Computers and Electronics in Agriculture 227 (2024) 109594

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Integrating remote sensing assimilation and SCE-UA to construct a grid-by-grid spatialized crop model can dramatically improve winter wheat yield estimate accuracy

Qiang Li [a], Maofang Gao [a,*], Sibo Duan [a], Guijun Yang [b,c], Zhao-Liang Li [a]

[a] State Key Laboratory of Efficient Utilization of Arid and Semi-arid Arable Land in Northern China, Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[b] College of Geological Engineering and Geomatics, Chang'an University, Xi' an 710054, China

[c] Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China

ARTICLE INFO

Keywords:
The grid-by-grid model
ENKF
4Dvar
SCE-UA
Winter wheat yield estimation

ABSTRACT

Grain yield estimation remains a focal point in agricultural research. It's well known that crop models have very high accuracy in field application, but their scalability to a regional level encounters formidable constraints attributed to stringent input parameter demands, challenges in data acquisition, and complexities in parameter calibration. In a concerted effort to overcome these aforementioned challenges, this study endevours to formulate a spatialized crop growth model, organized grid by grid, propelled by a myriad of data sources encompassing diverse remote sensing and statistical inputs. Our approach involves the integration of a machine learning technique—the shuffled complex evolution algorithm (SCE-UA) to propose an automatic parameter optimization method for model calibration, alongside two remote sensing assimilation methods: a four-dimensional variational assimilation algorithm (4Dvar) and ensemble Kalman filter (Enkf) to optimising model trajectories to improve crop yield estimation accuracy. This innovative methodology addresses the intricacies associated with regional-scale simulation and bridges the gap between the inherent limitations of conventional crop models and the demand for high-precision yield estimations. The results show that: (1) we improved the accuracy of the regional crop model from 0.53 to 0.94 for the coefficient of determination (R^2) and from 824.82 kg/ha to 148.48 kg/ha for root mean square error (RMSE), which greatly improved the accuracy of winter wheat yield estimation; (2) after comparing different optimization and assimilation strategies, the simulation strategy of complex shuffling algorithm (SCE-UA) combined with the four-dimensional variational algorithm (4Dvar) can enable the grid-by-grid model to estimate yield to achieve the highest simulation accuracy, with R^2 of 0.94 and RMSE of 148.48 kg/ha; (3) we evaluated the simulation effectiveness of the algorithm and discuss the shortcomings and uncertainties of the grid-by-grid model. This study has important practical implications for the development of spatialized models for estimating winter wheat yields and bolstering our capacity for informed decision-making in the realm of food production and agricultural management.

1. Introduction

The estimation of crop yields holds paramount significance in ensuring sustainable agricultural development and global food security (Ines, Hansen, and Robertson 2011). It facilitates governmental agricultural decision-making, optimizes resource allocation, and supports agricultural practices to enhance productivity (Zambrano et al. 2018). In the field of crop yield estimation study, the application of statistical models to estimate yields was dominant, which usually developed from regression models and agrometeorological models to establish relationships between crops and meteorological factors (e.g., temperature, precipitation) (Doraiswamy et al. 2005; Qian et al. 2009). This kind of statistical method needs a large number of field observation data to calibrate models but is only suitable for specific study areas, crop varieties, specific crop growth environments, and management strategies (Huang et al. 2015; Paola, Valentini, and Santini 2016). In recent decades, crop yield estimation technology developed more comprehensively with data sources diversity (e.g., field data database, statistical

* Corresponding author.
E-mail address: gaomaofang@caas.cn (M. Gao).

https://doi.org/10.1016/j.compag.2024.109594
Received 13 December 2023; Received in revised form 23 October 2024; Accepted 24 October 2024

data, remote sensing data) and method innovation (e.g., process-based crop growth models (CGM), crop growth environment-based models, machine learning and deep learning methods and remote sensing methods) (Roudier et al. 2016; Ceglar et al. 2018; Lecerf et al. 2018; Zhuo et al. 2019; Wu et al. 2021; Dhakar et al. 2022).

The process-based crop growth model, through decades of development, has proven to be an effective way for crop yield estimation by inputting precise weather, soil, crop and management parameters. Some of the best-known current crop models have had relevant applications in crop estimation, such as the Decision Support System for Agro-technology Transfer model (DSSAT), World Food Studies (WOFOST) and Agricultural Production Systems Simulator (APSIM), etc. Due to the difficulty in obtaining inputs (high cost, low accuracy), the application of crop models is mainly focused on a field scale, which constrains the application of crop models' spatialization (Doraiswamy et al. 2004). With the development of remote sensing technology, the improvement of remote sensing inversion accuracy and the emergence of a large number of high-precision remote sensing products have provided a reliable data source for spatialized crop models (Dumont et al. 2015; Huang et al. 2019). The integration of remote sensing data and crop growth-related statistics constitutes a multi-source input parameter dataset for crop models that can significantly improve the accuracy of crop models in yield estimation (Basso and Liu 2018). Integrating remote sensing data and crop growth models has been recognized as the most effective way to crop yield estimation at a large regional scale (Wit and Diepen 2007). Because both the crop growth model and remote sensing have uncertainty, the remote sensing assimilation method was put forward to decrease uncertainty and increase simulation accuracy.

Sequential and variational are the two most common methods applied in remote sensing assimilation. The most typical method of sequential strategy is the Ensemble Kalman Filter (EnKF), which corrects the minimizes and trajectories the uncertainties of state variables by combining crop growth model and remote sensing data. During the forward run of the crop model, remote sensing data are continuously incorporated to improve the simulation accuracy of the crop model by making the trajectory of the crop model more closely match the actual conditions of crop growth. The most typical methods of variational approach are Three-Dimensional Variational Data Assimilation (3DVar) and Four-Dimensional Variational Data Assimilation (4DVar), which reinitialize the crop growth model input parameters to optimize a given criterion by the minimization of a cost function to improve the crop model simulation accuracy. For example, Huang et al. (2016) used the EnKF algorithm to assimilate 30-m resolution leaf area index (LAI) time series remotely sensed data into the WOFOST model, improving the winter wheat yield estimation accuracy (Huang et al. 2016). Wu et al. 2021 based on the 4DVar algorithm referred to an improved variational approach named anisotropic background error and time window four-dimensional variation (ABT-4DVar) with WOFOST model to estimate winter wheat yield, which had good performance in regional yield estimation (Wu et al. 2021). Besides the WOFOST model, other mainstream crop growth models around the world have been studied in the field of data assimilation. For example, Ines et al. (2013) applied a bi-variate (soil moisture (SM) and LAI) assimilation into the DSSAT model by EnKF to estimate maize yield, which obtained the result that bi-variate assimilation was better than the single variate assimilation (Ines et al. 2013). Jin et al. (2017) assimilated multi-source medium-resolution optical and radar imaging into the AquaCrop model to estimate winter wheat yield (Jin et al. 2017). Machwitz et al. (2014) used a particle filter as an assimilation method, assimilating RapidEye image into the APSIM model and a radiative transfer model to predict maize yield (Machwitz et al. 2014). In light of the above, remote sensing assimilation studies combining crop models and remote sensing data have been developed with considerable advancement in the last two decades (Jin et al. 2018).

Remote sensing assimilation using a single parameter is called single-parameter assimilation, and when two or more parameters are used for assimilation, it is called multi-parameter assimilation. While multi-parameter assimilation may enhance accuracy, it can also introduce heightened algorithmic complexity. In investigations concerning crop yield, LAI often serves as the primary assimilation parameter due to its pivotal role in crop growth and yield conditions. Conversely, research on crop drought may choose a two-parameter assimilation approach, incorporating LAI and SM, to delve into both crop yield dynamics and the influence of soil moisture on vegetation water content (Huang et al., 2019a,b). Regardless of which crop models were used, which assimilation methods were adopted, and some new assimilation methods proposed based on the traditional assimilation methods, a high accuracy has been achieved in estimating and predicting crop yields, which is sufficient to meet the needs of macro-guided agricultural production.

The remote sensing assimilation method, although it solves the problem of model uncertainty in the simulation process still has lots of room to develop. In particular, the application of machine learning and deep learning methods to crop yield estimation demonstrates very high accuracy, which is a hot point in the field of yield estimation (Paudel et al. 2020; Feng et al. 2020). Because of the requirement for an increase in production, there is a necessity to apply machine learning to estimate crop yield so that necessary action can be taken to increase crop yield (Akkem, Biswas, and Varanasi 2023). Most of the existing machine learning methods for estimating yield only consider the non-linear relationship between environmental factors such as temperature, precipitation, soil, etc. and yield, but it is difficult to explain why these factors have an impact on crop yield mechanistically (Praveen and Sharma 2020; Iniyan and Jebakumar 2022). This is a shortcoming of such studies of single machine-learning methods for estimating crop yields. Crop modelling can tell you mechanistically how environmental factors affect crop yields. Over time, crop models can simulate how a crop grows step-by-step in a given environment, ultimately achieving the desired yield. Cao et al. (2022) demonstrated that stacking with multiple models provides better accuracy when compared with a single model (Cao et al. 2022). Coupling machine learning algorithms are a trend in the development of next-generation crop models (Hwang et al. 2017; Huang et al. 2019a,b). Therefore, machine learning methods coupled with crop models may further enhance the accuracy of crop models.

