., 2014). The climatic observation from weather station (http://www.cma.gov.cn/2011qxfw/2011qsjgx/), county-level yield records and averaged fertilizer amount from 1998 to 2010 statistical yearbooks (http://navi.cnki.net/KNavi/Yearbook.html) are used for parameter optimization and multi-model ensemble based on the methods in Section 3. Only parameters related to crop growth are selected for optimization and they are set in accordance with Yu's literature (Yu et al., 2018). The period 1998–2007 is used for calibration and 2008–2010 for validation.

In this case, 6000 CGs are applied collectively to generate 4000 samples for each crop in each county, with MCMC simulations for the three crop models taking approximately 18 h. We extract the predictions with maximum posterior PDF in each county and display the simulation performance in Fig. 6. And the parameter sets obtained here can be used for the real-time prediction afterwards.

Each individual model is shown to perform well at the national scale largely matching observations. However, for all the ensemble members, there are still significantly overestimated or underestimated results, especially in validation period. This implies great uncertainty of model structure for agricultural simulation and the limited skill of a single model for large-scale prediction. In comparison to individual models, the multi-model ensemble prediction in Fig. 6 shows greater accuracy in both calibration and validation periods with a higher coefficient-of-determination ($R^2$) value and lower root-mean-square error (RMSE) value. For example, the ensemble simulation of wheat is obviously better than any individual model. The over- and under-estimations of yield prediction decrease and the residual errors appears to have a normal distribution with smaller variance. These observations indicate that the Bayesian inference and BMA methods in this system effectively reduce the predictive uncertainty of crop growth on the basis of current crop models.

### 4.3. Dynamic drought impact assessment in North and Northeast China

Between April and July 2017, a severe drought occurred in North and Northeast China, including Jilin, Liaoning, Heilongjiang, Inner Mongolia, Shandong and Hebei provinces (Fig. 7). Combined, these regions produce over 60% of China's annual maize production, while any drought during this period is likely to cause significant decrease to the maize production later in the year. Collating daily climate data from June 1st, we monitored the current production losses and predicted the potential evolution of agricultural drought based on our three drought scenarios (Section 3.4).

In Fig. 8, the average yield for each scenario is predicted for July 5th and July 30th. On July 5th (Fig. 8(a)), given the study region has been in meteorological drought for over 2 months, the predictions in most scenarios are below the historic medium-level records. Under the

scenario where drought lasts less than 10 days, the final yield loss varies within 10% of the historical average. Theoretically, if future precipitation is sufficient, the final yield could even be slightly higher than the historic average. However, as maize growth just proceeds into the reproductive stage at this time of year, the uncertainty among these scenario predictions remains large due to the sensitivity of crop growth to subsequent weather conditions. On July 30th, the phenology of maize growth is closer to maturity than 25 days ago (July 5th), and the rate of photosynthetic assimilation typically decreases since this point. As a result, the range of yield of different scenarios in Fig. 8(b) converges to smaller range compared with Fig. 8(a). In the whole month in July, the drought condition still existed in many counties of this region, and it is inevitable that the average yield in 2017 will drop below the historical mean value, even with sufficient precipitation thereafter.

The distributions of up-to-date yield losses from June 15st to July 15th in about 10-day step are displayed in Fig. 9. The model simulations indicate the yield at any time before early June could still be recovered to the average level if the drought condition ends then. However, the sustained drought condition in late June and July was critical to the final grain formation, and the water stress in this period led to the irreversible damage to maize growth. This knowledge can be used by the local department of agriculture to advise mitigation actions to minimize the impacts of the drought. Fig. 9 further shows the agricultural loss in generally more severe in Northeast China compared to North China. We expect this is due to different meteorological conditions and irrigation levels.

### 4.4. The probability of yield losses

The marginal distributions of drought-induced yield loss can be estimated through the generalized extreme value (GEV) distribution and the joint distribution of regional yield losses by Gumbel copulas. The long-term yield simulations from 1960 to 2010 in each county are performed using the optimal model parameters in Section 4.2, based on which the related parameter values of probability distributions are estimated by maximum likelihood method. We calculate the return periods (the inverse of cumulative probability) for county and province levels in these regions in Fig. 10(a) and the joint cumulative probability for North and Northeast China in Fig. 10(b) based on the scenario-analysis results on July 30th.

