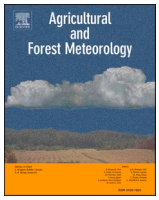
Agricultural and Forest Meteorology 355 (2024) 110137

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

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



Regional uncertainty analysis between crop phenology model structures and optimal parameters

Chenyao Yang [a,b,c,1], Na Lei [a,1], Christoph Menz [d], Andrej Ceglar [e], Jairo Arturo Torres-Matallana [f], Siqi Li [a], Yanling Jiang [a], Xianming Tan [a], Lei Tao [g], Fang He [g], Shigui Li [c,h], Bing Liu [i,*], Feng Yang [a,c,*], Helder Fraga [b], João A. Santos [b]

[a] College of Agronomy, Sichuan Agricultural University, Chengdu 611130, China
[b] Centre for the Research and Technology of Agro-Environmental and Biological Sciences (CITAB), Institute for Innovation, Capacity Building and Sustainability of Agri-food Production (Inov4Agro), Universidade de Trás-os-Montes e Alto Douro (UTAD), Vila Real 5000-801, Portugal
[c] Key Laboratory of Agricultural Bioinformatics, Ministry of Education, Sichuan Agricultural University, China
[d] Potsdam Institute for Climate Impact Research e. V. (PIK), Telegrafenberg A 31, Potsdam 14473, Germany
[e] Climate Change Centre of the European Central Bank, Sonnemannstrasse 20, Frankfurt am Main 60314, Germany
[f] Climate Service Center Germany (GERICS), Helmholtz-Zentrum Hereon, Fischertwiete 1, Hamburg 20095, Germany
[g] Sichuan Seed Station, No.5 Yulin North Road, Wuhou District, Chengdu 610044, China
[h] Rice Research Institute, Sichuan Agricultural University, Chengdu 611130, China
[i] Key Laboratory for Crop System Analysis and Decision Making, Jiangsu Key Laboratory for Information Agriculture, Ministry of Education, Ministry of Agriculture, National Engineering and Technology Center for Information Agriculture, Engineering Research Center of Smart Agriculture, Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, Nanjing, China

A R T I C L E I N F O

Keywords:
Crop Phenology Model
Ensemble Simulation
Parameter Uncertainty
Model Variability
Cross Validation
Regional Modelling

A B S T R A C T

Crop phenology models are pivotal for simulating crop development, predicting yields and guiding agricultural practices. However, uncertainties exist in simulations due to different model structures and variability in model parameters. Although quantifying these contributions to total variability is often conducted at a site-specific level, few attempts to address this for regional crop modelling using field-calibrated parameters. Our study employs six crop phenology models (APSIM, CERES, GDD, Richardson, Sigmoid and Wang) for simulating maturity timings of three representative rice cultivars using trial data within the Sichuan Basin, China. The Leave-One-Out Cross-Validation (LOOCV) is applied for model calibration with a global parameter optimization algorithm and evaluation. Calibrated models show robust prediction capabilities during LOOCV with R^2 of 0.68–0.95 and RMSE of 2–4 days, though a larger variance is found for evaluation data than for calibration data. Models calibrated with data from sites having frequent high-temperature ($T_{max} \geq 32$ °C) episodes tend to have better predictability than without high-temperature episodes. Parameter variability, calibrated with different subsets of each cultivar during LOOCV, is low-to-moderate (mostly CV≤20 %) except for the Sigmoid model's curve steepness parameter. For the early-maturity cultivar, parameter variability is spatially the main uncertainty factor, relating to its greater variability of site-specific calibrated parameter values. For the medium-maturity and late-maturity cultivars, the dominant uncertainty source arises from the interplay between model structures and parameters. Parameter variability notably influences the overall uncertainty more than the model structure variability across the region, except in areas prone to high-temperature extremes where divergent model responses predominate. These findings highlight the cultivar-specific nature of simulation uncertainty, but also the critical need to assess the spatial distribution of uncertainty sources. For parameter uncertainty, a broader conceptualization is essential for more accurate quantifications of uncertainty sources, paving the way for improved ensemble crop modelling, especially at a large spatial scale.

* Corresponding authors.
E-mail addresses: bingliu@njau.edu.cn (B. Liu), f.yang@sicau.edu.cn (F. Yang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.agrformet.2024.110137
Received 6 March 2024; Received in revised form 20 May 2024; Accepted 24 June 2024
Available online 2 July 2024

1. Introduction

Rice (*Oryza sativa* L.) is a staple food crop for more than half of the world's population (Prasad et al., 2017). Accurate prediction of the timing of rice maturity is crucial as it is a key phenology stage to determine the yield and its quality. Timely and accurately prediction of rice maturity can aid farmers in adapting and optimizing harvest management, such as properly organizing manpower, hence contributing to increased production efficiency and profitability. However, with ongoing climate change, it becomes increasingly complex, due to the difficulties to adequately describe crop development facing stronger climatic variability and more frequent weather extremes (IPCC, 2022). Thus, understanding the key drivers influencing maturity predictions under varying climatic conditions is essential.

Crop phenology models are valuable tools for simulating and forecasting crop development stages from sowing to maturity (Wallach et al., 2017; Zhang et al., 2017; Gao et al., 2020; Kawakita et al., 2020). These models describe crop development mainly as a function of temperature, photoperiod and vernalization effects (Gao et al., 1992; Bouman et al., 2001; Jones et al., 2003; Brisson et al., 2009; Liu et al., 2019). There are also other factors that can affect crop development, such as drought stress, but this is rarely considered (Brisson et al., 2009). Despite the critical role of temperature in driving phenology development, there is a considerable variability in the temperature response functions among phenology models, leading to uncertainties in model structures when simulating a target phenology stage (Zhang et al., 2017; Zhang and Tao, 2019; Kawakita et al., 2020; Yang et al., 2023b). Some models adopt the linear temperature accumulation function, such as the Growing Degree Days (GDD) approach (Arlo Richardson et al., 1974; Bonhomme, 2000). Other models express the temperature effects by exponential (Hänninen, 1990) or curvilinear functions (Wang and Engel, 1998). Others, such as CERES (Ritchie and Otter, 1985; Ritchie et al., 1998) and ORYZA2000 (Bouman et al., 2001), incorporate an hourly temperature interpolation scheme when calculating the phasic development (Wallach et al., 2017; Ceglar et al., 2019). The choice of the appropriate model often depends on the purpose of the study, the target region, the varieties simulated and available computational resources. To address the model structural variability, multi-model ensembles have been frequently adopted, to gain insights for reducing variability in simulations between models and improve the reliability of predictions (Wallach et al., 2021a, 2021b; Zheng and Zhang, 2023). The multi-model ensemble median or mean often outperforms individual models or achieves comparable accuracy to the best single model (Rötter et al., 2018; Wallach et al., 2021b).

Parameter uncertainty in crop model simulations is another critical aspect. Parameter uncertainty can be defined as the uncertainty in calibrated parameter values that could not be able to minimize the errors between observations and predictions (Seidel et al., 2018). This can be partly due to the lack of an established calibration methodology with detailed and standard procedures to guide model users (Wallach et al., 2021c). Because different choices of parameter optimization algorithms and objective functions, as well as the assumption for parameter distribution, can all lead to a considerable variability in obtained parameter values (Gao et al., 2020; Wallach et al., 2021c, 2023a). In particular, it is essential to have a robust and effective optimization algorithm capable of finding the optimal parameter values given the observed data for calibration (Seidel et al., 2018). Additionally, insufficient data or limited representativeness of calibration data for a target population can also contribute significantly to the parameter uncertainty (Kersebaum et al., 2015; Montesino-San Martin et al., 2018; Seidel et al., 2018; Wallach et al., 2021a). Calibration typically relies on limited sample data and uncertainties exist on whether the model fitting to calibration data (goodness-of-fit) is indicative of the model's predictive performance in new and different situations (out-of-sample predictions) (Wallach et al., 2017, 2021a, 2021b, 2021c). Ideally, while assuming the same optimization algorithm or calibration approach in general, the estimated model parameters and prediction performance should be consistent or marginally different when using different data subsets of the same cultivar from multiple site × year combinations for calibration. However, this is often not the case. Again, this can be potentially due to limited representativeness of data subsets, as the optimal site × year combinations of a given cultivar is situation-dependent. Besides, the fact that no model can exhaustively take into account all explanatory variables affecting crop development and growth, can be another cause (Montesino-San Martin et al., 2018; Seidel et al., 2018; Kawakita et al., 2020). Indeed, parameter uncertainty is intricately linked with the model structural uncertainty. To address the parameter uncertainty resulting from the uncertainty in cultivar-specific observations for calibration, the cross-validation technique can be employed, as it allows iterative trainings and evaluations of models across varied data subsets, thereby quantifying parameter and associated prediction uncertainties in relation to different observation subsets for a given crop cultivar.

Identifying the dominant source of uncertainty in model simulations, whether due to model structure or parameter, is a critical research direction within the crop modeling community (Asseng et al., 2013; Wallach et al., 2017; Zhang et al., 2017; Tao et al., 2018; Kawakita et al., 2020). This identification is key to directing model improvements and reducing simulation uncertainties. It can be achieved through experiments to improve model equations to reduce the structural uncertainty or more effective use of existing data (and more and better data) in the calibration procedure to minimize parameter uncertainty (Seidel et al., 2018). Such advancements could lead to more accurate and robust simulation outputs. In rice phenology simulations, model structural uncertainties often outweigh parameter uncertainties (Wallach et al., 2017; Zhang et al., 2017). However, these studies' definition of parameter uncertainty does not consider the potential variability in parameter values due to different data subsets of a given cultivar used for calibration. Furthermore, previous studies primarily focused on decomposing total simulation variance at the site-specific scale (Asseng et al., 2013; Wallach et al., 2017, 2021b; Zhang et al., 2017; Tao et al., 2018; Kawakita et al., 2020; Wang et al., 2020), with few attempts to identify the main uncertainty factor affecting regional-scale simulations. Identifying the dominant source of uncertainty at regional-scale is crucial, as it is a fundamental step towards more accurate regional crop modelling, which then can lead to improved decision-makings and agricultural policy formulations for a specific production region.

In this study, six phenology models of varying complexity in temperature response functions are calibrated and evaluated for simulations of the maturity timings of three representative rice cultivars in the Sichuan Basin, southwest China. Observed data from multiple sites over 2017–2022 was measured and collected for each cultivar to capture their spatial-temporal distribution within the region. A cross-validation technique, coupled with a global parameter optimization algorithm, was adopted in a multi-model framework. Our objectives are twofold: 1) to estimate parameter uncertainties and evaluate the predictive performance of calibrated models across different sites for each cultivar; 2) to identify the dominant source of uncertainty in simulating the maturity timings of representative cultivars for main rice production area of Sichuan Basin, based on the ensemble model simulations with obtained site-specific parameterizations.