To improve winter wheat yield estimation accuracy, a grid-by-grid regional crop model was constructed, integrating remote sensing assimilation and machine learning methods. This article achieves that aim step by step: (i) collection of multiple raster data sets to build a grid-by-grid model with multi-source data coupling; (ii) based on the grid-by-grid model, a kind of machine learning methods (SCE-UA) was added into the spatialized crop model to optimize crop parameters; (iii) based on the optimized crop model, two kinds remote data assimilation methods were applied to further improve simulation accuracy; (iv) finally, the results of winter wheat yield simulation under each simulation strategy were validated to select the optimal winter wheat yield estimation method.

2. Materials and methods

2.1. Study area

We selected Hebei province, China as the study area because it is located in the North China Plain, which is a typical winter wheat production area in China. Hebei Province has a typical temperate continental monsoon climate, with an average temperature is 15 °C, an average annual precipitation of 484.5 mm, and soils mainly alluvial, which are suitable for winter wheat cultivation (https://www.hebei.gov.cn/hebei/14462058/14462085/14471224/index.html). Winter wheat is mainly distributed in the south of Hebei Province, because the terrain in the northwest of Hebei Province is high, with many plateaus and mountains, and the latitude is high, with a cold and dry climate, which is unfavourable for winter wheat cultivation. The terrain of the

Hebei Plain in the southeast is low, with many rivers merging into the sea, forming a large area of alluvial plains, which is conducive to the cultivation of winter wheat (Fig. 1). Winter wheat in Heibei Province usually planted on September or October and harvested on mid-June (Fig. 2).

2.2. Technical route

We estimated winter wheat yield in Hebei Province from 2010 to 2020 with a spatial resolution of 1 km and a simulation step size of days. We proposed four simulation/optimisation/assimilation strategies (Fig. 3) to explore the combination of machine learning algorithms and remote sensing assimilation methods on the grid-by-grid spatialized crop model, whose aim is to improve the accuracy of regional modelling for winter wheat yield estimation. (i) The reference crop parameters, as well as meteorological, soil, and management parameters, are input directly into the crop model without any methodological processing. Then, we obtained the original winter wheat yield simulation result. (ii) Using the ten-year average statistical yield data as a benchmark, we used a machine learning approach, the Shuffled Complex Evolution Algorithm (SCE-UA), to recalibrate crop parameters. The calibrated crop parameters were input into the grid-by-grid model along with meteorological, soil, and management data. (iii) Based on (ii), the MODIS leaf area index (LAI) product was applied for sequential remote sensing assimilation using the Ensemble Kalman filter (ENKF) to intervene in regional model simulation trajectories. (iv) Based on (ii), constructing a 4Dvar cost function based on MODIS LAI and again using the SCE-UA algorithm to recalibrate the crop parameters to obtain 4Dvar assimilated yield estimation. We matched the county-scale statistics by averaging the yields of all grids within each county for calibration and verification. Finally, we compared and analysed four winter wheat yield estimation results.

2.3. Grid-by-grid regional crop model construction

2.3.1. Fundamental model

The Python version of the WOFOST model (PCSE 5.4.2, Wofost72_WLP_CWB) was selected as the foundational model for building the spatialized grid-by-grid model. This decision was based on the availability of a Python implementation of the WOFOST model, distinguishing it from the DSSAT and APSIM models. Leveraging its Python compatibility, the WOFOST model facilitates seamless integration with open-source resources and various Python packages, including those for machine learning algorithms and remote sensing tools. The WOFOST model is a photosynthesis-driven crop growth model that is capable of simulating the crop growth process under a wide range of natural and anthropogenic conditions and calculating the final yield at maturity, and its simulation accuracy has been verified by scholars in various countries and regions (Laar et al., 1992). To transform a field model into a spatialized model, we defined a grid as the smallest simulation unit and constructed a spatial database of the regional crop model. Various remote sensing data and regional multiple sources are collected according to the requirement of PCSE (climate data, crop data, soil data and management data), which aim is to realize the simulation results with high precision, and time continuity. After PCSE version 5.2, the WOFOST model provides a simple module to simulate N/P/K

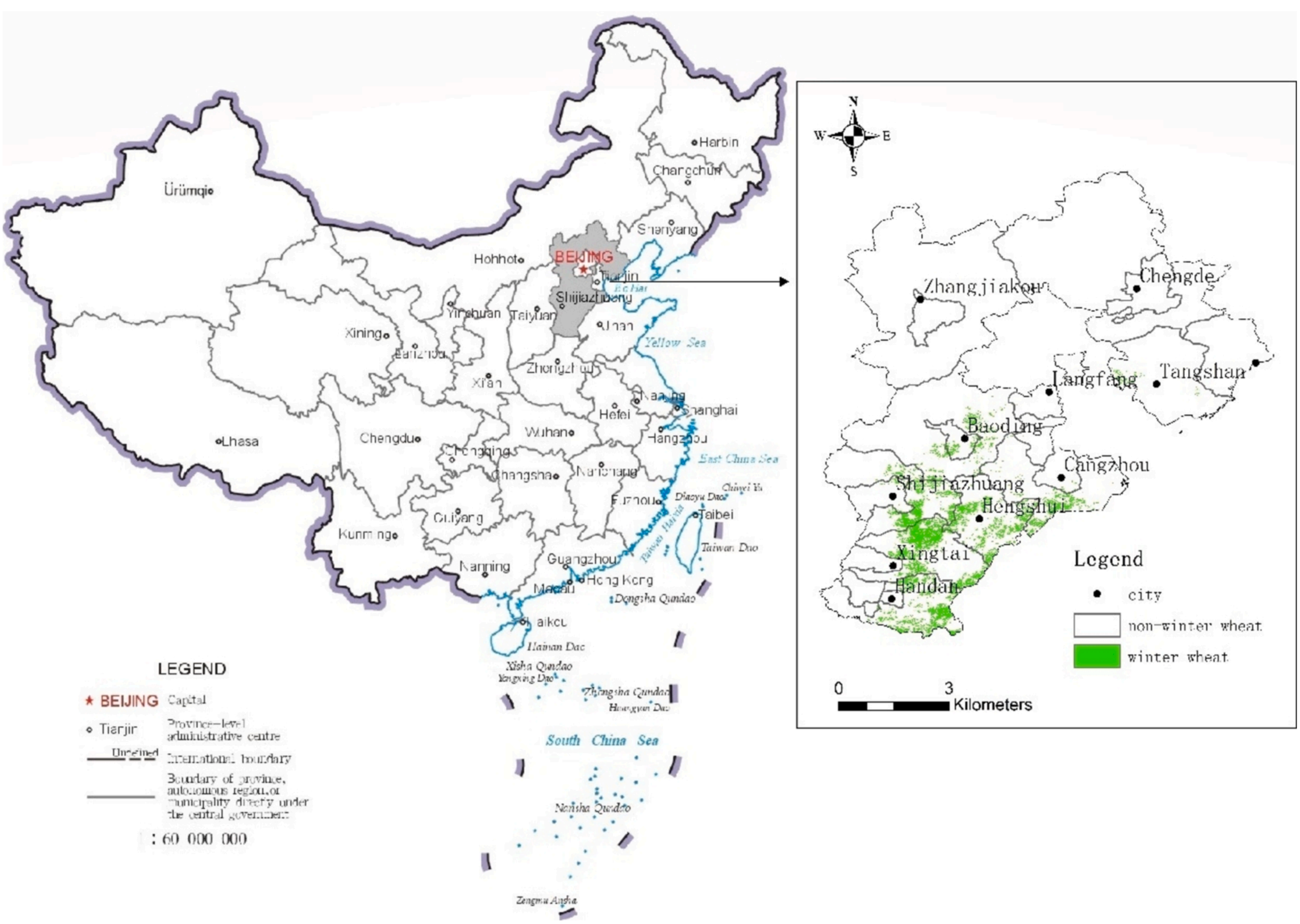

Fig. 1. The location of the study area (winter wheat distribution above is for 2020).