Fig. 10(a) shows that the drought caused the most severe damage to agricultural production in Inner Mongolia, with the return periods over 50 and 100 years in many counties as well as the province level. The magnitude of yield loss seems a bit overestimated as the irrigation level in our database is actually lower than the current condition in this region. But generally this result is consistent with the meteorological drought in Fig. 7, because precipitation deficit in 2017 in this province shows a higher spatial and temporal constancy than other provinces. It is important to note that some provinces or counties with relatively low irrigation levels (e.g. Jilin and Heilongjiang provinces) have lower return period values than those with higher irrigation levels (e.g. Hebei and Shandong provinces). It is mainly because the lack of irrigation is more likely to cause the variability in historical yield simulation more frequent and severe, which makes the yield loss in 2017 less sensitive. The difference between yield loss and return period consequently implies the advantage of risk analysis for drought assessment in longer time series than then absolute value.

Fig. 10(b) clearly demonstrates the trajectory of joint probability in different scenarios. For decision makers concerned with food security, instead of treating those regions independently, we suggest using information on the joint condition of different production areas as the increase of production in some regions can compensate for the loss of others. In this case, for example, the marginal cumulative probabilities of current yield loss on July 30th are 0.1585 and 0.1393 for North and Northeast China respectively, while the joint cumulative probability is about 0.0311 (approx. 32-year return period). Compared with the joint distribution, the product (independent assumption) of the marginal probability 0.0221 (approx. 45-year return period) seems to overestimate the yield loss significantly. As the overall meteorological conditions for crop growth in North and Northeast China are similar due to geographic proximity, the spatial dependency of yield loss cannot be ignored for risk analysis, and our system provides an objective evaluation by Copula method.

## 5. Concluding remarks

In this study, we have developed an integrated agricultural risk system for daily agriculture dynamics on the Sunway TaihuLight supercomputer, including the parallelization and acceleration for existing crop models, model optimization and multi-model ensemble for yield prediction and multi-scale analysis of yield loss risk.

While traditional crop models are relatively computationally inefficient, through the detailed classification of model algorithms, we have observed an approximate 3× decrease in run times using the SW26010 processor. With over 40,000 processors available in our platform, we can conduct parameter uncertainty analyses and multi-model ensemble based on the MCMC method to improve the accuracy and precision of our model predictions. We have demonstrated the feasibility of using our approach for monitoring the effect of drought evolution on agricultural production across large scales through application to the May–July 2017 drought in North and Northeast China. The combination of yield losses and corresponding return periods at different scales is able to provide objective and useful evaluation of loss magnitude for the public and the local decision makers (e.g. local Departments of Agriculture) while the probability of joint distribution is helpful for higher-level decision makers such as the Ministry of Agriculture to coordinate mitigation or response actions in different regions.

To reduce the cost of I/O processes, which account for a considerable share of the total run time (especially when the number of processes increases), future improvements can be made by optimizing the format of I/O files (e.g. mergence of files in small size, binary format) of our crop models and the introduction of new techniques for the parallel I/O (e.g. PnetCDF) into our system. The quality and temporal frequency of input regional data during drought conditions is critical for the accuracy of the dynamic prediction. Such real-world information is not limited to replanting schedules, and changes to crop species, irrigation levels, and planting times. Potential techniques for collecting such information need to be explored, such as using online big data collection techniques to continually update the database. Data assimilation using satellite observations should also be adopted to improve predictive accuracy.

## Acknowledgement

This study was financially supported by the Chinese National Basic Research Program (2017YFA0603602). We thank Guangwen Yang, Haohuan Fu, Wei Xue, Zhao Liu, Junfeng Liao, Bin Yang and Lin Gan from National Supercomputing Center in Wuxi for their technical assistant to the modelling system. Thanks also go to the two anonymous reviewers for their constructive comments.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compag.2018.07.027.