2. Material and methods

2.1. Study region and data collection

2.1.1. Characteristics of the study region

The study was conducted within the Sichuan Basin, located in southwest China (Fig. 1). The region has a subtropical humid climate with an annual mean temperature between 14 and 19 °C. The cumulative daily mean temperature (≥10 °C) throughout the year ranges from 4200 °C to 6100 °C (Shao et al., 2012). Due to its favorable agro-climatic conditions and abundant resources, the Sichuan Basin has been a

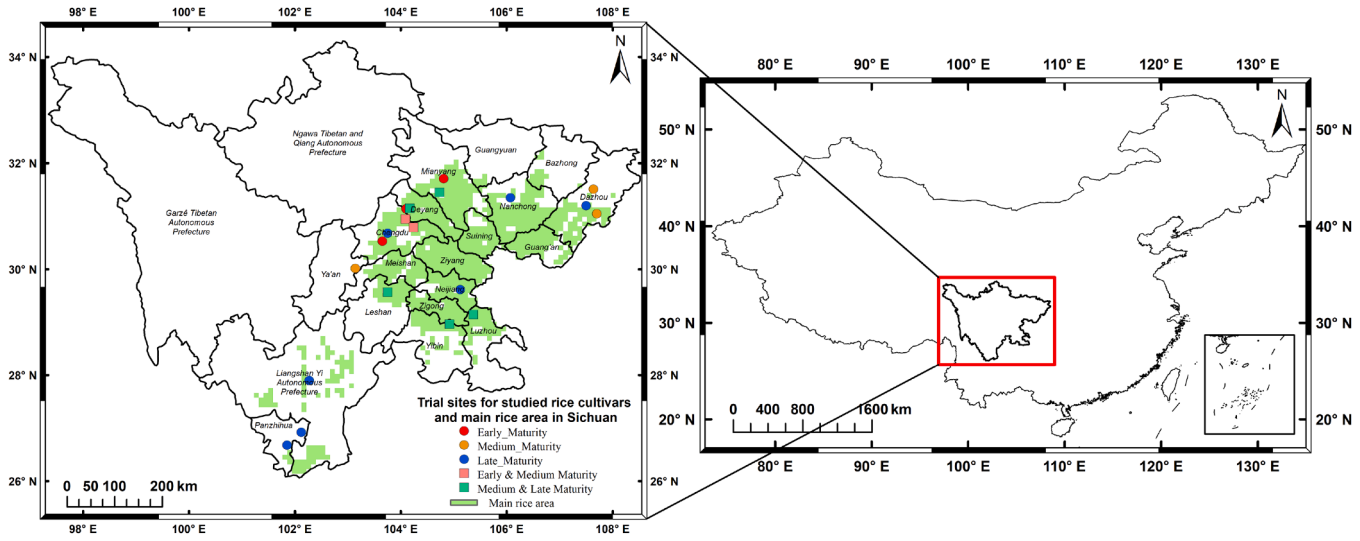


Fig. 1. Geographic distribution of regional trial sites for CK (Control-Check) rice cultivars from early-maturity type (ChuanZuoYou8727), medium-maturity type (FuYou838) and late-maturity type (YiXiangYou2115) during 2017—2022 in Sichuan province (main rice production area is also shown). The same site for both early- and medium-maturity types or both medium- and late-maturity types is denoted. Note the individual grid cell of main rice area corresponds to that with ≥25 % of land use for growing rice crop.

significant rice production area in China for a long time. The spatial distribution of major paddy rice areas has been identified (Fig. 1), which was derived from high-resolution (10 m) rice maps during 2017–2022 (Shen et al., 2023).

2.1.2. *Cultivar data*

This study incorporated data from three Control-Check (CK) reference rice cultivars, selected from regional rice cultivar trials of Sichuan province. These included ChuanZuoYou8727, FuYou838 and YiXiangYou2115, representing early-maturity, medium-maturity and late-maturity cultivar type, respectively. Among them, YiXiangYou2115 had extensive cultivations, covering more than a million hectares. The geographic distribution of trial sites for each maturity type is depicted in Fig. 1. These trials (with three replicates), conducted from 2017 to 2022, encompassed locations with a diverse range of environmental conditions. Table 1 provides detailed information on the number of trial sites per maturity type, site names (20 unique sites), geographic coordinates, altitudes and the availability of site-specific data. In general, the rather short time span of the data was compensated by a high number of trial sites, hence the total amount of data was comparable to those for other crop phenology modelling studies (Wallach et al., 2017, 2023b; Gao et al., 2020; Kawakita et al., 2020).

For each trial site, we measured and collected data on heading and maturity stages. The heading stage was defined as the date when 50 % of panicles emerged from the boot, while the physiological maturity was defined as the date on which grains attain their maximum dry weight

Table 1
Study sites for associated rice cultivar maturity types

Sites number	Sites name	Longitude (E)	Latitude (N)	Altitude (m)	Data available years	Rice cultivar maturity type (cultivar name)
site1	chongzhou1	103.65	30.53	504	2017–2022	Early-maturity type (ChuanZuoYou8727)
site2	jiangyou	104.81	31.71	586	2017–2022	
site3	pengzhou	104.09	30.95	515	2017–2019, 2021	
site4	shifang	104.10	31.14	550	2017–2020	
site5	xindu	104.24	30.79	491	2017–2022	
site1	dachuan2	107.70	31.05	366	2018–2022	Medium-maturity type (FuYou838)
site2	leshan	103.75	29.57	446	2017–2022	
site3	luxian	105.37	29.15	353	2017–2022	
site4	mianyang	104.73	31.45	530	2017,2018,2021,2022	
site5	pengzhou	104.09	30.95	515	2017–2019, 2021, 2022	
site6	shuangliu	104.17	31.15	506	2017–2022	
site7	yaan	103.14	30.02	557	2017–2020	
site8	yibin	104.92	28.97	375	2017–2022	
site9	xindu	104.24	30.79	491	2017–2022	
site10	xuanhan	107.64	31.51	392	2017–2022	
site1	dachuan1	107.50	31.20	288	2017–2022	Late-maturity type (YiXiangYou2115)
site2	leshan	103.75	29.57	446	2017–2022	
site3	luzhou(luxian)	105.37	29.15	295	2017–2022	
site4	mianyang	104.73	31.45	487	2017–2022	
site5	nanchong	106.07	31.35	360	2017–2022	
site6	neijiang	105.12	29.62	331	2017–2022	
site7	shuangliu	104.17	31.15	529	2017–2019,2021,2022	
site8	chongzhou2	103.75	30.68	540	2017–2022	
site9	yibin	104.92	28.97	368	2017–2022	
site10	miyi	102.12	26.92	1224	2017–2019	
site11	xichang	102.27	27.90	1574	2017–2019	
site12	yanbian	101.85	26.68	1165	2018,2019	

over the course of development (Shi et al., 2015). The empirical cumulative distribution of these stages, along with the computed GDD for the heading-maturity phase, distinctively highlighted the differences among the studied rice maturity types, as illustrated in **Fig. S1**. The heading-maturity GDD for early-, medium-, and late-maturity cultivars ranged approximately between 430–600 °C d^{-1}, 420–690 °C d^{-1} and 450–720 °C d^{-1}, respectively (**Fig. S1**). The accumulated GDD related to the regional characteristics of the sowing date: the early-maturity cultivar was typically sown later (around April 20th), compared to the medium-maturity (early April) and late-maturity type (around March 20th). This practice attempted to align sowing times with the optimal growing conditions of each season.

2.1.3. Weather data

During the trial period 2017–2022, daily observed minimum (T_{min}), mean (T_{mean}) and maximum (T_{max}) temperatures at the study sites were primarily retrieved from the ERA5-Land hourly reanalysis dataset. Hourly values were then aggregated into daily values. The ERA5-Land,

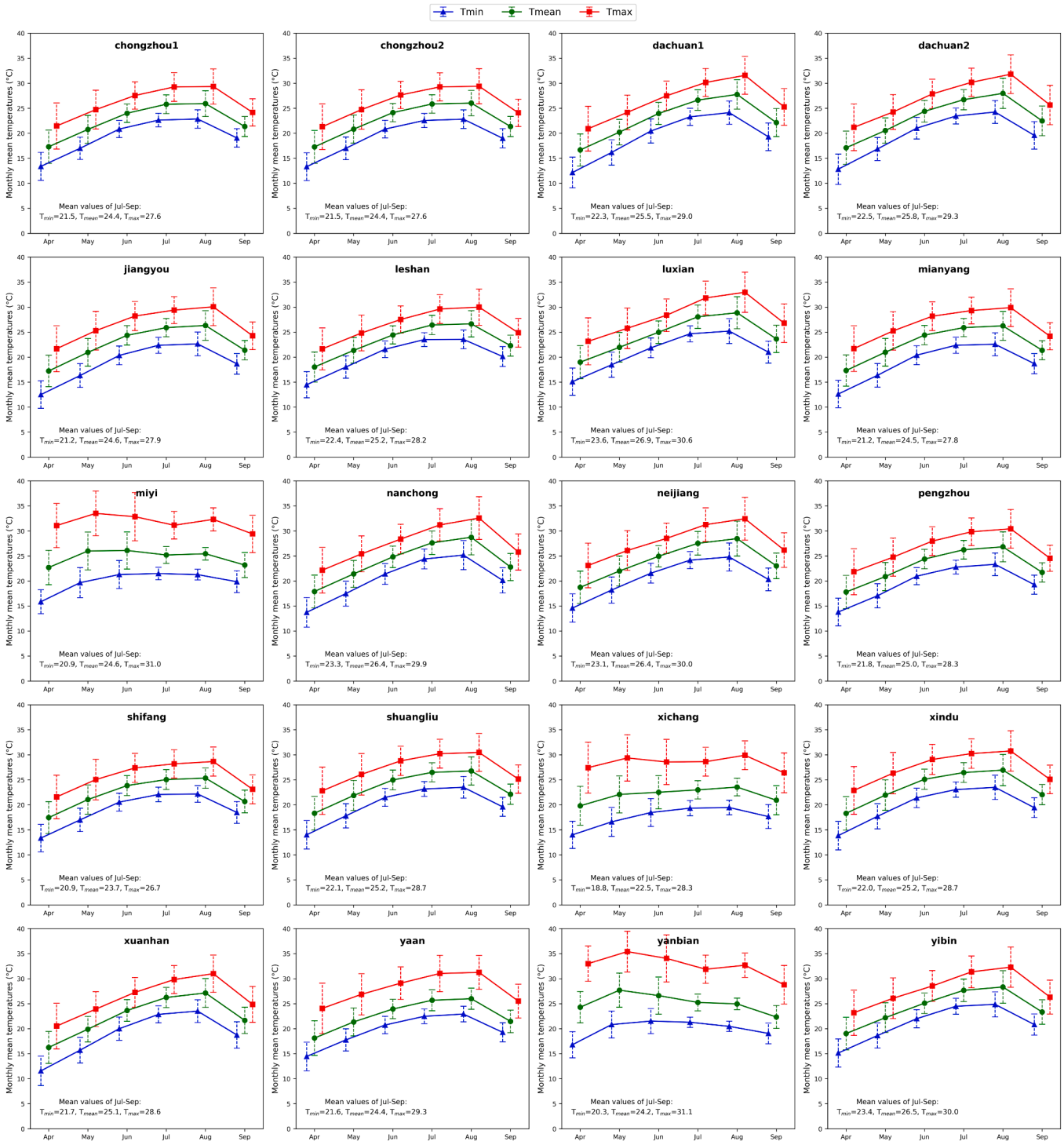


Fig. 2. Monthly average of daily minimum (T_{min}), mean (T_{mean}) and maximum (T_{max}) temperature during the rice growing season (April to September) for study sites over 2017—2022. Error bars indicate standard deviations over years. Mean temperature from July to September is also shown, within which the heading-maturity phase occurs.

integrating model data with observations globally, provides a gridded dataset at a resolution of 0.1° × 0.1° (~9 km) from 1950 to the present, detailing key variables of the water and energy cycles over land surfaces (Muñoz-Sabater et al., 2021). Despite the general accuracy of ERA5-Land dataset, its reliability at the local site scale is critical to assess. Therefore, for 20 study sites, daily T_{min}, T_{mean}, and T_{max} data from ERA5-Land were compared with the nearest (nearest to each study site with the maximum distance<20 km) weather station data from the China Meteorological Administration (CMA) over 1985—2014 (note CMA data was only accessible until 2015 for most of these weather stations). The comparison indicated that ERA5-Land data could accurately reflect the reference weather station data, evidenced by consistently high R^2 values above 0.98 (**Fig. S2**). Notable discrepancies were only discovered in T_{max} for a few southern sites, i.e., miyi, xichang, yaan and yanbian (**Fig. S2**). Consequently, CMA weather station data from 2017 to 2022 was utilized for these four sites.