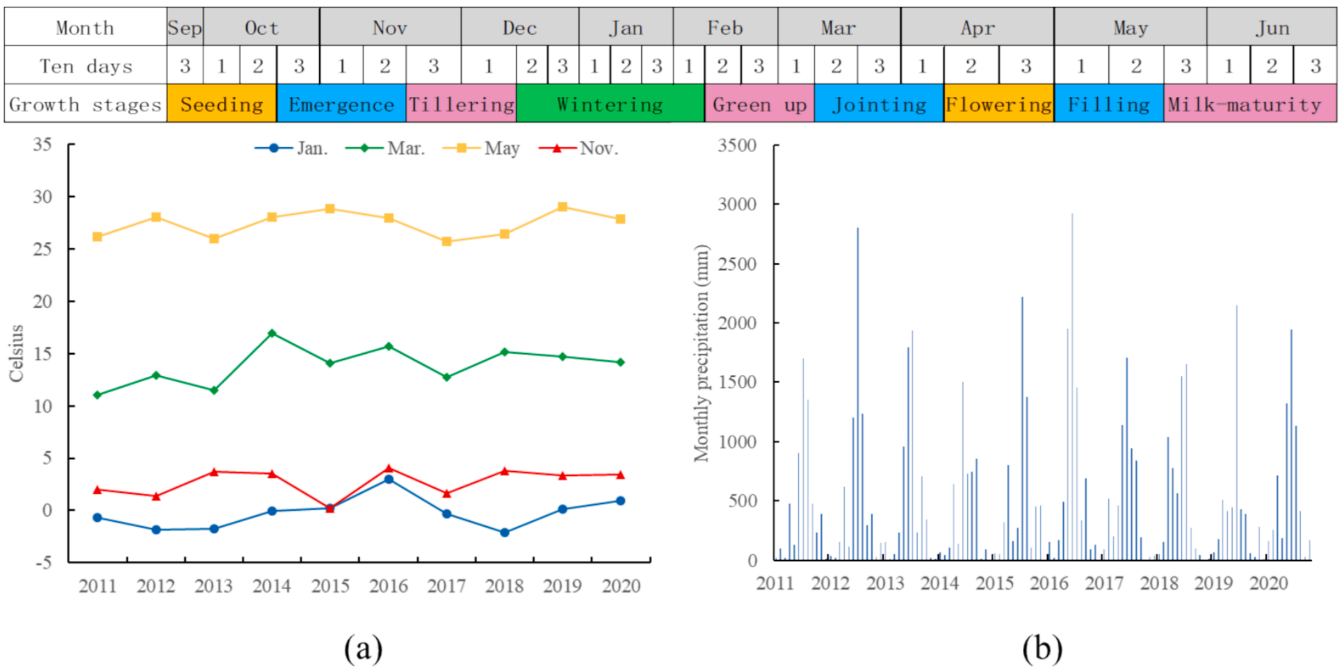

Fig. 2. Average temperature for the main winter wheat growing season (a), monthly precipitation (b) and winter wheat phenology in Hebei Province.

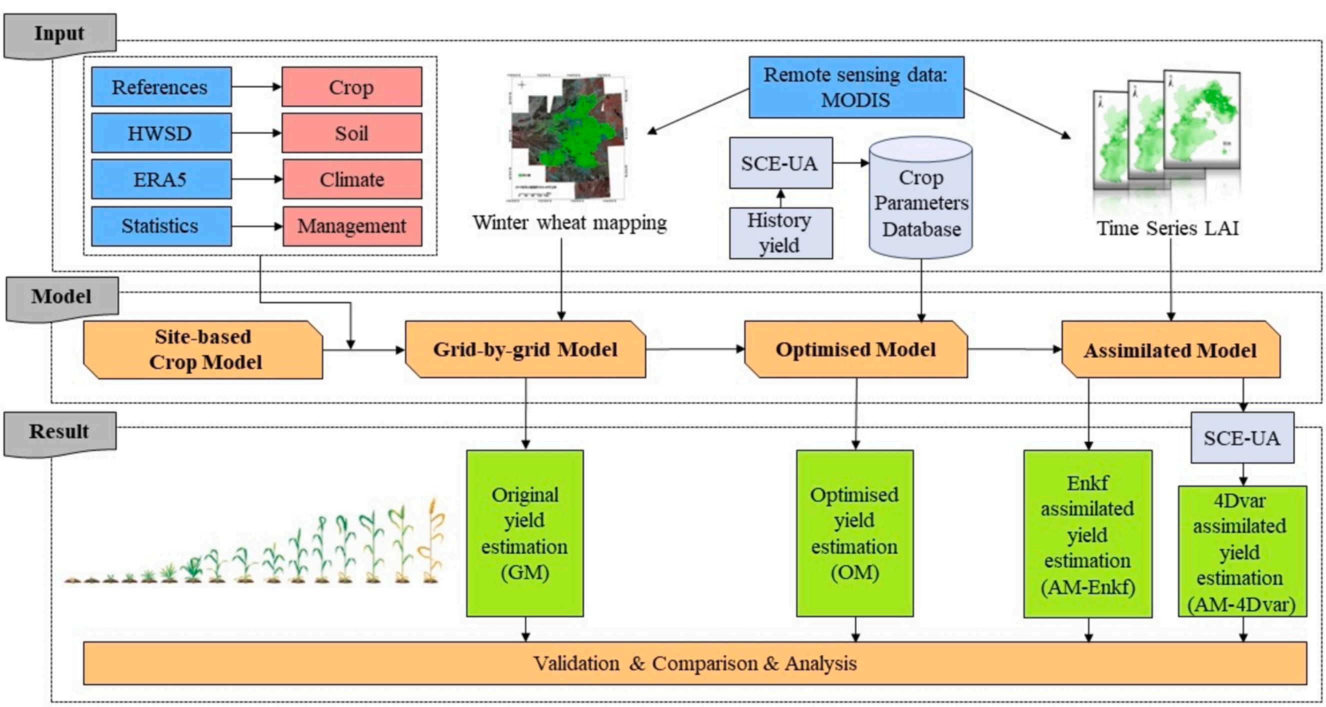

Fig. 3. Technology roadmap.

limitations on crop growth. We used this module to simulate the effect of N-fertilizer on yield.

2.3.2. Grid-by-grid dataset

High-precision extraction of winter wheat cultivation area is a prerequisite for winter wheat regional model construction, directly affecting the study's correctness and model simulation accuracy. We used the published winter wheat classification dataset from the Chinese Academy of Agricultural Sciences (CAAS). For this dataset, the 16-day synthetic NDVI data MOD13Q1 at 250 m under MODIS was the main data source, and the threshold method was used for winter wheat planted area extraction. The spatial resolution of this data set was 1 km, and the overall accuracy of the results was 87.4 % with a kappa coefficient of 0.61.

We selected the ERA5 dataset as the meteorological database. However, some meteorological parameters required by the PCSE are not available in the ERA5 dataset but can be converted by other parameters in ERA5. Water vapour pressure could be calculated from 2 m dew point temperature in ERA5 data (Liao et al. 2020):

$$e_a = a \times \exp(\frac{b \times T_d}{T_d + c}) \quad (1)$$

where e_a is the water vapour pressure (kPa), T_d is the dew point temperature (∘C), a, b and c coefficients take the values of 0.611 kPa, 17.502 and 240.97∘C respectively. The 2 m wind speed at ground level can be transferred from the 10 m wind speed at ground level (Liao et al. 2020):

$$u(z) = \frac{u_*}{k}\ln(\frac{z-d}{z_0}) \quad (2)$$

$$d = 0.65h \quad (3)$$

$$z_0 = 0.1h \quad (4)$$

where z_0 and d are physical quantities describing the aerodynamic properties of the surface, h is the height (m), and k is von Karman's constant (dimensionless, typically between 0.39 and 0.41). After calculation, the surface 2 m wind speed and 10 m wind speed conversion relationship are:

$$u_2 = 0.35u_{10} \quad (5)$$

In terms of temporal resolution, the grid-by-grid model requires meteorological data in daily steps as meteorological input parameters, whereas the ERA5 data are hourly data, which need to be synthesized from the ERA5 hourly data into daily scale data. The hourly climatic data are filtered to the maximum and minimum values as the daily maximum and daily minimum temperatures for the PCSE temperature data; the hourly precipitation is summed up as the daily precipitation for each day; the hourly radiation is summed up as the total daily radiation; the hourly water vapour pressure is averaged out as the daily mean water vapour pressure; and the hourly wind speed is averaged out as the daily mean surface 2 m wind speed. The temporal resolution of the ERA5 data was resampled to one day using the above methods.

Besides, we collected the gridded database about soil and management data. Because the spatial resolution of each dataset is not the same, we reclassified the spatial resolution of all datasets to 1 km to ensure that the number of grid rows and columns and the geographic coordinate system are uniform for each dataset. Regarding precipitation, we investigated the irrigation areas and irrigation habits in Hebei Province and added the irrigation coefficients to the grid-by-grid model. Details of these data are shown in Table 1.

The MODIS LAI product (MCD15A2H) was chosen as the remote sensing data for assimilation. This product offers an 8-day composite at a resolution of 500 m. The data were selected from the time when wheat entered the rejuvenation stage, so that the selected data dates were distributed evenly from the rejuvenation stage to the maturity stage of wheat, and the LAI time series was obtained for winter wheat in Hebei from 2010 to 2020 in March to June each year (Fig. 4).