## References

Adams, R.M., Hurd, B.H., Lenhart, S., Leary, N., 1998. Effects of global climate change on agriculture: an interpretative review. Clim. Res. 11, 19–30.
AghaKouchak, A., Bárdossy, A., Habib, E., 2010. Conditional simulation of remotely

sensed rainfall data using a non-Gaussian v-transformed copula. Adv. Water Resour. 33, 624–634.
Báez-González, A.D., Chen, P., Tiscareño-López, M., Srinivasan, R., 2002. Using satellite and field data with crop growth modeling to monitor and estimate corn yield in Mexico. Crop Sci. 42, 1943–1949.
Baigorria, G.A., Jones, J.W., O Brien, J.J., 2008. Potential predictability of crop yield using an ensemble climate forecast by a regional circulation model. AGR Forest Meteorol. 148, 1353–1361.
Bárdossy, A., Pegram, G., 2009. Copula based multisite model for daily precipitation simulation. Hydrol. Earth Syst. Sci. 13, 2299.
Bloom, D.E., 2011. 7 billion and counting. Science 333, 562–569.
Bonaccorso, B., Cancelliere, A., Rossi, G., 2003. An analytical formulation of return period of drought severity. Stoch. Environ. Res. Risk A 17, 157–174.
Box, G.E., Tiao, G.C., 2011. Bayesian Inference in Statistical Analysis. John Wiley & Sons.
De Wit, A.D., Van Diepen, C.A., 2007. Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts. AGR Forest Meteorol. 146, 38–56.
Dongarra, J.J., Luszczek, P., Petitet, A., 2003. The LINPACK benchmark: past, present and future. Concurr. Comput. Pract. Exp. 15, 803–820.
Dumont, B., Leemans, V., Mansouri, M., Bodson, B., Destain, J., Destain, M., 2014. Parameter identification of the STICS crop model, using an accelerated formal MCMC approach. Environ. Modell. Softw. 52, 121–135.
Elliott, J., Kelly, D., Chryssanthacopoulos, J., Glotter, M., Jhunjhnuwala, K., Best, N., Wilde, M., Foster, I., 2014. The parallel system for integrating impact models and sectors (pSIMS). Environ. Modell. Softw. 62, 509–516.
Ewert, F., Rötter, R.P., Bindi, M., Webber, H., Trnka, M., Kersebaum, K.C., Olesen, J.E., van Ittersum, M.K., Janssen, S., Rivington, M., 2015. Crop modelling for integrated assessment of risk to food production from climate change. Environ. Modell. Softw. 72, 287–303.
Fang, J., Fu, H., Zhao, W., Chen, B., Zheng, W., Yang, G., 2017. swDNN: a library for accelerating deep learning applications on Sunway TaihuLight. In: 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 615–624.
Fernández, B., Salas, J.D., 1999. Return period and risk of hydrologic events. I: mathematical formulation. J. Hydrol. Eng. 4, 297–307.
Field, C.B., 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
Fu, H., Liao, J., Ding, N., Duan, X., Gan, L., Liang, Y., Wang, X., Yang, J., Zheng, Y., Liu, W., 2017. Redesigning CAM-SE for peta-scale climate modeling performance and ultra-high resolution on Sunway TaihuLight. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, pp. 1.
Fu, H., Liao, J., Yang, J., Wang, L., Song, Z., Huang, X., Yang, C., Xue, W., Liu, F., Qiao, F., 2016. The Sunway TaihuLight supercomputer: system and applications. Sci. China Inform. Sci. 59, 72001.
Gaupp, F., Pflug, G., Hochrainer Stigler, S., Hall, J., Dadson, S., 2017. Dependency of crop production between global breadbaskets: a copula approach for the assessment of global and regional risk pools. Risk Anal. 37 (11), 2212–2228.
Guo, Y., Ma, Y., Zhan, Z., Li, B., Dingkuhn, M., Luquet, D., De Reffye, P., 2006. Parameter optimization and field validation of the functional–structural model GREENLAB for maize. Ann. Bot.-Lond. 97, 217–230.
Hansen, J.W., Challinor, A., Ines, A., Wheeler, T., Moron, V., 2006. Translating climate forecasts into agricultural terms: advances and challenges. Clim. Res. 33, 27–41.