During the months of July to September, which typically encompasses the heading-maturity phase, the mean T_{min}, T_{mean} and T_{max} for 2017–2022 across trial sites of the early-maturity rice type was 20.9–22.0 °C, 23.7–25.2 °C and 26.7–28.7 °C, respectively (Fig. 2). For the medium-maturity type, the corresponding ranges were 21.2–23.6 °C, 24.4–26.9 °C and 27.8–30.6 °C (Fig. 2). The late-maturity type showed ranges of 18.8–23.6 °C, 22.5–26.9 °C and 27.6–31.1 °C, respectively (Fig. 2). The mean number of Hot Days (HD) during July-September, defined as days with $T_{max}\geq 32$ °C that could adversely impact rice development (Zhang and Tao, 2013), varied roughly between 15–20 days, 14–36 days and 12–48 days across associated sites for early-, medium- and late-maturity type, respectively (**Fig. S3**). Sites with a relatively high frequency of HD (≥30 days) were predominantly under the late-maturity (6 sites), followed by the medium-maturity (3 sites), with no such sites found for the early-maturity (**Fig. S3**). Furthermore, a significant warming trend was observed at all trial sites, where the annual mean T_{min}, T_{mean} and T_{max} respectively showed a 10-year increment of 0.2–0.39 °C, 0.23–0.56 °C and 0.36–0.74 °C, clearly evidencing the climate warming (**Fig. S4**). With such a warming trend, a higher HD can be expected, which can further impose negative impacts on crop development and growth, particularly during the heading-maturity phase. Hence, it is necessary to examine how the current crop phenology models perform at sites with different HD.

2.2. Description of phenology models

In our study, the phenology stage was determined based on the accumulation of thermal forcing temperatures from a specified onset date (t_0) to a target date (t_s), where the state of thermal forcing (S_f) reaches a specific threshold value (F^*), typically cultivar-dependent:

$$S_f(t_s) = \sum_{t_0}^{t_s} R_f(x_t) \geq F^* \quad (1)$$

where x_t represented either daily mean temperature or hourly temperature, depending on the model. Since this study focused on the heading-maturity phase, t_0, t_s corresponded to the heading and maturity stage respectively. Besides, the post-heading photoperiod effects were considered negligible, and consequently, only temperature response functions specific to the phenology models were utilized. Six models were applied and categorized into three groups based on their structural attributes: (1) GDD and GDD-Richardson, with a relatively simple and near-linear temperature response function; (2) Sigmoid and Wang, with a non-linear structure; and (3) APSIM and CERES, with a non-linear structure and hourly temperature interpolation scheme.

2.2.1. GDD and GDD-Richardson

Both GDD and GDD-Richardson (Richardson thereafter) models adopt the GDD concept. The temperature response function for GDD has a single parameter ($T_{\_base}$) (Table 2), assuming a linear accumulation of daily effective thermal forcing when $x_t > T_{\_base}$ (Bonhomme, 2000):

$$R_f(x_t) = Max(x_t - T_{\_base}, 0) \quad (2)$$

The GDD-Richardson is a slightly modified version of the GDD model, which additionally considers a high-temperature threshold ($T_{\_max}$) (Table 2), above which the daily development rate is constant (Arlo Richardson et al., 1974):

$$R_f(x_t) = Max(Min(x_t - T_{\_base}, T_{\_max} - T_{\_base}), 0) \quad (3)$$

2.2.2. Sigmoid and Wang

The Sigmoid model expresses the temperature response as a logistic function with two parameters (d, e) (Table 2) (Hänninen, 1990):

$$R_f(x_t) = 1 \,/\, (1 + exp(d * (x_t - e))) \quad (4)$$

Table 2
Applied phenology models and associated parameters for calibrations. The " × " symbol indicates the calibration parameter of a given model.

Parameter Abbreviation	Description	Unit	Lower bound	Upper bound	Model Name					
					APSIM	CERES	GDD	Richardson	Sigmoid	Wang
F^*	Critical state of thermal temperature	degree days^{-1}* or temperature ratios *	0	1500* or 150*			×	×	×	×
DVRR	Threshold parameter of development rate to determine the cumulative thermal time necessary to accomplish heading-maturity phase	degree days^{-1}	0.0005	0.005	×					
P5	Threshold parameter to determine the cumulative thermal time to accomplish heading-maturity phase	degree days^{-1}	0	2000		×				
$T_{\_base}$	Minimum development temperature (base temperature)	°C	5	15	×	×	×	×		×
$T_{\_opt}$	Optimum development temperature	°C	15	35	×	×				×
$T_{\_max}$	Maximum development temperature	°C	35	45				×		×
$T_{\_lim}$	Limiting temperature above which development rate is null	°C	35	45	×					
d	Fitted parameter of sharpness of response curve	/	−20	0					×	
e	Fitted parameter of mid-response temperature	°C	0	30					×	

* GDD and Richardson compute F^* with degree days^{-1}, while the other models compute F^* with temperature ratios (between 0 and 1). Accordingly, the search boundary for F^* is 0–1500 for GDD and Richardson and 0–150 for other models.

The Wang model (named as Wang-Engel model, Wang model for short) follows a curvilinear temperature response function with three cardinal temperatures ($T_{\_base}$, $T_{\_opt}$, $T_{\_max}$) (Table 2) to define the shape of the response curve (Wang and Engel, 1998):

$$R_f(x_t)=\begin{cases}\dfrac{2(x_t-T_{\_base})^{\alpha}(T_{\_opt}-T_{\_base})^{\alpha}-(x_t-T_{\_base})^{2\alpha}}{(T_{\_opt}-T_{\_base})^{2\alpha}} & T_{\_base}\le x_t\le T_{\_\max}\\ 0 & x_t<T_{\_base}\ or\ x_t>T_{\_\max}\end{cases}$$
$$\text{with}$$
$$\alpha=\frac{\ln 2}{\ln\left(\dfrac{T_{\_\max}-T_{\_base}}{T_{\_opt}-T_{\_base}}\right)} \tag{5}$$

2.2.3. APSIM and CERES

The APSIM-Oryza model, hereafter referred to as APSIM, is a process-based rice model for simulating crop development, growth and yield formation in different environmental conditions (Bouman et al., 2001; Liu et al., 2019). Here we are only interested in its phenology routines when simulating the heading-maturity phase. Similar to the Wang model, it considers three cardinal temperatures ($T_{\_base}$, $T_{\_opt}$, $T_{\_lim}$) (Table 2) to compute degree days (DD), but it is based on hourly temperature (x_{ti}) using a sinusoidal interpolation between the daily minimum (x_{tmin}) and maximum (x_{tmax}) temperatures:

$$x_{ti}=\frac{(x_{tmin}+x_{tmax})}{2}+\frac{(x_{tmax}-x_{tmin})\cos(0.2618\times(i-14))}{2}$$
$$\text{for } i=1,\ 2,\ 3\cdots,\ 24 \tag{6}$$

and then,

$$DD_i=\begin{cases}x_{ti}-T_{\_base} & T_{\_base}\le x_{ti}\le T_{\_opt}\\ (T_{\_opt}-T_{\_base})-(x_{ti}-T_{\_opt})\dfrac{(T_{\_opt}-T_{\_base})}{(T_{\_\lim}-T_{\_opt})} & T_{\_opt}<x_{ti}<T_{\_\lim}\\ 0 & x_{ti}<T_{\_base}\ or\ x_{ti}\ge T_{\_\lim}\end{cases}$$
$$R_f(x_t=x_{ti})=\sum_{i=1}^{24}DD_i \tag{7}$$

In addition, we have adopted the approach given by Wallach et al., (2017) to convert the fractional development sum into the cumulative DD value required to accomplish the heading-maturity phase, leading to $F^*=1/DVRR$ (Table 2).

The CERES-Rice (hereafter CERES) model in DSSAT (Jones et al., 2003) was developed by Singh et al., (1993) for simulating both lowland and upland rice development and growth. Crop development was simulated with a thermal time accumulation approach. When the plant has more than 10 leaves, the model uses a daily minimum (x_{tmin}) and maximum (x_{tmax}) temperature to compute the phasic development with two cardinal temperatures ($T_{\_base}$, and $T_{\_opt}$) (Table 2) (Singh et al., 1993).

$$R_f(x_t=x_{ti})=\begin{cases}\dfrac{(x_{tmin}+x_{tmax})}{2}-T_{\_base} & T_{\_base}<x_{tmin}\ and\ x_{tmax}<T_{\_opt}\\ \left(\dfrac{1}{24}\right)\sum_{i=1}^{24}(x_{ti}-T_{\_base}) & x_{tmin}\le T_{\_base}\ or\ x_{tmax}\ge T_{\_opt}\end{cases} \tag{8}$$

Similar to APSIM, hourly temperature (x_{ti}) could be estimated from x_{tmin} and x_{tmax}, but with a different interpolation method:

$$x_{ti}=\begin{cases}\dfrac{(x_{tmin}+x_{tmax})}{2}+\dfrac{(x_{tmax}-x_{tmin})\sin\left(\dfrac{\pi\times i}{12}\right)}{2} & T_{\_base}\le x_{ti}\le T_{\_opt}\\ x_{ti}=T_{\_base} & x_{ti}<T_{\_base}\\ x_{ti}=T_{\_opt}-(x_{ti}-T_{\_opt}) & x_{ti}>T_{\_opt}\end{cases}$$
$$\text{for } i=1,2,3\cdots.24 \tag{9}$$

where phenology development can be somewhat delayed if the x_{ti} exceeds $T_{\_opt}$ (Table 2) (Boote et al., 2005). For CERES, the last two phases from heading to grain-filling and grain-filling to maturity were combined into the heading to the maturity phase, following Wallach et al., (2017). Accordingly, the fractional development sum was converted into the cumulative DD value, leading to $F^*=170+P5$ (Table 2) (Wallach et al., 2017).

The phenology models were implemented in Python according to Equations 1–9 and can be found in the Zenodo repository (https://zenodo.org/record/7875044).

2.3. Global optimization algorithm

Model calibration is a crucial process in adapting models to local conditions by estimating model-specific parameters. The goal is to find parameters that could minimize prediction errors, a task typically achieved using an optimization algorithm (Seidel et al., 2018; Wallach et al., 2021c). The Shuffled Complex Evolution, developed at the University of Arizona (SCE-UA) has been employed for estimating the optimal parameter values in our study. It is an established, robust and efficient optimization algorithm (Duan et al., 1993, 1994), which has been widely applied in both hydrological (Duan et al., 1994; Muttil and Jayawardena, 2008) and crop models, including the crop phenology models (Zhang et al., 2017; Jin et al., 2022; Yang et al., 2023b). As indicated by Duan et al., (1994), SCE-UA's strengths include: (1) achieve global convergence despite the presence of multiple local optimums; (2) avoid trapping in suboptimal regions of the objective function surface; (3) remain effective across varying parameter sensitivities and interdependencies; (4) operate independently of explicit expressions for the objective function or its derivatives; (5) handle efficiently high-dimensional parameter spaces.

For parameter optimization, a uniform distribution was assumed for each parameter with equal sampling probability within predefined bounds (Table 2). These predefined ranges were generally set broader than those in our previous study, ensuring that the initial boundary settings had little or no influence on the final estimated parameter values (Yang et al., 2023b). The selected objective function for SCE-UA was the minimization of the Root Mean Squared Error (RMSE) between the observed and simulated values, which provided a quantitative measure of the agreement between the model and the actual data.