Table 1
Spatial database of grid-by-grid regional crop model.

Abbr. name	Full name	Time resolution	Spatial resolution	Parameters	Reference
–	Winter wheat map from CAAS	Year	1 km	winter wheat map	(Yanyan, Qing, and Jianqiang 2020)
ERA5	–	hour	0.1°×0.1°	2 m temperature, 2 m dew point temperature, 10 m wind speed, precipitation, solar radiation	(Olauson 2018)
HWSD	Harmonized World Soil Database	–	1 km	soil texturesoil bulk weight, wilting coefficient, field water holding capacity, saturated soil water content and saturated hydraulic conductivity	(Shi et al. 2004)
–	China Statistical Yearbook	year	county area	fertilization, irrigation, phenology	–

*1) The average application of N fertilizer in the counties of Hebei Province was 96.6 ± 12.83 kg N/ha (active ingredient). 2) The irrigated area was irrigated 2 ~ 4 times between March and June. The average single irrigation was 3.45 ± 1.64 cm. The rainfed areas were not irrigated.

2.4. Optimization and assimilation methods

2.4.1. Initial crop parameters calibration and sensitivity analysis

Initial crop parameters calibration is one of the most important things in related crop model study because these crop parameters have a great impact on the final simulation results of the crop growth model that they need to be sensitivity analysed and calibrated. The initial ranges of crop parameters were mainly derived from related literature with the same study area (Huang et al. 2015; Wu et al. 2021) and model default values (Appendix 1).

The Sobol sensitivity analysis method was chosen to analyse the parameter sensitivity of the PCSE, the results of the sensitivity analysis are used to help with parameter optimization. The Sobol is a kind of Monte Carlo analysis method based on variance which can reflect the influence level of parameter changes to model simulation results (Sobol and Kucherenko. 2005; Yang 2011). The Sobol methods can decompose the PCSE into a sum of incremental terms, calculate the total and each order variance in the model by sampling, and then calculate the sensitivities of multiple input parameters. We put the sensitivity analysis result in the PCSE with the simulation LAI_{max} here because the following Shuffled Complex Evolution Algorithm (SCE-UA), a machine learning parameter optimization method, was based on it. In Fig. 5, the first 14 most sensitivity parameters for PCSE were listed in order.

2.4.2. Optimized model estimate yield via SCE-UA

Parameter localisation is an important step in model simulation. The crop parameter database in the study was reconstructed on a grid-by-grid basis based on historical average yields (2000 ~ 2009) by the SCE-UA algorithm, whose aim is to match the parameters of the crop model to the productivity of the study area. The average yield of winter wheat in Hebei Province from 2000 to 2009 was 6158.58 ± 141.24 kg/ha. The SCE-UA method is a kind of automatic model optimization algorithm, which was first applied in the hydrology model (Duan, Sorooshian, and Gupta 1994). SCE-UA sets the crop parameters to be adjusted in the crop model into a dataset containing many sampling points according to the range of parameter values and the adjustment step. It samples generates sample points randomly in the feasible parameter space and computes the criterion value at each point. In the absence of prior information on the approximate location of the global optimum, use a uniform probability distribution to generate a sample. Shuffle complexes combine the points in the evolved complexes into a single sample population, and sort the sample population in order of increasing criterion value. If the newly generated point is better than the worst point, replace the worst point with the new point. If any of the pre-specified convergence criteria are satisfied, stop; otherwise, continue. The following are three pre-specified convergence criteria for this study: (1) the initial parameters of the model have converged to within a predetermined small range after 5 successive iterations; (2) the cost function value does not improve by more than 0.1 % after 5 successive iterations; or (3) the objective function has been calculated more than

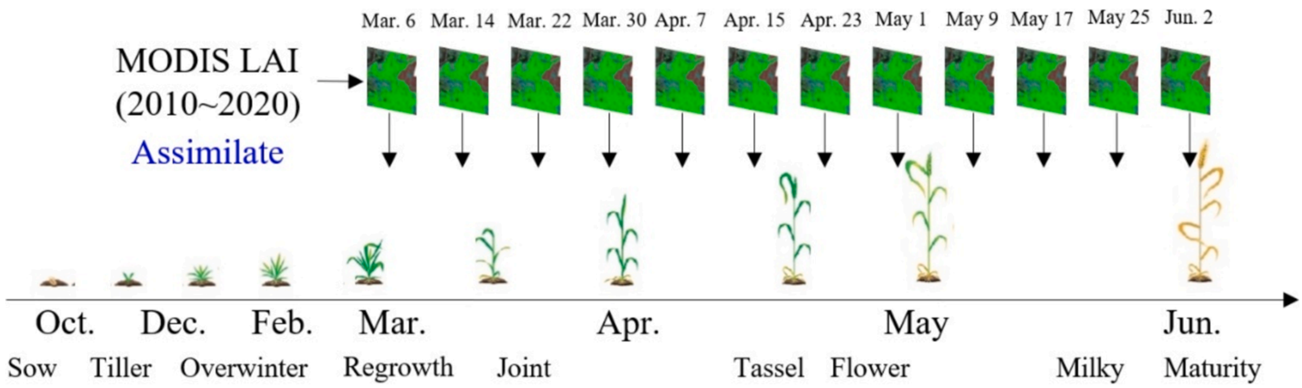

Fig. 4. MODIS LAI data (MCD15A2H) for remote sensing assimilation.

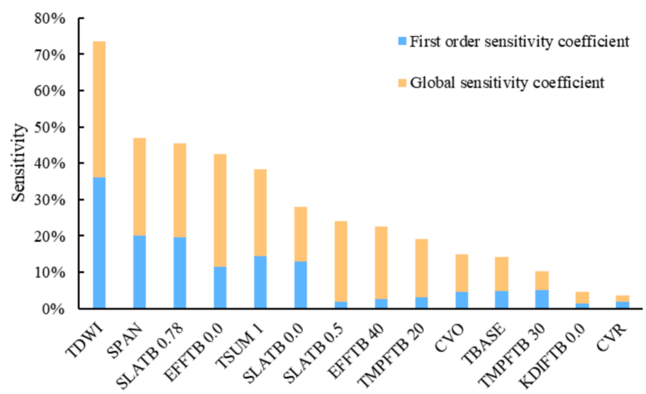

Fig. 5. Sensitivity analysis results of crop parameters in the PCSE with the simulation LAI_{max}.

100 times. The three forced exit loop settings above are configured to safeguard against non-convergence or overfitting within the constructor function. The maximum number of cycles is defined as 100 based on the results of our tests.

Up to 9 most sensitivity parameters for PCSE were included for optimization and we set the lower/upper bound and sample size according to both PCSE and SCE-UA requirements (Table 2). The historical average winter wheat yields in various counties of Hebei Province were used as objective functions because county-scale yield data are the most accurate statistics available in China. Based on the spatial heterogeneity of individual grids, winter wheat crop parameters were calibrated for each grid, which constructed the optimized crop dataset. We brought the optimized crop dataset into the grid-by-grid model to obtain winter wheat yield results from the optimized model.

Table 2
Setting of parameter optimization for Optimized model.

Rank	Sensitivity/ %	Parameter	Lower bound	Upper bound	Sample size
1	73.5	TDWI	180	330	0.5
2	47.0	SPAN	30	60	1
3	45.5	SLATB 0.78	0.004	0.006	0.001
4	42.5	EFFTB 0.0	0.2	0.7	0.01
5	38.3	TSUM1	500	700	0.5
6	28.0	SLATB 0.00	0.004	0.006	0.001
7	24.0	SLATB 0.50	0.004	0.006	0.001
8	22.6	EFFTB 40	0.2	0.7	0.01
9	19.0	TMPFTB 20	0	1	0.1

2.4.3. Yield estimation process based on the Enkf

The Ensemble Kalman filter is a sequential assimilation algorithm proposed based on the Monte Carlo method. It was an improved algorithm of the Kalman filter and was first proposed in 1994(Evensen 1994). The set Kalman filter treats the simulation error of the crop model and the error of remote sensing inversion as the data set satisfying the Gaussian distribution, which satisfies the probability distribution of the approximate prior state. Compared with the Kalman filter, the ensemble Kalman filter can deal with nonlinear data sets better, can adapt to the situation that the data does not meet the Gaussian distribution, and can greatly reduce the algorithm complexity of the yield estimation process compared with the variational method. The core formula is as follows:

$$E_0 = E + P_t\left(HP_tH^T + R_t\right)^{-1}(D - HE) \tag{6}$$

$$K = P_tH^T\left(HP_tH^T + R_t\right)^{-1} \tag{7}$$

where E_0 repents the analytical matrix of the state variables set; E repents forecast matrix of state variables set; D is the observation matrix; K is the Kalman gain matrix; P_t repents the error covariance of the background field; H is the observation operator.