Heng, L.K., Hsiao, T., Evett, S., Howell, T., Steduto, P., 2009. Validating the FAO AquaCrop model for irrigated and water deficient field maize. Agron. J. 101, 488–498.
Holzworth, D.P., Snow, V., Janssen, S., Athanasiadis, I.N., Donatelli, M., Hoogenboom, G., White, J.W., Thorburn, P., 2015. Agricultural production systems modelling and software: current status and future prospects. Environ. Modell. Softw. 72, 276–286.
Howden, S.M., Soussana, J., Tubiello, F.N., Chhetri, N., Dunlop, M., Meinke, H., 2007. Adapting agriculture to climate change. Proc. Nat. Acad. Sci. 104, 19691–19696.
Huang, J., Ma, H., Su, W., Zhang, X., Huang, Y., Fan, J., Wu, W., 2015. Jointly assimilating MODIS LAI and ET products into the SWAP model for winter wheat yield estimation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8 (8), 4060–4071. https://doi.org/10.1109/JSTARS.2015.2403135.. https://ieeexplore.ieee.org/document/7063257/.
Huang, X., Huang, G., Yu, C., Ni, S., Yu, L., 2017. A multiple crop model ensemble for improving broad-scale yield prediction using Bayesian model averaging. Field Crops Res. 211, 114–124.
Iizumi, T., Yokozawa, M., Nishimori, M., 2009. Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: application of a Bayesian approach. AGR Forest Meteorol. 149, 333–348.
Jakku, E., Thorburn, P.J., 2010. A conceptual framework for guiding the participatory development of agricultural decision support systems. Agric. Syst. 103, 675–682.
Jiang, L., Yang, C., Ao, Y., Yin, W., Ma, W., Sun, Q., Liu, F., Lin, R., Zhang, P., 2017. Towards highly efficient DGEMM on the emerging SW26010 many-core processor. In: 2017 46th International Conference on Parallel Processing (ICPP). IEEE, pp. 422–431.
Kroes, J.G., Van Dam, J.C., Groenendijk, P., Hendriks, R., Jacobs, C., 2009. SWAP version 3.2. Theory description and user manual. Alterra.
Kroes, J.G., Wesseling, J.G., Van Dam, J.C., 2000. Integrated modelling of the soil–water–atmosphere–plant system using the model SWAP 2·0 an overview of theory and an application. Hydrol. Process. 14, 1993–2002.
Launay, M., Guerif, M., 2005. Assimilating remote sensing data into a crop model to improve predictive performance for spatial applications. Agric. Ecosyst. Environ. 111, 321–339.
Lawrence, D.M., Oleson, K.W., Flanner, M.G., Thornton, P.E., Swenson, S.C., Lawrence, P.J., Zeng, X., Yang, Z.L., Levis, S., Sakaguchi, K., 2011. Parameterization improvements and functional and structural advances in version 4 of the Community Land Model. J. Adv. Model. Earth Syst. 3.
Li, B., Li, B., Qian, D., 2017. PFSI. sw: A programming framework for sea ice model algorithms based on Sunway many-core processor. In: 2017 IEEE 28th International Conference on Application-specific Systems, Architectures and Processors (ASAP). IEEE, pp. 119–126.
Li, C., Frolking, S., Frolking, T.A., 1992. A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity. J. Geophys. Res. Atmosp. 97, 9759–9776.
Lobell, D.B., Burke, M.B., Tebaldi, C., Mastrandrea, M.D., Falcon, W.P., Naylor, R.L., 2008. Prioritizing climate change adaptation needs for food security in 2030. Science 319, 607–610.
Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rötter, R.P., Boote, K.J., Ruane, A.C., Thorburn, P.J., Cammarano, D., 2015. Multimodel ensembles of wheat growth: many models are better than one. Global Change Biol. 21, 911–925.
Neale, R.B., Chen, C., Gettelman, A., Lauritzen, P.H., Park, S., Williamson, D.L., Conley, A. J., Garcia, R., Kinnison, D., Lamarque, J., 2010. Description of the NCAR community atmosphere model (CAM 5.0). NCAR Tech. Note NCAR/TN-486+ STR.
Nelsen, R.B., 1999. Introduction. An Introduction to Copulas. Springer, pp. 1–4.
Oleson, K.W., Lawrence, D.M., Gordon, B., Flanner, M.G., Kluzek, E., Peter, J., Levis, S., Swenson, S.C., Thornton, E., Feddema, J., 2010. Technical description of version 4.0 of the Community Land Model (CLM).