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(Simulation_i-Observation_i)^2}{n}} \tag{10}$$

where n is the number of total observations included. For SCE-UA implementation, we utilized the open-source package Statistical Parameter Optimization Tool (SPOTPY, v1.5.14) (Houska et al., 2015). SPOTPY offers a comprehensive set of methods for parameter optimization and sensitivity analysis (Houska et al., 2015). For each optimization run, the maximum 30,000 iterations were set. More details of how the SCE-UA was applied can be found in Yang et al., (2023b).

2.4. Model calibration and evaluation

2.4.1. Leave-One-Out Cross-Validation (LOOCV)

Evaluating the goodness-of-fit and predictive ability of models is essential. Goodness-of-fit assesses how well a calibrated model could fit

the data used for calibration, while predictive ability refers to how well a calibrated model can predict new data that were not involved in calibration (Wallach et al., 2023b). To balance these assessments, the Leave-One-Out Cross-Validation (LOOCV) technique was employed. In this approach, each site's data within a cultivar is iteratively used for evaluation/prediction, while the rest forms the calibration dataset for estimating optimal parameter values via the SCE-UA algorithm. This method treated each site-specific observation equally, assuming both calibration and evaluation data during LOOCV were drawn from the same underlying target population for a given cultivar, i.e., observations over a range of environments that we aimed to assess (**Fig. S1**), consistent with the approach suggested by Wallach et al., (2021a). LOOCV was carried out per model for each cultivar.

2.4.2. Statistical metrics

The variability in optimized parameter value during LOOCV, measured by the Coefficient of Variation (CV), was used to estimate the parameter uncertainty (Kawakita et al., 2020). Model performance, both for goodness-of-fit and out-of-sample predictions, was quantified using RMSE. Besides, following a recently proposed calibration protocol for crop phenology models that utilized the Bayesian Information Criterion (BIC) for model comparison (Wallach et al., 2023b), BIC was applied in our study and calculated as follow:

$$BIC = n\ln(MSE) + p\ln(n) \tag{11}$$

where n is the number of observations and p is the number of calibrated parameters. *MSE* stands for Mean Squared Error between calibration data and simulations. *BIC* is a criterion for model selection by quantitatively assessing the balance between model complexity and accuracy (Wallach et al., 2023b). Models with a lower *BIC* during LOOCV are generally preferred. Apart from BIC, an overall evaluation of calibrated models against observations was conducted across all sites for each maturity type. This evaluation utilized a suite of conventional statistical metrics, including Mean Absolute Error (MAE), Mean Biased Error (MBE), RMSE, coefficient of determination (R^2) and Nash-Sutcliffe modeling efficiency (EF) (Wallach et al., 2017, 2021a, 2023b; Zhang and Tao, 2019; Kawakita et al., 2020).

2.5. Identifying the main source of variability in regional-scale simulations

For each maturity type, six models were employed for simulations over 1993–2022 (a 30-year period to capture the recent climate variability) across the major paddy rice area for the Sichuan basin, as depicted in Fig. 1. These simulations were based on site-specific optimized parameters during LOOCV. The empirical heading onset was used to initialize the simulations according to the median observed value of each maturity type (**Fig. S1**). The ERA5-Land gridded dataset, used for model calibration and evaluation, also served as the weather input for these regional runs, ensuring consistency in the meteorological data across both site and regional scales.

The sources of variability in these simulations could apparently arise from differences in model structures (inter-model variability) and variance in the optimized model parameter values (intra-model variability). To disaggregate the overall variability, a repeated measures Analysis of Variance (ANOVA) with a balanced design (Table 3), similar to Wallach et al., (2017), was applied at each grid point, which was implemented with the python class AnovaRM (Repeated Measures ANOVA using least squares regression) from the statsmodels package. This approach considered the influence of both model structures and parameters, including their interaction effects. Each year within the 30-year span was treated as a repeated measure to examine the influence of these factors and their interactions. The contribution of uncertainty source (factors and their interactive term) to total simulation variance was quantified based on the computed *F*-values, effectively calculating the ratio of mean square for each factor (and their interactions) to the sum of mean squares of all sources but excluding the residual term (Table 3). As such, this approach estimated their contributions to the overall explained variance. This analytical framework was applied to two pivotal output variables: the timing of rice maturity and the GDD of the heading-maturity phase. Simulations for the latter variable can be useful to characterize the thermal conditions during a critical phase for different rice maturity types at a regional level.

Table 3
Design of the repeated measures ANOVA accounting for different crop phenology models (*models*), site-specific optimized parameter vectors (*parameters*) and their interaction effects (*models* × *parameters*). Simulations across different years are treated as repeated measures.

Sum of squares (SS)	Degree of freedom (df)
$SS_{total} = \sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{p=1}^{P}(Y_{sm} - \overline{Y})^2$	$df_{total} = T \times M \times P - 1$
$SS_{models} = \sum_{m=1}^{M} n_m(\overline{Y}_m - \overline{Y})^2$	$df_{models} = M - 1$
$SS_{parameters} = \sum_{p=1}^{P} n_p(\overline{Y}_p - \overline{Y})^2$	$df_{parameters} = P - 1$
$SS_{models \times parameters} = \sum_{m=1}^{M}\sum_{p=1}^{P} n_{mp}(\overline{Y}_{mp} - \overline{Y}_m - \overline{Y}_p + \overline{Y})^2$	$df_{models \times parameters} = (M-1) \times (P-1)$

Y_{sm} denotes the simulated value of the t-th year, m-th model and p-th parameter vector; $\overline{Y}$, $\overline{Y}_m$, $\overline{Y}_p$, $\overline{Y}_{mp}$ respectively denotes the mean simulate value of all simulations (grand mean), of the m-th model across all years and parameter vectors, of the p-th parameter vector across all years and models, of the combination of the m-th model and the p-th parameter vector across all years; n_m, n_p, n_{mp} respectively denotes the total number of simulated values for the m-th model across all years and parameter vectors, for the p-th parameter vector across all years and models, for the combination of the m-th model and the p-th parameter vector across all years; T, M, P respectively denotes the total number of years, models and parameter vectors.

3. Results

3.1. Model performance during LOOCV

3.1.1. Goodness-of-fit

When fitting models to calibration subsets during LOOCV, all models demonstrate satisfactory performance, with the median RMSE consistently below 4 days (mainly 2–4 days) for all cultivar types (Fig. 3**a**). The variation in median RMSE among models is minimal (<1 day), although APSIM tends to have marginally higher value and APSIM also exhibits a greater variance in RMSE compared to the other models (Fig. 3**a**). Models generally are better fitted to the early-maturity type than to the medium-maturity and late-maturity type (Fig. 3**a**).

For the early-maturity type, a notably lower BIC (<70) is discovered, in contrast to higher values (BIC>110) obtained for the other two types (Fig. 3**b**). GDD, Richardson, Wang and CERES, and to some extent Sigmoid (excluding the late-maturity cultivar), demonstrate comparable skills, with slight advantages observed for GDD and CERES (Fig. 3**b**). APSIM consistently shows the highest BIC values across all types (Fig. 3**b**).

3.1.2. Out-of-sample predictions

For out-of-sample predictions (evaluations), the median RMSE also varies between 2 and 4 days (Fig. 3**a**). Compared to the fitting performance, an increase in median RMSE is observed for the early-maturity type, while a decrease is noted for the medium-maturity and late-maturity type (Fig. 3**a**). However, a substantial increase is found for the variance of RMSE across cultivar types, indicating much larger prediction uncertainties than those for calibration data (Fig. 3**a**). Across cultivar type, the best predictive performance is found for the late-maturity type, with a median RMSE well below 3 days, followed by the medium-maturity and early-maturity type, respectively (Fig. 3**a**).

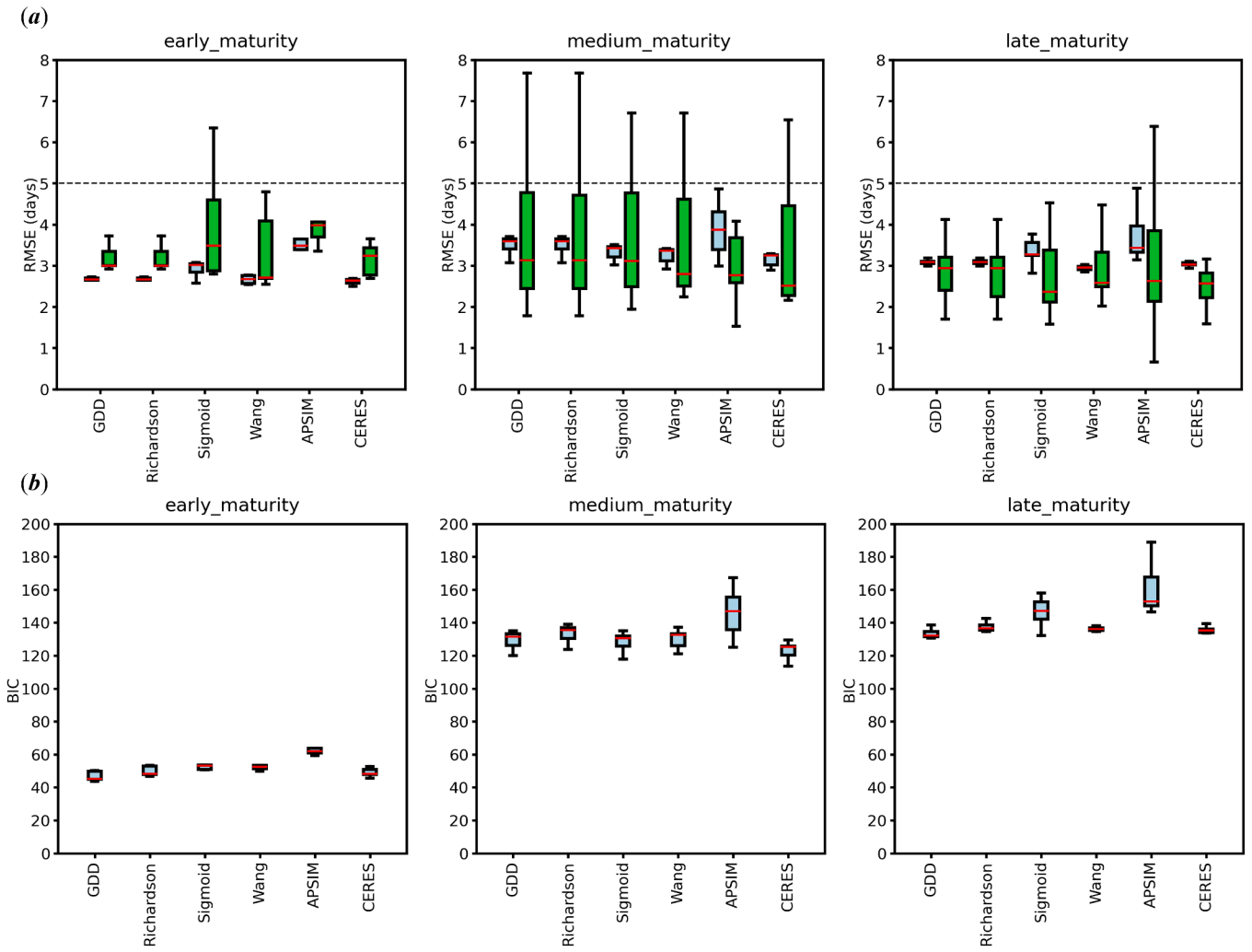


Fig. 3. Variability in (*a*) models' goodness-of-fit (blue) and out-of-sample predictions (green) and in (*b*) Bayesian Information Criterion (BIC) during the Leave-One-Out Cross-Validation (LOOCV). The box horizontal lines respectively represent the first (Q1), second (median denoted by red line) and third (Q3) quartile, while the lower and upper whisker respectively denotes values extending from Q1 and Q3 by 1.5 times the interquartile range (outliers are not shown). The dashline for RMSE represents the threshold value (RMSE=5 days) above which predictions are considered inaccurate.