Based on the joint probability distribution of TDWI and SPAN, Wit A. et al. showed that there is no obvious correlation between TDWI and SPAN, so in this study, TDWI and SPAN are selected as the source of model uncertainty, and the uncertainty of both of them is set to 10 %, and the uncertainty of MODIS LAI is set to 30 % (Wit et al., 2012). Using the PCSE model, TDWI and SPAN are concurrently Gaussian-perturbed to generate the parameter set Para {Para 1, Para 2, …, Para N}, where N represents the set size (N = 30 in this study). If remote sensing observations are available at time t, simulated LAIs under different model parameter runs are extracted accordingly to generate the forecast set M_{LAI} {M_{LAI} 1, M_{LAI} 2, …, M_{LAI} N}. Simultaneously, observed LAI data undergo Gaussian perturbation to create an observation set of identical size, O_{LAI} {O_{LAI}1, O_{LAI}2, …, O_{LAI}N}. The Enkf algorithm is then employed to assimilate the forecast and observation sets, yielding the assimilated set and synchronizing the forecasts. If no remotely sensed LAI is available, the WOFOST model is executed directly until the winter wheat harvest. This process repeats, culminating in Enkf assimilation results derived from averaging the 30 ensemble members.

2.4.4. Yield estimation process based on the 4Dvar

The Four-dimensional Variational algorithm (4Dvar) assimilation algorithm was proposed in 1987 (Philippe et al. 1987), which is a variational assimilation algorithm. 4Dvar takes into account the change of the background field information over time, and better reflects the complex nonlinear constraints between the model variables. 4Dvar algorithm is a global optimization algorithm, and the core idea is to obtain the optimal solution by defining the objective function within the assimilation window and calculating the gap between the model simulated values and remote sensing observations by obtaining the minimum value of the objective function. The core idea is to define the objective function in the assimilation time window, calculate the gap between the simulated values of the model and the observed values of remote sensing, and obtain the optimal solution by obtaining the objective function. The cost function $D(X_0)$ as constructed as follows:

$$D(X_0) = \frac{1}{2}\left(X_0 - X_0^b\right)^T B^{-1}\left(X_0 - X_0^b\right) + \frac{1}{2}\sum_{i=0}^{n}\left(H(X_t) - X_t^g\right)^T R^{-1}\left(H(X_t) - X_t^g\right) \tag{8}$$

where X_0 is the initial state variable of the assimilation window established by 4Dvar; X_0^b is the background value at the initial moment; X_t is the value at the moment t obtained by substituting X_0 into the model; X_t^g is the remote sensing inversion value at the moment t; B is the error covariance matrix of the crop model simulation; and R is the error covariance matrix of remote sensing observation values.

The SCE-UA algorithm was used again to minimize the cost function between observation values (MODIS LAI) and model-simulated LAI by reinitializing crop parameters in the 4Dvar assimilation model. Based on sensitivity analysis results, three parameters—TDWI, SPAN, and SLATB 0.78 were identified as highly sensitive to both LAI and crop yield. These parameters were selected for reinitialization, and their optimization ranges were aligned with those specified in Table 2. Markov Chain Monte Carlo approach was used to estimate the uncertainties. The error covariance of MODIS LAI was set to 0.01 at the green-up stage (DOY (day of year) 41–69), 0.02 at the jointing stage (DOY 70–100) and 0.04 at the flowering stage (DOY 101–120) according to a related study (Ma et al. 2013; Zhuo et al. 2022). R^2, RMSE and RMSE% were used to evaluate model accuracy (Appendix 3).

3. Results

3.1. Yield description of simulation results

The split-half violin plot serves as a graphical depiction of the data distribution state. In Fig. 6, the left half of the plot illustrates the statistical data of winter wheat yield in Hebei Province from 2010 to 2020, while the right half showcases the simulated data under each distinct simulation strategy. The assessment of simulated yield accuracy hinges on the symmetry observed between the left and right halves of the graph.

For the original grid-by-grid model (GM), as depicted in the violin plot, the average simulated yield (5782.8 kg/ha) surpasses the average of statistical data (5140.2 kg/ha). This discrepancy underscores the overestimation inherent in the grid-by-grid model's yield simulations. Despite utilizing a meticulously calibrated set of crop parameters specific to the study area, the accuracy of large-area simulations over an extended time series remains a challenge. This motivates the transition to a grid-by-grid model, acknowledging the complexity of improving regional simulation accuracy with a single set of crop parameters.

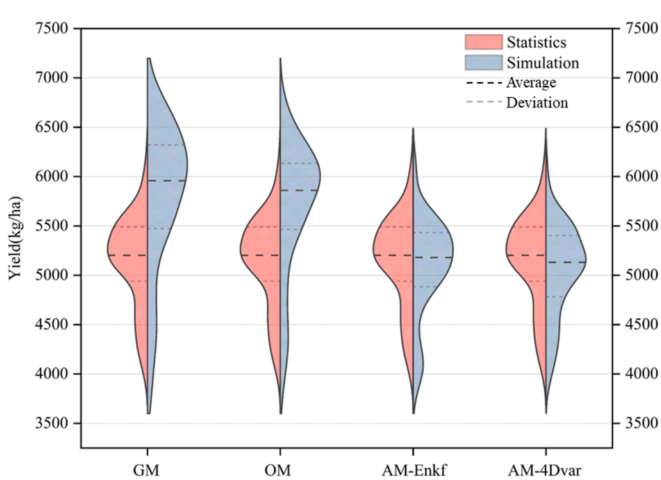

Fig. 6. Data description of winter wheat yield under different simulation strategies.

After optimizing the grid-by-grid model using the SCE-UA method (OM), the average simulated yield shows a slight decrease compared to the unoptimized model. This decline suggests improved accuracy in yield simulation with SCE-UA optimization, effectively addressing the overestimation typical in the regional model.

Integrating the remote sensing assimilation method into the optimized model further enhances simulated yields. When AM-Enkf combines with the optimized model, the average simulated yield notably decreases, leading to improved symmetry between the left and right sides of the violin plot. This improvement underscores the significant enhancement in yield simulation accuracy achieved by Enkf.

Subsequent assimilation of the optimized model using 4Dvar (AM-4Dvar) results in a further decrease in the average simulated yield. This reduction indicates that 4Dvar provides superior simulation results compared to Enkf when integrated with the optimized model. The symmetrical alignment of the left and right sides of the violin plot confirms enhanced accuracy in the simulation outcomes. These findings collectively emphasize the iterative refinement achieved through the integration of optimization and assimilation methodologies. They underscore the importance of these enhancements in ensuring robust and accurate yield simulations.

3.2. Assimilation MODIS LAI into the grid-by-grid model

In Fig. 7, the LAI curve, whether simulated, assimilated, or observed, adeptly captures the climatic nuances of winter wheat across its growth stages. The LAI value, near zero before the greening period, ascends post-greening, peaks during flowering, and gradually declines into the ripening period. Importantly, this LAI curve faithfully mirrors the climatic patterns characteristic of winter wheat in Hebei Province.

Comparing the grid-by-grid model (GM) and optimised model (OM) results for each year revealed significant discrepancies when crop parameters were not automatically optimized using the SCE-UA algorithm. Notably, overestimation/underestimation phenomena were evident in 2011, 2012, 2016, 2017, 2018, and 2020 (corresponding to Figures (a), (b), (f), (h), and (i), respectively). However, post-SCE-UA algorithm optimization, the simulation outcomes were markedly refined, showcasing a substantial improvement in the agreement between simulated crop growth trajectories and actual conditions. The LAI values, when optimized, are closely aligned with observed values, indicating enhanced parameter quality and simulation accuracy.

As depicted in Fig. 7, AM-Enkf and AM-4Dvar demonstrated closer proximity to observed values compared to OM. The latter exhibited significant deviations from assimilated LAI values, suggesting overestimation or underestimation in the regional model's LAI simulations before sequential assimilation. Post-sequential assimilation, the trajectory of the grid-by-grid model adjusted, narrowing the gap between assimilated and observed values, thus enhancing accuracy.

Fig. 7 highlights that both AM-Enkf and AM-4Dvar assimilation methods outperform OM. However, AM-4Dvar assimilation yielded results closer to observed values than AM-Enkf emphasizing the efficacy of the four-dimensional variational assimilation algorithm in the grid-by-grid model. Whether employing the ensemble Kalman filter or the 4Dvar algorithm, both methodologies enhanced simulation accuracy, aligning the model trajectory with observed values. This approach effectively reduces uncertainty in the regional model, illustrating the merit of assimilation post-automatic optimization of crop parameters using the SCE-UA algorithm.