Qiao, F., Zhao, W., Yin, X., Huang, X., Liu, X., Shu, Q., Wang, G., Song, Z., Li, X., Liu, H., 2016. A highly effective global surface wave numerical simulation with ultra-high resolution. In: SC16: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, pp. 46–56.
Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155–1174.
Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A.C., Müller, C., Arneth, A., Boote, K.J., Folberth, C., Glotter, M., Khabarov, N., 2014. Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. Proc. Nat. Acad. Sci. 111, 3268–3273.
Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P., Antle, J.M., Nelson, G.C., Porter, C., Janssen, S., 2013. The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies. AGR Forest Meteorol. 170, 166–182.
Rosenzweig, C., Tubiello, F.N., 2007. Adaptation and mitigation strategies in agriculture: an analysis of potential synergies. Mitig. Adapt. Strat. Global 12, 855–873.
Shangguan, W., Dai, Y., Duan, Q., Liu, B., Yuan, H., 2014. A global soil data set for earth system modeling. J. Adv. Model. Earth Syst. 6, 249–263.
Skakun, S., Kussul, N., Shelestov, A., Kussul, O., 2016. The use of satellite data for agriculture drought risk quantification in Ukraine. Geomat. Nat. Hazards Risk 7, 901–917.
Sklar, M., 1959. Fonctions de repartition an dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8, 229–231.
Steduto, P., Hsiao, T.C., Raes, D., Fereres, E., 2009. AquaCrop—the FAO crop model to simulate yield response to water: I. Concepts and underlying principles. Agron. J. 101, 426–437.
Tao, F., Yokozawa, M., Zhang, Z., 2009. Modelling the impacts of weather and climate variability on crop productivity over a large area: a new process-based model development, optimization, and uncertainties analysis. AGR Forest Meteorol. 149, 831–850.
Van Ittersum, M.K., Ewert, F., Heckelei, T., Wery, J., Olsson, J.A., Andersen, E., Bezlepkina, I., Brouwer, F., Donatelli, M., Flichman, G., 2008. Integrated assessment of agricultural systems–a component-based framework for the European Union (SEAMLESS). Agric. Syst. 96, 150–165.
Vedenov, D., 2008. Application of copulas to estimation of joint crop yield distributions. In: American Agricultural Economics Association Annual Meeting, Orlando, FL, pp. 27–29.
Vital, J., Gaurut, M., Lardy, R., Viovy, N., Soussana, J., Bellocchi, G., Martin, R., 2013. High-performance computing for climate change impact studies with the Pasture Simulation model. Comput. Electron. Agric. 98, 131–135.
Vrugt, J.A., Ter Braak, C., Diks, C., Robinson, B.A., Hyman, J.M., Higdon, D., 2009. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int. J. Nonlin. Sci. Num. 10, 273–290.
Yu, C., Huang, X., Chen, H., Huang, G., Ni, S., Wright, J.S., Hall, J., Ciais, P., Zhang, J., Xiao, Y., Sun, Z., Wang, X., Yu, L., 2018. Assessing the Impacts of Extreme Agricultural Droughts in China Under Climate and Socioeconomic Changes. Earth's Future.
Yu, C., Li, C., Xin, Q., Chen, H., Zhang, J., Zhang, F., Li, X., Clinton, N., Huang, X., Yue, Y., 2014. Dynamic assessment of the impact of drought on agricultural yield and scale-dependent return periods over large geographic regions. Environ. Modell. Softw. 62, 454–464.
Zhao, G., Bryan, B.A., King, D., Luo, Z., Wang, E., Bende-Michl, U., Song, X., Yu, Q., 2013. Large-scale, high-resolution agricultural systems modeling using a hybrid approach combining grid computing and parallel processing. Environ. Modell. Softw. 41, 231–238.
Zhao, W., Fu, H., Fang, J., Zheng, W., Gan, L., Yang, G., 2018. Optimizing Convolutional Neural Networks on the Sunway TaihuLight Supercomputer. ACM Trans. Archit. Code Optim. 15, 1–26.