Site-specific predictions show that the largest prediction errors tend to occur in sites experiencing a high number of HD (>30 days) during July-September, such as miyi, neijiang, yaan, and yanbian (**Fig.S5**).

When aggregating individual site predictions, MAE and RMSE remain within 3–4 days, with a negligible mean bias (MBE≤1) between models and across cultivar types (Fig. 4). The calibrated models show a good predictive skill and ability to capture observed data variability, as evidenced by EF (R^2) values, ranging between 0.67–0.80 (0.68–0.80), 0.80–0.87 (0.80–0.88) and 0.94–0.95 (0.94–0.95) for the early-maturity, medium-maturity and late-maturity, respectively (Fig. 4). Furthermore, the median of the ensemble model simulations shows nearly or equally as good as the best individual model across all maturity types (**Fig.S6**).

3.2. *Variability in estimated parameter values during LOOCV*

For the studied rice cultivars, most model parameters, except for one in the Sigmoid model, display low-to-moderate variability in their estimated optimal values, with CV≤20 % generally during LOOCV (Fig. 5**a**). Notably, the T_{base} parameter, common in five out of the six models (Table 2), exhibits higher variability compared to other threshold parameters like T_{opt}, T_{max} and T_{lim} (Fig. 5**a**). Parameters with CV<5 % can be considered as stable parameters, there are 7, 11 and 14 ones for the early-, medium- and late-maturity type respectively (Fig. 5**a**). In contrast, the Sigmoid models *d* parameter (defining curve sharpness) shows a significant variability, with a CV reaching up to 200 % (Fig. 5**a**), indicating high uncertainty for its parameter value. Across different cultivar types, the late-maturity type displays generally smaller variability in estimated parameter values compared to the other types, except for the Sigmoids *d* and Wangs T_{base} parameter (Fig. 5**a**).

For each model, the temperature response functions resulting from the variability of calibrated parameter values are presented in Fig. 5**b**. The Sigmoid model shows distinct response curves across cultivars, directly relating to its high variability of the shape parameter *d* (Fig. 5**b**). In contrast, the GDD and Richardson models maintain a similar linear response curve. More specifically, the Richardson model is calibrated with high plateaus that are seldom reached for local conditions (Fig. 5**b**). Wang and CERES models exhibit a notable variability in their response functions during LOOCV across maturity types (Fig. 5**b**). The APSIM model shows the least variability in its response curves, in association with its little variability of calibrated parameter values (Fig. 5**b**). Moreover, for the late-maturity type, the temperature functions of differently calibrated parameters are more stable than those of other types, where the model like GDD, Richardson, Wang and APSIM tends to converge into similar forms, indicative of their low parameter variability as shown in Fig. 5**a**.

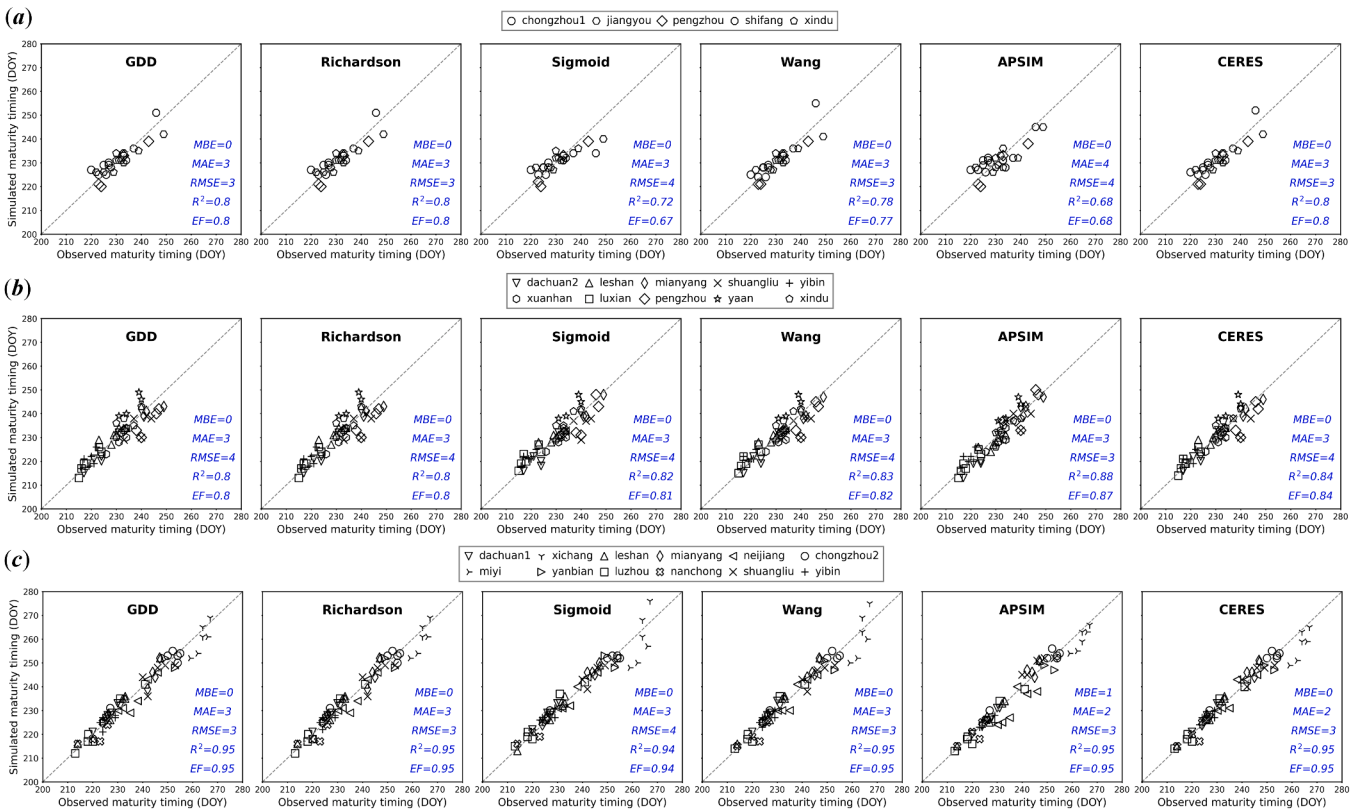


Fig. 4. Comparison between observed and predicted maturity timing (DOY) of calibrated models during the Leave-One-Out Cross-Validation (LOOCV) for (*a*) early-maturity, (*b*) medium-maturity and (*c*) late-maturity rice cultivar. Individual model predictions at each site are based on optimized parameters using the rest of sites' data of a given rice cultivar. Mean Bias Error (MBE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), the coefficient of determination (R^2) and Nash Sutcliffe modeling efficiency (EF) are shown.

3.3. Variability in regional simulations between model structures and parameters

3.3.1. Simulations of multi-models with different parameterizations

Regional simulated maturity timings over 1993–2022 exhibit a considerable variability between models and their estimated optimal parameters during LOOCV (**Fig.S7–9**). The median maturity stage (DOY) ranges from about 220–235, 225–240, 230–245 in the northeast areas (latitude>28°N, longitude>103°E) to 240–260, 240–260, 240–260 in the southwest areas (latitude<29°N, longitude≤103°E) for the early-maturity (**Fig.S7**), medium-maturity (**Fig.S8**) and late-maturity (**Fig.S9**) type respectively. The APSIM model, exhibiting the least variability in its temperature response functions (Fig. 5**b**), presents the largest variability in its regional simulations using these functions, especially for the medium- and late-maturity types (**Fig.S8–9**). In contrast, the Sigmoid model, showing the greatest variability in its response curves (Fig. 5**b**), leads to relatively small variability for regional simulations (**Fig.S8–9**).

3.3.2. ANOVA results

ANOVA dissects the total variance of regional maturity simulations into contributions from model structures, parameters and their interaction effects (Fig. 6). Variability due to calibrated model parameters accounts for approximately 15–90 % of the total variance for the early-maturity type, and 15–45 % and 30–60 % for the medium- and late-maturity types, respectively (Fig. 6**a**). Inter-model variability contributes about 0–45 % for the early-maturity type and 0–75 % and 0–45 % for the medium- and late-maturity types, respectively (Fig. 6**b**). The interaction between models and parameters contributes to around 15–45 %, 15–75 % and 15–60 % of the total variance across the three maturity types (Fig. 6**c**). As a result, parameter variability emerges as the dominant uncertainty source for the early-maturity type in most areas, except some in the northeast (Fig. 6**d**). For the medium- and late-maturity types, the interactive effects represent the main uncertainty source, except in the southwest (mainly for the medium-maturity type) where the inter-model variability tends to mostly affects the simulations (Fig. 6**d**). A similar pattern is discovered for the computed GDD of the heading-maturity phase, with parameter variability maintaining its dominant role in uncertainty analysis for the early-maturity type in most production areas (**Fig.S10**). For the medium-maturity and late-maturity types, interaction effects are still the dominant uncertainty source in the northeast areas, whereas inter-model variability plays a more significant role in the southwest for both maturity types (**Fig.S10**).

4. Discussion

4.1. Model goodness-of-fit

Assessment of how well calibrated models can fit the observed data used for calibration is an integral part of crop phenology modelling studies (Wallach et al., 2017; Zhang et al., 2017; Liu et al., 2018; Gao et al., 2020; Yang et al., 2021). Our findings highlight the usefulness of SCE-UA in optimizing model parameters across diverse rice cultivars, achieving a similar fitting performance among models with median RMSE ranging from 2 to 4 days, echoing the calibration accuracy by Liu et al., (2018) and Wallach et al., (2021a). Reliable estimations of parameter values play a pivotal role, which can ensure that the remaining discrepancy between observations and simulations can be attributed to limitations in the model structure or data, rather than the sub-optimal parameters (Liu et al., 2018).