Fundamentally, the remote sensing assimilation paradigm considers observed values of remote sensing data as ground truth to guide crop

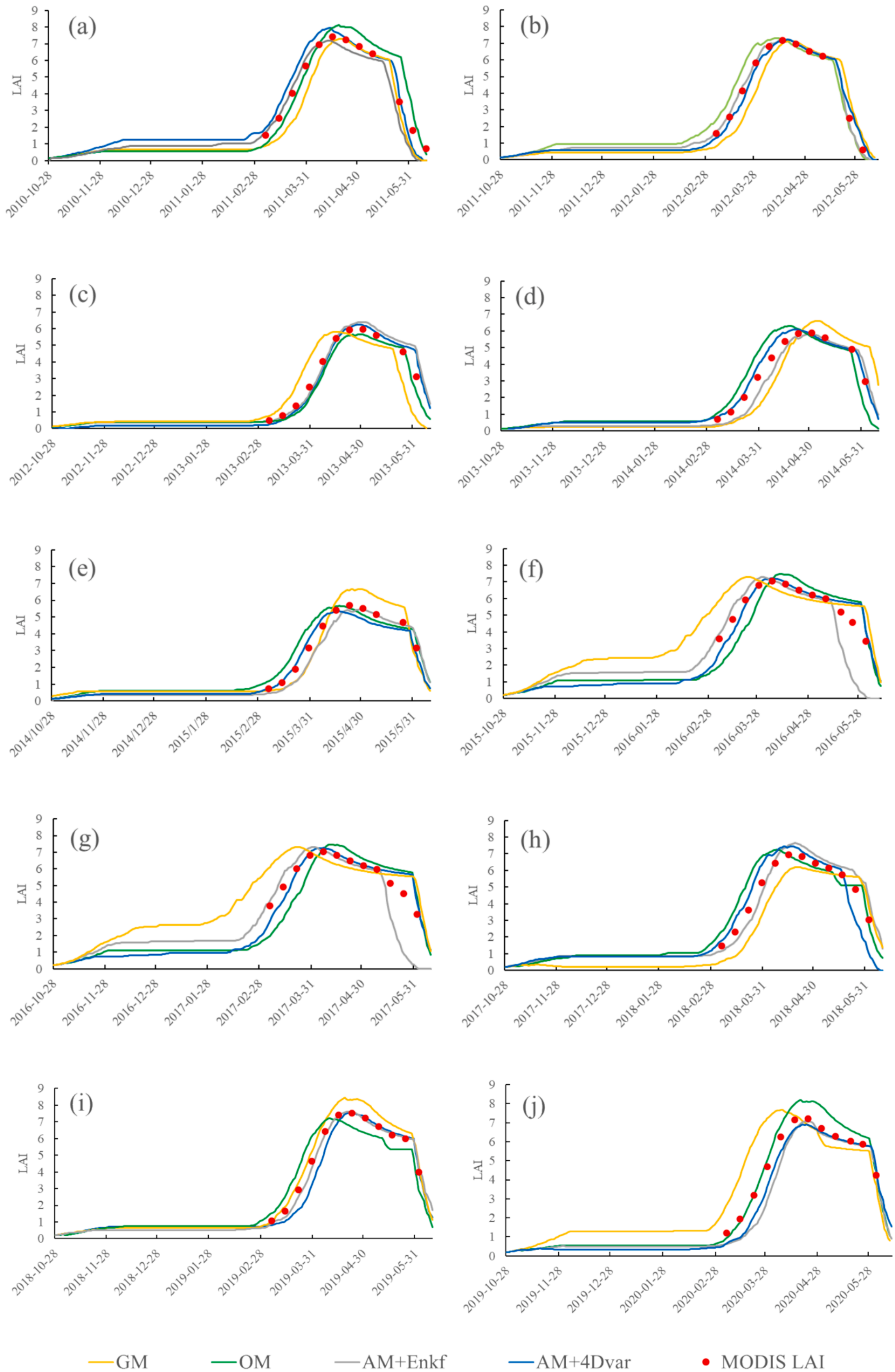

Fig. 7. Comparison of LAI simulated values, assimilated values and observed values before and after different simulation/assimilation methods for winter wheat ((a-j) represent the winter wheat cropping period from 2011 to 2020, respectively.).

model trajectories. In this study, MODIS leaf area index (LAI) values were treated as absolute and correct, intervening in the model trajectory with the assumption of 100 % accuracy in LAI inversion. Fig. 8 illustrates the correlation between LAI results from site observations and MODIS values. The validation, using discontinuous station LAI, revealed a strong correlation post-data assimilation, with R^2 increasing to 0.84 at the site scale and 0.90 at the regional scale. Importantly, the regional scale R^2 surpassed the site scale, potentially attributed to scale differences in the study.

3.3. Winter wheat yield estimation verification

The accuracy of winter wheat grain yield simulations in Hebei Province for the period 2011 to 2020 underwent thorough validation against statistical data, utilizing the county as the smallest analytical unit. Statistical yield data, sourced from the statistical yearbooks of each city in Hebei Province, formed the basis for this validation. The model's simulated values, assimilated values, and observed values at each grid were meticulously aggregated based on distinct years and administrative units of counties. Finally, we got 152 valid validation points. Fig. 7 presents the outcomes of this comprehensive yield validation.

Upon OM, as depicted in Fig. 9(a-b), a discernible enhancement in simulation efficacy was observed compared to GM. The coefficient of determination (R^2) surged from 0.53 to 0.84, concurrently with a reduction in the root mean square error (RMSE) from 824.82 to 477.96 kg/ha. The normalized RMSE (NRMSE) exhibited a decrease from 16.04 % to 9.03 %, indicating a robust alignment between statistical and simulated values, regardless of parameter optimization using the SCE-UA algorithm.

After OM and AM-Enkf, illustrated in Fig. 9(c), the model's simulation further improved. The R2 increased to 0.87, the RMSE decreased from 477.96 kg/ha to 256.04 kg/ha, and the NRMSE diminished from 9.03 % to 4.98 %.

Furthermore, OM and AM-4Dvar, illustrated in Fig. 9(d), a notable refinement in the model's simulation emerged. The R2 further increased to 0.94, the RMSE decreased from 256.04 kg/ha to 148.48 kg/ha, and the NRMSE reduced from 4.89 % to 2.89 %. Among the diverse simulation methodologies explored, the AM-4Dvar approach demonstrated the highest accuracy in simulating winter wheat yield in Hebei Province.

3.4. Winter wheat yield mapping

Following a comprehensive assessment of various spatialized crop model simulation strategies, the AM-4Dvar emerged as the optimal optimization-assimilation method for the grid-by-grid model. The application of this strategy produced the winter wheat yield map for the period 2011 to 2020, as illustrated in Fig. 10. Subsequent analysis of the simulated data revealed an average winter wheat grain yield of 5840.69 ± 467.27 kg/ha over the specified timeframe.

The highest simulated grain yield was observed in Linxi County, Xingtai City, Hebei Province, in 2011, reaching 7132.35 kg/ha. In contrast, the lowest simulated yield occurred in Gucheng County, Hengshui City, Hebei Province, in 2018, with a value of 3728.79 kg/ha.

Fig. 11 presents the temporal evolution of winter wheat production in Hebei Province from 2011 to 2020. Using the baseline yield in 2011 (6086.4 kg/ha), fluctuations in subsequent years are evident. Notably, a decrease of 5.5 % (5822.6 kg/ha) was observed in 2015, followed by a further decline of 2.81 % (6229.2 kg/ha) in 2018. Conversely, 2017 witnessed a surge in production, reflecting an increase of 7.78 % (6409.5 kg/ha). These fluctuations in production are intricately linked to natural disasters or policy regulations.

Spatially, the southeastern regions of Hebei Province demonstrated a consistent upward trend in winter wheat production (Slope > 0). In contrast, some northern and western areas experienced a slight reduction in production (Slope < 0). Despite localized reductions, the overall trend over the past decade indicates a net increase in winter wheat production in Hebei Province, underscoring the resilience and growth of this vital agricultural sector.

4. Discussions

4.1. Improvement of SCE-UA efficiency by adjusting algorithm setting

To further improve the simulation accuracy of the regional crop model, this study uses the SCE-UA algorithm to automatically adjust the parameters of the model. However, this machine learning algorithm is inefficient to run since it needs to call the crop model several times for adjustment, and it is impractical to optimise each input parameter as the PCSE model has hundreds of input parameters. Therefore, to improve the running efficiency of the regional crop model, taking into account the simulation accuracy improvement, this study conducted a sensitivity analysis of the crop parameters of the model before parameter optimisation and selected the crop parameters with the top 9 sensitivity rankings for optimisation. All settings of the algorithms should be adjusted when dealing with optimization problems. The setting of the algorithms for this study is depicted in Table 3.