BIC analysis, assessing the balance between model complexity and accuracy, complements RMSE results and is an important criterion when

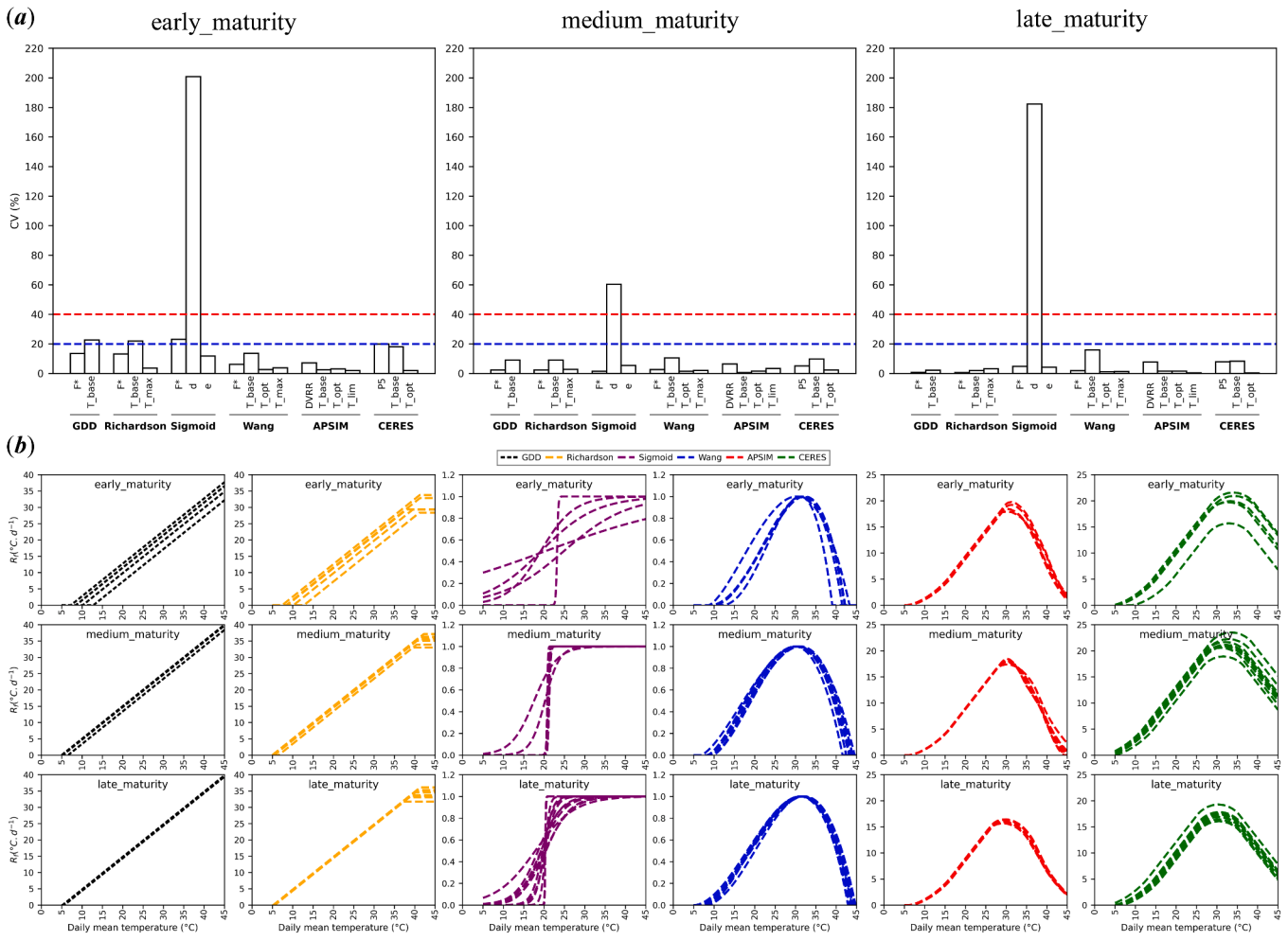


Fig. 5. Variability in optimized model parameters and corresponding temperature response functions during the Leave-One-Out Cross-Validation (LOOCV). (*a*) The Coefficient of Variation (CV, %) in optimized model parameters using all sites' data except for one for a given cultivar (blue and red dotted lines respectively denote 20 % and 40 % of CV values); (*b*) different temperature response functions due to variability in optimized parameters.

comparing models calibrated using the same data (Wallach et al., 2023b). Models generally showcase equivalent performance, except for APSIM that exhibits a slightly higher BIC, relating to its higher number of parameters (Table 2) and marginally higher calibration errors. Models with more parameters tend to express crop development in a more complex way (Kawakita et al., 2020), but it does not necessarily translate to superior fitting performance. All models have a considerably lower BIC for the early-maturity type than the other two types (Fig. 3**b**), attributed to the fewer calibration sites that result in a low *n* value (eq. 11).

4.2. *Model out-of-sample predictions*

The best predictive performance is found for the late-maturity type with the highest number of trial sites, followed by the medium-maturity and early-maturity type respectively. This can be due to that the prediction skill of models can be asymptotically improved when more observational data is utilized for model calibration (Montesino-San Martin et al., 2018). Moreover, models calibrated with data having characteristics of averagely warmer conditions (Fig. 2) and higher instances of HD episodes (**Fig.S3**) notably excel, such as those for the late-maturity type. This pattern is evident when examining individual site predictions. Models calibrated using sites' data with less frequent HD underperform in predicting sites with prevalent high-temperature periods (HD>30) (**Fig.S5**). This could be related to that the upper temperature thresholds are not properly fit for these models with the lack of HD events, thus leading to inaccurate description of crop response when HD becomes more frequent. In contrast, good predictability is discovered at sites (e.g. chongzhou1, chongzhou2, jiangyou, leshan, mianyang) with less frequent HD (HD<15) when they are trained with data subject to more HD (**Fig.S5**). Several studies point out model performance is limited under high-temperature conditions, and there is a need to conduct temperature-controlled experiments to better understand how the above-optimal high temperatures affect crop development rate (Shi et al., 2015; Zhang et al., 2017; Zhang and Tao, 2019).

Aggregating LOOCV predictions across all sites reveals an overall satisfactory performance across models and cultivar types, with MAE and RMSE within a 3–4 days range and EF (R^2) ranging between 0.67–0.95 (0.68–0.95) (Fig. 4). The performance is similar to a study simulating rice maturity at two different zones in China using five phenology models (Zhang et al., 2017). Moreover, the best model can vary in different situations, but the ensemble median of all model simulations always gives nearly as good ($R^2 \geq 0.79$) as the best model results (**Fig.S6**), advocating for ensemble approaches in enhancing prediction reliability, the finding aligned with Wallach et al., (2021b, 2021a). Despite overall accuracy, prediction errors significantly fluctuate across sites, leading to a considerable prediction uncertainty, i.e., a consistently larger variance of RMSE in the evaluation data than in the calibration data (Fig. 3**a**). Partly this is associated with differently calibrated temperature response functions of the models when predicting for individual sites. On the other hand, a good fit to the current data does not guarantee its repeatability of predictability in other situations, even if

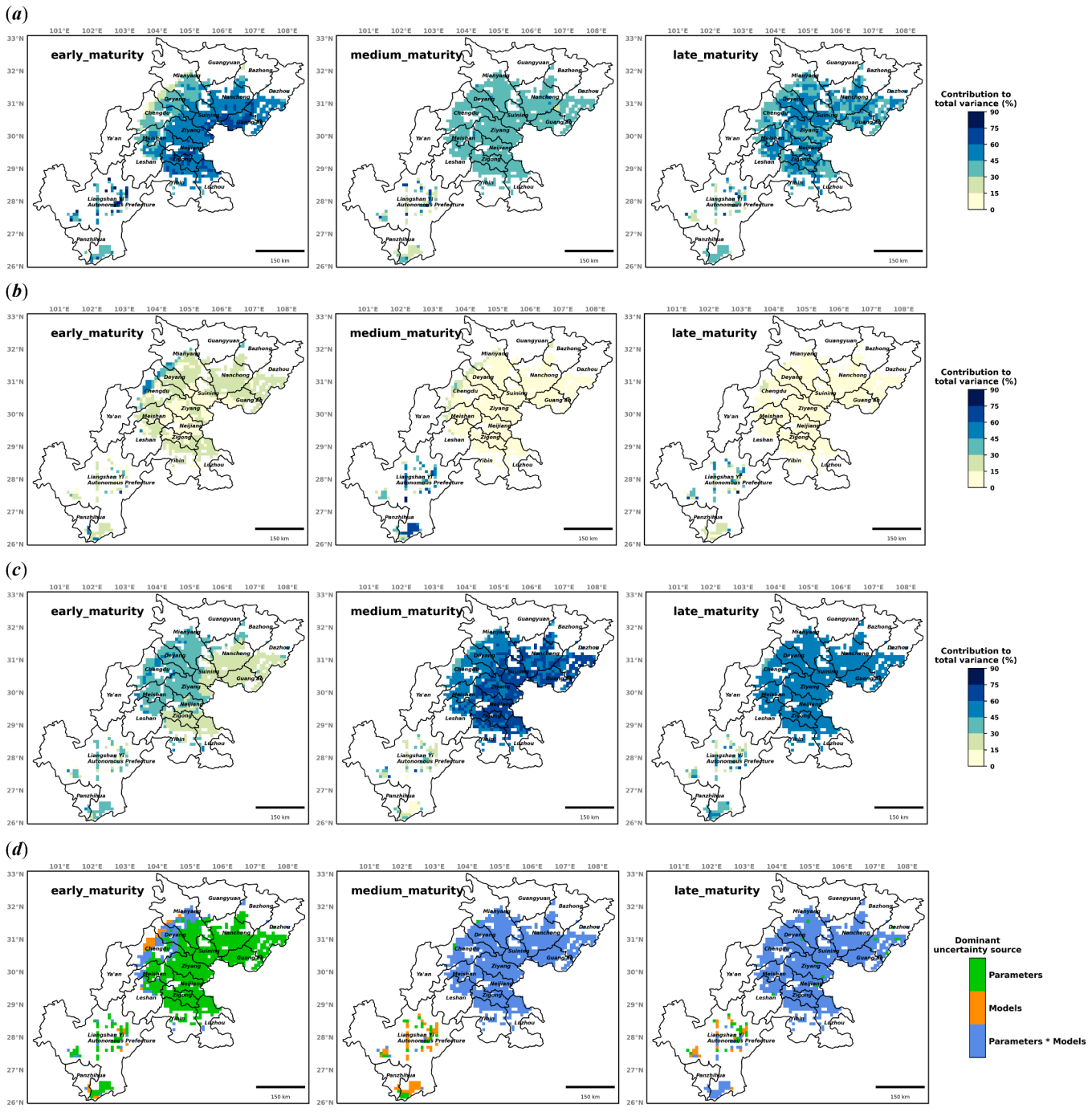


Fig. 6. The contributions (%) to the total variance in simulating the rice maturity timings (DOY) of different cultivar types in Sichuan Province's main rice production area over 1993–2022 by (*a*) calibrated model parameters (Parameters) and (*b*) model structures (Models), as well as (*c*) their interactions (Parameters * Models). (*d*) The dominant source of uncertainty is identified. The Analysis of Variance (ANOVA) treats different years as repeated measures to assess the impacts of different factors and their interactions

the occurrence of a new situation shares similar characteristics to those incorporated in the model calibration (Seidel et al., 2018; Kawakita et al., 2020; Wallach et al., 2021c, 2021b, 2021a). This highlights the inherent uncertainty in model predictions and the crucial role of not just evaluating model predictions, but also in terms of prediction consistency across different environmental conditions.

For the early-maturity, medium-maturity and late-maturity types, GDD/Richardson, APSIM and CERES are considered as the best choice of models, respectively (Fig. 3 and Fig. 4). As the best predictive model is contingent upon the cultivar type, considering the balance between the prediction accuracy and consistency. For rice phenology modelling, GDD and Richardson are rarely chosen, probably because of their simple linear structure. In the present study, these two models are demonstrated to yield reliable simulations, as they perform equally to or even exceed the other non-linear models, along with robust parameterizations. For predicting the heading-maturity phase, during which the photoperiod effects can be ignored, utilizing relatively complex models (APSIM and CERES) with a daily interpolation routine might lead to a more precise prediction on the crop development, but a simpler thermal-time accumulation approach (GDD and Richardson) demanding less computational resources could already offer enough predictability.

4.3. Parameter uncertainty and extended regional simulations

Obtaining reliable parameter values are key for analyzing the relationship between parameters and cultivar phenotype. Most parameters exhibit a low-to-moderate variability, underscoring the overall robustness of model calibrations during LOOCV (Fig. 5). More stable (CV<5 %) parameters in the late-maturity than in the other two types can again illustrate the importance of adequate data amount for calibration, which might be useful for reducing parameter uncertainties. The $T_{\_base}$ parameter among models shows a higher variability than the other cardinal temperature thresholds. Cardinal temperature parameters are variable between cultivars and regions (Zhang et al., 2017), and particular attentions should be paid to the parameter triggering the initial crop development. The higher variability of $T_{\_base}$ in the early-maturity cultivar compared to the others (Fig. 5**a**) underscores the need for more observational data to achieve robust estimations for this critical parameter. Exceptionally, the Sigmoid model's *d* parameter consistently shows a high variability up to 200 %, leading to distinct temperature response curves for the same cultivar. This, however, well aligns with prior studies indicating that the Sigmoid model's curve sharpness parameter is highly uncertain and sensitive, especially when observational data are not sufficient enough (Kawakita et al., 2020; Yang et al., 2023b, 2023a). Site-specific estimations underscore the sensitivity of this parameter, and uncertainties from random site effects, as referred to by Wallach et al., (2017), can significantly influence its estimations. Such variability also suggests that while the Sigmoid model performs adequately across calibration and evaluation data in general, it can potentially have overly steep temperature response functions, such as those in the medium-maturity cultivar (Fig. 5**b**). Specifically for this cultivar, two distinctively calibrated curves for the Sigmoid model (Fig. 5**b**) are obtained from mianyang and yaan, which have the lowest T_{min} (21.2–21.6 °C) and T_{mean} (24.4–24.5 °C) among associated sites (Fig. 2). The mean growing degree days of July–August for these two coolest sites and the other sites are 982–997 °C d^{-1} and 1025–1145 °C d^{-1}, respectively. This might suggest when temperature conditions exceed a certain threshold, the cultivar development can respond similarly as described by the Sigmoid model.

This study extends the site-specific calibrated parameters of the model ensemble from LOOCV to regional simulations for each maturity type. The simulations adeptly capture the inherent maturity timing characteristics of each cultivar type, accurately reflecting the expected regional maturity patterns (**Fig.S7–9**). For the medium-maturity and late-maturity types, the APSIM model, noted for its minimal function curve uncertainty, exhibits the biggest variability in its regional simulations (spatial patterns) between different parameterizations. This can be due to the model's sensitivity and responsiveness, especially at higher temperatures. For APSIM, the difference in calibrated temperature functions is only notable for temperatures above the optimum threshold (Fig. 5**b**). Hence, the model can have a larger variability in simulations over 1993–2022 with a broad spectrum of temperature conditions, including frequent above-optimal temperatures, than the calibration period. This can already be seen in Fig. 3**a**, where APSIM shows the biggest prediction variability for the late-maturity type that has more sites subject to frequent HD (>30). Conversely, the Sigmoid model, despite its pronounced temperature function variability, shows relatively small variability in its regional simulations. High parameter variability does not necessarily contribute to high prediction variability. This is often associated with parameter correlation, or equifinality (Beven, 2006), where parameters with higher uncertainties tend to relate with other parameters, resulting in different parameter combinations for similar/same simulation results (Liu et al., 2018; Kawakita et al., 2020; Wallach et al., 2021c; Yang et al., 2023b).

4.4. Source of uncertainty in regional simulations

The contribution of uncertainty sources to rice maturity simulations varies with the cultivar maturity type. For the early-maturity type, parameter variability is predominant, which is associated with its notable variability in calibrated model parameters during LOOCV. As such, more observational data is needed to refine the parameter estimates and reduce parameter uncertainties for this cultivar type. For the medium-maturity and late-maturity types, a clear spatial pattern emerges that delineate the relative contributions of uncertainty sources. In the northeast, parameter variability overshadows the model structural differences (Fig. 6**a–b**). Yet, in the southwest areas prone to high-temperature extremes, model structural variability becomes paramount (Fig. 6**d**), hinting at models' divergent responses to temperature extremes. Uncertainties exist among models in accurately describing the abrupt decline in development rate when temperature exceeds the optimal threshold (Zhang et al., 2017; Zhang and Tao, 2019). For the calculated GDD of the heading-maturity phase, similar results are obtained, particularly since the model variability is more pronounced in the southwest areas (**Fig.S10**).

Our findings elucidate that while different combinations of models and parameterizations can be the dominant source of uncertainty, parameter variability consistently plays a more important role than model variability for a major portion of the study region. Model variability is more important in only limited areas prone to high-temperature extremes. Zhang et al., (2017) report that the model structures and parameters could explain 92.15 % and 7.85 % of the total variability in simulated timing to maturity for rice in the southwestern China. Similarly, the variability in simulations between models can contribute twice as much as parameter variability to the total variability for rice phenology simulations (Wallach et al., 2017). However, parameter uncertainties in these studies often do not incorporate the variability of observational dataset for calibration (Wallach et al., 2017; Zhang et al., 2017). In general, the main source of uncertainty in simulations not only depends on the cultivar and study region, but also on how the parameter uncertainty is defined. It is of interest to treat parameter uncertainty with a broader perspective. For instance, the uncertainty perspective can incorporate sources not only from different site × year combinations of a give cultivar, but also from different calibration approaches (e.g., multiple optimization algorithms, objective functions) and choices of estimating parameters (Seidel et al., 2018; Wallach et al., 2021b, 2021a). It is found that when the uncertainty in the choice of parameters is considered, in most cases, parameter variability accounts for much more of the total simulation variability than the model structural variability (Wallach et al., 2023a). A careful model calibration and evaluation is the prerequisite for regional simulations that often use parameters obtained from field data. Our study demonstrates the variability in observed data of a given cultivar at field scale can be an important source of uncertainties for regional runs.

4.5. Limitations of the study

Firstly, this study focuses exclusively on the heading-maturity phase rather than the entire growing period. While this phase is crucial for rice yield formation, an analysis that includes the entire growing cycle would provide a more comprehensive understanding of uncertainties in crop phenology simulations. However, models such as GDD, GDD-Richardson, Sigmoid and Wang inherently omit photoperiod effects in their equations, making simulations before the heading or anthesis stage difficult.

Secondly, the impacts of transplanting dates and management were not explicitly addressed. Both transplanting date and management practice (e.g., planting date) can have a greater effect on rice maturity timing in China than climate change (Zhao et al., 2016; Wang et al., 2017). For instance, Zhao et al., (2016) indicate that the maturity timing of single, early and late rice in China's primary producing regions is closely linked to transplanting and heading dates, while climatic factors are of secondary importance. Although the variability in the observed heading stage (where simulations start) in this study (**Fig.S1**) may

implicitly reflect the influence of different transplanting dates and managements, a more explicit analysis of these variables and their interaction with different model structures and parameters is required.

While our study focuses on temperature response during an important heading-maturity phase, future research can build on these findings and delve into an entire growth cycle to investigate how to further reduce uncertainties in crop phenology simulations.

5. Conclusions

In this study, we dissect the contributions of model structural uncertainties and parameter uncertainties that are derived from field data to the overall uncertainties in regional simulations. Utilizing six phenology models, they are calibrated with a global parameter optimization algorithm and evaluated for simulating the maturity timing of three representative rice cultivars using trial data within the Sichuan Basin (China). The employed Leave-One-Out Cross-Validation (LOOCV) allows to quantify parameter uncertainties across various calibration data subsets for each cultivar. The median of the RMSE range is mainly 2–4 days for both goodness-of-fit and out-of-sample predictions during LOOCV. Models calibrated with data from sites experiencing warmer conditions with frequent high-temperature episodes tend to have better predictability. Despite considerable prediction uncertainty due to the fluctuating prediction errors across sites, an overall satisfactory prediction performance is achieved, with the ensemble median of all model simulations explaining over 80 % of the total variation among the cultivars. Notably, the most effective predictive model differs by cultivar, with simpler thermal-time accumulation approaches (GDD and Richardson) performing comparably to relatively more complex models (APSIM and CERES). During LOOCV, most model parameters exhibit a low-to-moderate variability except one parameter in the Sigmoid model.

For regional simulations, the parameter variability obtained from LOOCV consistently shows a more important role than the model structure variability for most of the region. Specifically, it emerges as the predominant uncertainty factor for the early-maturity cultivar, while for medium- and late-maturity cultivars, though the parameter variability outweighs the importance of model variability, a nuanced interplay between model structures and parameters predominates for most of the rice production area. However, for these two cultivars, model structural variability became the principal uncertainty factor in areas subjected to frequent high-temperature extremes, reflecting models' divergent responses to temperature extremes. These findings illustrate that the relative contribution of uncertainty sources depends on both the study cultivar and region. In this study, parameter uncertainty is estimated only based on the variability in calibration subsets per cultivar, which can be further amplified when the inherent variability in calibration approaches is considered. Therefore, a broader conceptualization of parameter uncertainty is advocated. Accurate identification of the dominant sources of uncertainty is fundamental for refined calibration strategies and model improvements, which is vital for improving the accuracy and robustness of regional crop modelling.

CRediT authorship contribution statement

Chenyao Yang: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Na Lei:** Writing – review & editing, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Christoph Menz:** Writing – review & editing, Validation, Supervision, Software, Methodology, Investigation, Data curation, Conceptualization. **Andrej Ceglar:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Jairo Arturo Torres-Matallana:** Writing – review & editing, Software, Methodology, Investigation. **Siqi Li:** Validation, Resources, Investigation, Formal analysis. **Yanling Jiang:** Resources, Investigation. **Xianming Tan:** Resources, Investigation. **Lei Tao:** Resources, Investigation, Data curation, Conceptualization. **Fang He:** Supervision, Resources, Project administration, Conceptualization. **Shigui Li:** Supervision, Resources, Project administration, Conceptualization. **Bing Liu:** Formal analysis, Resources, Validation, Visualization, Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Feng Yang:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Helder Fraga:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **João A. Santos:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Funding

This study is supported and funded by FCT—Portuguese Foundation for Science and Technology, under the project UIDB/04033/2020 (https://doi.org/10.54499/UIDB/04033/2020) and Inov4Agro (LA/P/0126/2020), the Jiangsu Provincial Department of Science and Technology (BE2023400), the Natural Science Foundation of Jiangsu Province (BK20220146), and the Jiangsu Independent Innovation Fund Project of Agricultural Science and Technology [CX(23)3121]. The study is also funded by the project "Crop Growth Model and Smart Decision Support System" with grant No.12322349002 was developed at the Smart Agriculture Innovation Lab, co-founded by China Telecom and Sichuan Agricultural University (SAU), and funded by SAU's National/Provincial College Students' Innovation and Entrepreneurship Training Program Sponsorship.

Acknowledgements

We kindly acknowledge the data provision from the local meteorological bureau. Helder Fraga thanks the FCT for 2022.02317.CEECIND (https://doi.org/10.54499/2022.02317.CEECIND/CP1749/CT0002) and WaterQB 2022.04553.PTDC.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.agrformet.2024.110137.

References

Arlo Richardson, E., Seeley, S.D., Walker, D.R, 1974. A model for estimating the completion of rest for 'Redhaven' and 'Elberta' Peach Trees1. HortScience 9, 331–332. https://doi.org/10.21273/HORTSCI.9.4.331.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., et al., 2013. Uncertainty in simulating wheat yields under climate change. Nat. Clim. Change 3, 827–832. https://doi.org/10.1038/nclimate1916.

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320, 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007.

Bonhomme, R., 2000. Bases and limits to using 'degree.day' units. Eur. J. Agron. 13, 1–10. https://doi.org/10.1016/S1161-0301(00)00058-7.