The average grain yield of Hebei Province from 2000 to 2009 was used as the tutor for parameter optimisation. The crop parameters for

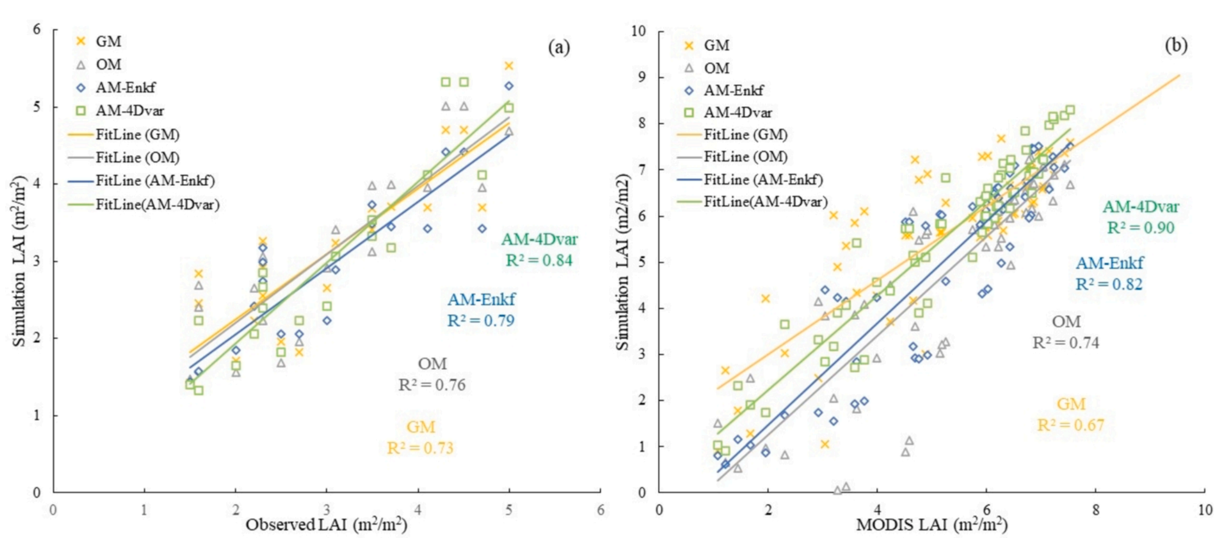

Fig. 8. Correlation between LAI results at the site observation and MODIS value.

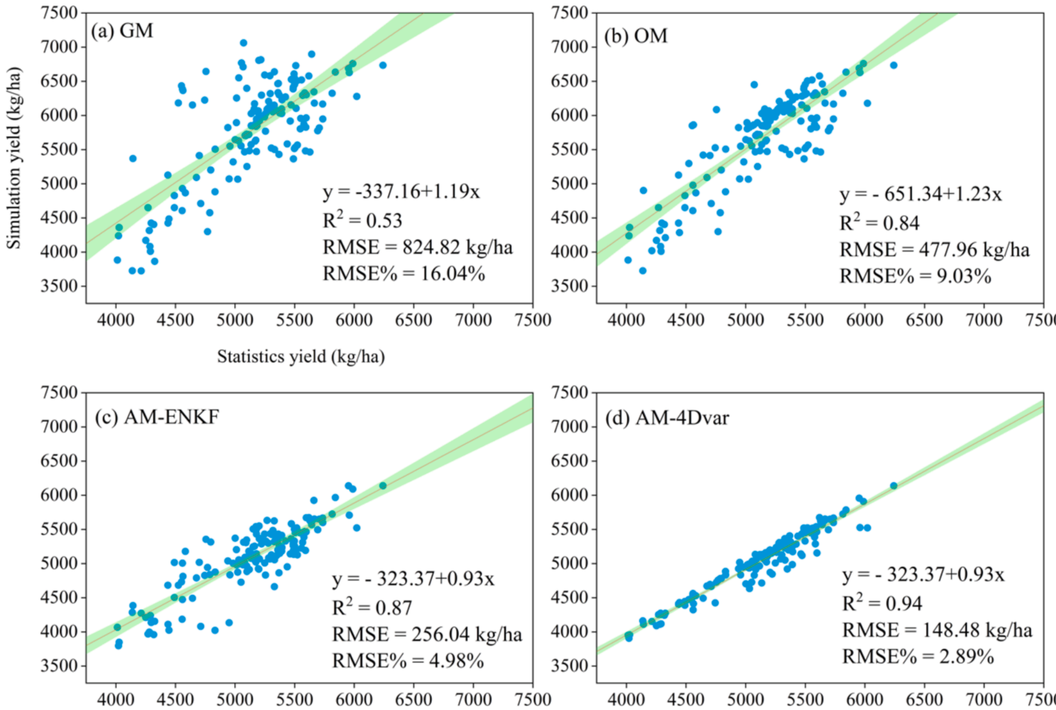

Fig. 9. Verification results of crop parameters with four simulation strategies.

each grid will be recalibrated based on the variability of the grid. In the crop parameter optimization process, the climate data is the 10-year average meteorological data of each grid point, while the soil parameter is regarded as a stable constant. We set up experiments with different maximum numbers of loops to investigate how this parameter should be defined for area crop models (Table 4). We found that when the maximum number of simulations was set to 100, the model was able to balance accuracy and efficiency.

The maximum number of optimisations for each grid was set to 100 times, and when the maximum number of loops was reached, the optimisation of this grid was ended. When the loop ends, the set of parameters with the minimum deviation of grain yield from the tutor is selected as the optimal crop parameters for this grid point. When crop parameters are optimised for all grids, the crop parameters for each grid are stored as a crop parameter database.

The efficiency of the SCE-UA algorithm is shown in Fig. 12. During 100 loops, the algorithm randomly selects random variables within the given parameter values by shuffling-reshuffling bringing them into the grid-by-grid model for simulation, and calculates the deviation and RMSE between the simulated yield and the target yield (tutor). Oscillate in waves, with the maximum deviation of 2929.9 kg/ha and the minimum deviation of −4202.7 kg/ha; the maximum RMSE appears in the 54th loop, with an RMSE of 4411.94 kg/ha; and the minimum RMSE appears in the 95th loop, with an RMSE of 53.75 kg/ha. It can be seen that, within 100 loops, the SCE-UA algorithm can find the optimal parameter optimisation result automatically. Finally, the set of parameter variables with the smallest deviation and the smallest RMSE was defined as the optimal crop parameters for the grid point.

While we've managed to reduce the time required for optimizing a single parameter to 26.8 s through iterative experimentation, conducting simulations on a large regional scale still demands considerable time. This prolonged duration severely restricts the decision-support capabilities of the grid-by-grid model, particularly in addressing time-sensitive applications such as yield forecasting. Consequently, we suggest two potential initiatives to augment the efficiency of model simulations further. The feasibility of these measures will be examined in our forthcoming research endeavours. The first strategy we devised involves constructing a nationwide database of winter wheat parameters utilizing our grid model. Leveraging a supervised optimization algorithm, highly accurate crop parameters can be derived provided there is sufficient prior knowledge, such as statistical yield or productivity data from specific years. These parameters are stored within a spatial database. When conducting time-sensitive simulations, this spatial database enables direct access to crop parameters, eliminating the need for repetitive parameter optimization steps and thereby indirectly enhancing simulation efficiency. As for the second approach, we propose enhancing and refining the SCE-UA algorithm. This could entail introducing new variations and crossover operations, as well as rectifying any existing algorithmic shortcomings to bolster efficiency and performance. Additionally, local optimization of the algorithm for each solution via local search or alternative optimization strategies during each iteration could expedite algorithm convergence. Moreover, parallelizing computational tasks within the algorithm using multi-core or distributed computing resources stands to accelerate computation speed. After further improving the execution speed of the model, our grid-by-grid model will be able to be applied to time-critical agricultural monitoring services, such as winter wheat growth monitoring and early warning of drought. In addition, we are curious about the performance of our grid point model on small-scale aboveground biomass estimation, and whether it can be coupled with UAV imagery and ground spectral observation information to further improve the accuracy of our model .

4.2. Comparison with relevant spatialized models

In this study, we conducted a comprehensive literature review on wheat yield estimation employing spatialized crop modelling approaches in recent years, comparing our findings with existing studies (Table 5). The predominant research theme across these articles involves the utilization of multi-year continuous field observation data to establish rate-setting crop model parameters. Subsequently, spatial simulations are conducted using these rate-determined parameters. However, this conventional approach is both resource-intensive and spatially limiting, restricting the model parameters to the immediate study area.