Boote, K.J., Allen, L.H., Prasad, P.V.V, Baker, J.T., Gesch, R.W., Snyder, A.M., et al., 2005. Elevated temperature and CO2 impacts on pollination, reproductive growth,

and yield of several globally important crops. J. Agric. Meteorol. 60, 469–474. https://doi.org/10.2480/agrmet.469.
Bouman, B., Kropff, M., Tuong, T.P., Wopereis, M.C., Berge, ten, et al., 2001. ORYZA2000: modeling lowland rice. International Rice Research Institute, Los Baños, Philippines.
Brisson, N., Launay, M., Mary, B., Beaudoin, N., 2009. Conceptual basis, formalisations and parameterization of the STICS crop model. Editions Quae, Versailles, France.
Ceglar, A., van der Wijngaart, R., de Wit, A., Lecerf, R., Boogaard, H., Seguini, L., et al., 2019. Improving WOFOST model to simulate winter wheat phenology in Europe: Evaluation and effects on yield. Agric. Syst. 168, 168–180. https://doi.org/10.1016/j.agsy.2018.05.002.
Duan, Q., Sorooshian, S., Gupta, V.K., 1994. Optimal use of the SCE-UA global optimization method for calibrating watershed models. J. Hydrol. 158, 265–284. https://doi.org/10.1016/0022-1694(94)90057-4.
Duan, Q.Y., Gupta, V.K., Sorooshian, S., 1993. Shuffled complex evolution approach for effective and efficient global minimization. J. Optim. Theory Appl. 76, 501–521. https://doi.org/10.1007/BF00939380.
Gao, L., Jin, Z., Huang, Y., Zhang, L., 1992. Rice clock model—a computer model to simulate rice development. Agric. For. Meteorol. 60, 1–16. https://doi.org/10.1016/0168-1923(92)90071-B.
Gao, Y., Wallach, D., Liu, B., Dingkuhn, M., Boote, K.J., Singh, U., et al., 2020. Comparison of three calibration methods for modeling rice phenology. Agric. For. Meteorol. 280, 107785 https://doi.org/10.1016/j.agrformet.2019.107785.
Hänninen, H. (1990). Modelling bud dormancy release in trees from cool and temperate regions.
Houska, T., Kraft, P., Chamorro-Chavez, A., Breuer, L., 2015. SPOTting model parameters using a ready-made python package. PLoS One 10, e0145180. https://doi.org/10.1371/journal.pone.0145180. Available at.
IPCC, 2022. Climate change 2022: impacts, adaptation, and vulnerability. In: Pörtner, B. R.H.-O., Roberts, D.C., Tignor, M., Poloczanska, E.S., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A. (Eds.), Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UKNew York, NY, USA. https://doi.org/10.1017/9781009325844 andCambridge University Press.
Jin, N., Tao, B., Ren, W., He, L., Zhang, D., Wang, D., et al., 2022. Assimilating remote sensing data into a crop model improves winter wheat yield estimation based on regional irrigation data. Agric. Water Manag. 266, 107583 https://doi.org/10.1016/j.agwat.2022.107583.
Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., et al., 2003. The DSSAT cropping system model. Eur. J. Agron. 18, 235–265. https://doi.org/10.1016/S1161-0301(02)00107-7.
Kawakita, S., Takahashi, H., Moriya, K., 2020. Prediction and parameter uncertainty for winter wheat phenology models depend on model and parameterization method differences. Agric. For. Meteorol. 290, 107998 https://doi.org/10.1016/j.agrformet.2020.107998.
Kersebaum, K.C., Boote, K.J., Jorgenson, J.S., Nendel, C., Bindi, M., Frühauf, C., et al., 2015. Analysis and classification of data sets for calibration and validation of agro-ecosystem models. Environ. Model. Softw. 72, 402–417. https://doi.org/10.1016/j.envsoft.2015.05.009.
Liu, J., Liu, Z., Zhu, A.-X., Shen, F., Lei, Q., Duan, Z., 2019. Global sensitivity analysis of the APSIM-Oryza rice growth model under different environmental conditions. Sci. Total Environ. 651, 953–968. https://doi.org/10.1016/j.scitotenv.2018.09.254.
Liu, L., Wallach, D., Li, J., Liu, B., Zhang, L., Tang, L., et al., 2018. Uncertainty in wheat phenology simulation induced by cultivar parameterization under climate warming. Eur. J. Agron. https://doi.org/10.1016/j.eja.2017.12.001.
Montesino-San Martin, M., Wallach, D., Olesen, J.E., Challinor, A.J., Hoffman, M.P., Koehler, A.K., et al., 2018. Data requirements for crop modelling—Applying the learning curve approach to the simulation of winter wheat flowering time under climate change. Eur. J. Agron. 95, 33–44. https://doi.org/10.1016/j.eja.2018.02.003.
Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al., 2021. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data 13, 4349–4383. https://doi.org/10.5194/essd-13-4349-2021.
Muttil, N., Jayawardena, A.W., 2008. Shuffled Complex Evolution model calibrating algorithm: enhancing its robustness and efficiency. Hydrol. Process. 22, 4628–4638. https://doi.org/10.1002/hyp.7082.
Prasad, R., Shivay, Y. S., and Kumar, D. (2017). "Current status, challenges, and opportunities in rice production BT - rice production worldwide," in, eds. B. S. Chauhan, K. Jabran, and G. Mahajan (Cham: Springer International Publishing), 1–32. doi: 10.1007/978-3-319-47516-5_1.
Ritchie, J.R., Otter, S., 1985. Description and performance of CERES-wheat: a user-oriented wheat yield model. ARS-United States Dep. Agric. Agric. Res. Serv.
Ritchie, J.T., Singh, U., Godwin, D.C., Bowen, W.T., 1998. Cereal growth, development and yield - understanding options for agricultural production. In: Tsuji, G.Y., Hoogenboom, G., Thornton, P.K. (Eds.), Systems Approaches for Sustainable Agricultural Development. Springer Netherlands, Dordrecht, pp. 79–98. https://doi.org/10.1007/978-94-017-3624-4_5.
Rötter, R.P., Hoffmann, M.P., Koch, M., Müller, C., 2018. Progress in modelling agricultural impacts of and adaptations to climate change. Curr. Opin. Plant Biol. 45, 255–261. https://doi.org/10.1016/j.pbi.2018.05.009.
Seidel, S.J., Palosuo, T., Thorburn, P., Wallach, D, 2018. Towards improved calibration of crop models – Where are we now and where should we go? Eur. J. Agron. https://doi.org/10.1016/j.eja.2018.01.006.
Shao, J., Li, Y., Ni, J., 2012. The characteristics of temperature variability with terrain, latitude and longitude in Sichuan-Chongqing Region. J. Geogr. Sci. 22, 223–244. https://doi.org/10.1007/s11442-012-0923-4.
Shen, R., Pan, B., Peng, Q., Dong, J., Chen, X., Zhang, X., et al., 2023. High-resolution distribution maps of single-season rice in China from 2017 to 2022. Earth Syst. Sci. Data 15, 3203–3222. https://doi.org/10.5194/essd-15-3203-2023.
Shi, P., Tang, L., Lin, C., Liu, L., Wang, H., Cao, W., et al., 2015. Modeling the effects of post-anthesis heat stress on rice phenology. F. Crop. Res. 177, 26–36. https://doi.org/10.1016/j.fcr.2015.02.023.
Singh, U.(Upendra), Ritchie, J.T.(Joe T., Godwin, D.C., 1993. A user's guide to CERES rice, V2.10. International Fertilizer Development Center, Muscle Shoals, Alabama.
Tao, F., Rötter, R.P., Palosuo, T., Gregorio Hernández Díaz-Ambrona, C., Mínguez, M.I., Semenov, M.A., et al., 2018. Contribution of crop model structure, parameters and climate projections to uncertainty in climate change impact assessments. Glob. Change Biol. 24, 1291–1307. https://doi.org/10.1111/gcb.14019.
Wallach, D., Nissanka, S.P., Karunaratne, A.S., Weerakoon, W.M.W., Thorburn, P.J., Boote, K.J., et al., 2017. Accounting for both parameter and model structure uncertainty in crop model predictions of phenology: a case study on rice. Eur. J. Agron. 88, 53–62. https://doi.org/10.1016/j.eja.2016.05.013.
Wallach, D., Palosuo, T., Mielenz, H., Buis, S., Thorburn, P., Asseng, S., et al. (2023a). Uncertainty in crop phenology simulations is driven primarily by parameter variability. bioRxiv, 2023.02.03.526931. doi: 10.1101/2023.02.03.526931.
Wallach, D., Palosuo, T., Thorburn, P., Gourdain, E., Asseng, S., Basso, B., et al., 2021a. How well do crop modeling groups predict wheat phenology, given calibration data from the target population? Eur. J. Agron. 124, 126195 https://doi.org/10.1016/j.eja.2020.126195.
Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Andrianasolo, F., Asseng, S., et al., 2021b. Multi-model evaluation of phenology prediction for wheat in Australia. Agric. For. Meteorol. 298–299, 108289 https://doi.org/10.1016/j.agrformet.2020.108289.
Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Gourdain, E., Andrianasolo, F., et al., 2021c. The chaos in calibrating crop models: lessons learned from a multi-model calibration exercise. Environ. Model. Softw. 145, 105206 https://doi.org/10.1016/j.envsoft.2021.105206.
Wallach, D., Palosuo, T., Thorburn, P., Mielenz, H., Buis, S., Hochman, Z., et al., 2023b. Proposal and extensive test of a calibration protocol for crop phenology models. Agron. Sustain. Dev. 43, 46. https://doi.org/10.1007/s13593-023-00900-0.
Wang, B., Feng, P., Liu, D.L., O'Leary, G.J., Macadam, I., Waters, C., et al., 2020. Sources of uncertainty for wheat yield projections under future climate are site-specific. Nat. Food 1, 720–728. https://doi.org/10.1038/s43016-020-00181-w.
Wang, E., Engel, T., 1998. Simulation of phenological development of wheat crops. Agric. Syst. 58, 1–24. https://doi.org/10.1016/S0308-521X(98)00028-6.
Wang, X., Ciais, P., Li, L., Ruget, F., Vuichard, N., Viovy, N., et al., 2017. Management outweighs climate change on affecting length of rice growing period for early rice and single rice in China during 1991–2012. Agric. For. Meteorol. 233, 1–11. https://doi.org/10.1016/j.agrformet.2016.10.016.
Yang, C., Ceglar, A., Menz, C., Martins, J., Fraga, H., Santos, J.A., 2023a. Performance of seasonal forecasts for the flowering and veraison of two major Portuguese grapevine varieties. Agric. For. Meteorol. 331, 109342 https://doi.org/10.1016/j.agrformet.2023.109342.
Yang, C., Menz, C., Fraga, H., Reis, S., Machado, N., Malheiro, A.C., et al., 2021. Simultaneous calibration of grapevine phenology and yield with a soil–plant–atmosphere system model using the frequentist method. Agronomy 11. https://doi.org/10.3390/agronomy11081659.
Yang, C., Menz, C., Reis, S., Machado, N., Santos, J.A., Torres-Matallana, J.A., 2023b. Calibration for an ensemble of grapevine phenology models under different optimization algorithms. Agronomy 13. https://doi.org/10.3390/agronomy13030679.
Zhang, S., Tao, F., 2013. Modeling the response of rice phenology to climate change and variability in different climatic zones: comparisons of five models. Eur. J. Agron. 45, 165–176. https://doi.org/10.1016/j.eja.2012.10.005.
Zhang, S., Tao, F., 2019. Improving rice development and phenology prediction across contrasting climate zones of China. Agric. For. Meteorol. 268, 224–233. https://doi.org/10.1016/j.agrformet.2019.01.019.
Zhang, S., Tao, F., Zhang, Z., 2017. Uncertainty from model structure is larger than that from model parameters in simulating rice phenology in China. Eur. J. Agron. 87, 30–39. https://doi.org/10.1016/j.eja.2017.04.004.
Zhao, H., Fu, Y.H., Wang, X., Zhao, C., Zeng, Z., Piao, S., 2016. Timing of rice maturity in China is affected more by transplanting date than by climate change. Agric. For. Meteorol. 216, 215–220. https://doi.org/10.1016/j.agrformet.2015.11.001.
Zheng, J., Zhang, S., 2023. Improving rice phenology simulations based on the Bayesian model averaging method. Eur. J. Agron. 142, 126646 https://doi.org/10.1016/j.eja.2022.126646.