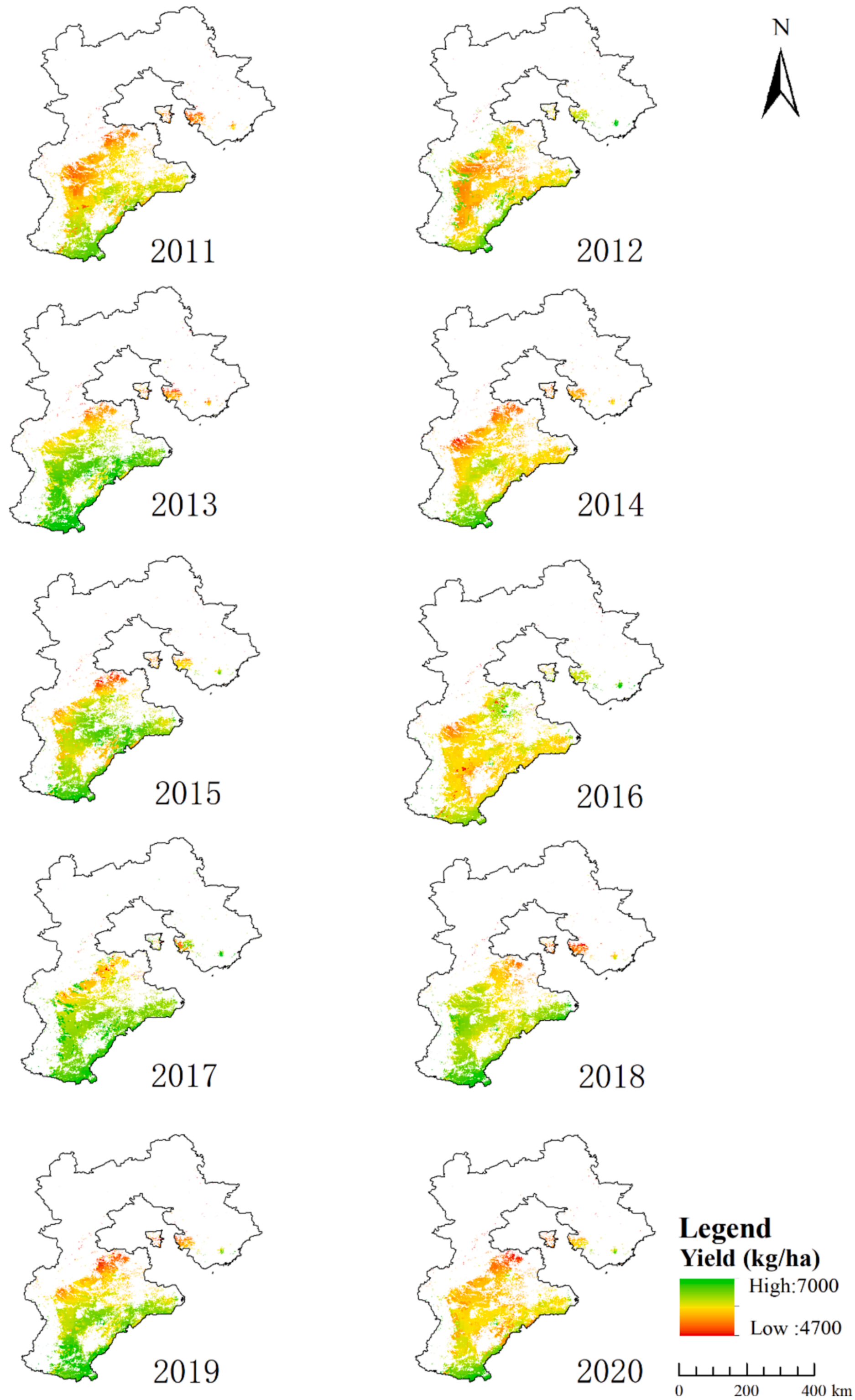

Fig. 10. Distribution of winter wheat grain production from 2011 to 2020.

In contrast, our proposed method employs a machine learning approach for the parameter rate-setting of spatial crop models, enabling grid-by-grid parameterization based on average yield information from historical records. This innovative methodology not only circumvents the need for laborious field experiments but also facilitates spatial adaptability. Our research encompasses the entire Hebei Province over the period 2010–2020, providing a broader scope and a more extended timeframe, thus enhancing its significance in guiding macroscopic decision-making processes.

Regarding input data for spatial crop models, several studies tend to oversimplify parameters such as soil, fertilizer, and irrigation information due to constraints inherent in the employed models, such as SAFY and AquaCrop. Our spatial model, however, meticulously collects and spatializes detailed information on soil, fertilizer, and irrigation, reconstructing the crop's growth and management processes in real-world scenarios and mitigating spatial model uncertainty.

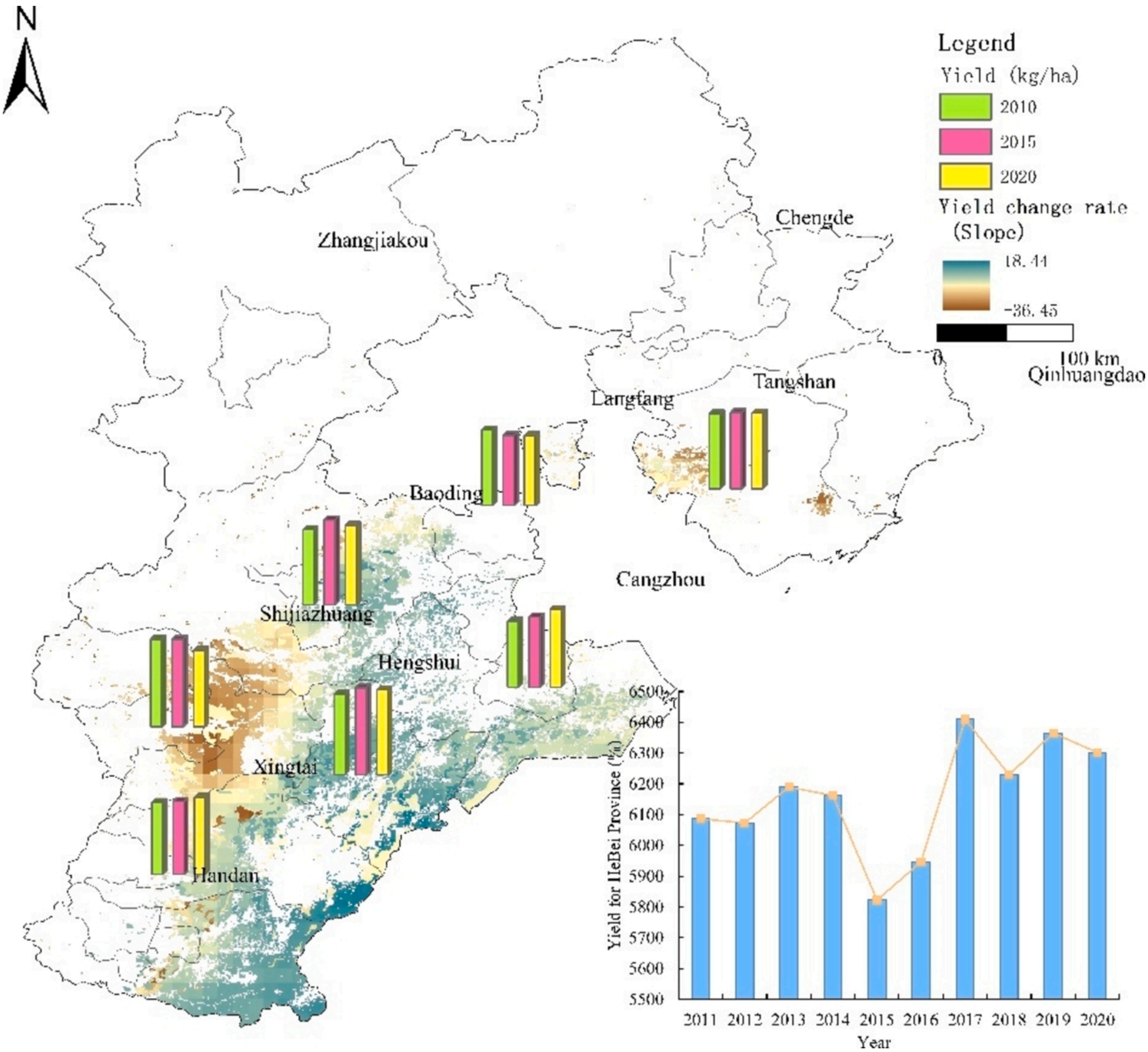

Fig. 11. Changes in the spatial and temporal distribution of winter wheat production in Hebei Province (2011–2020).

Table 3
The detailed settings of the SCE-UA algorithm.

Setting	Value
Number of parameters	9
Number of complexes	2(dim)
Maximum number of evolution loops before convergence	100
The percentage change allowed in loops before convergence	10^{-5}
Convergence criterion	10^{-4}

Table 4
Different Maximum number of evolution loops for model efficiency.

Maximum number of evolution loops	20	50	100	200	500
Bais (kg/ha)	−2869.3	704.8	**830.8**	830.8	652.6
RMSE (kg/ha)	195.8	187.0	**53**	53	47
Time cost (s)	5.6	12.4	**26.8**	54.3	186.2

In the realm of spatialization methods, while many studies rely on meteorological station data, our approach utilizes remote sensing inversion products and spatialized raster data products. Each raster corresponds to a set of refined input parameters, surpassing the granularity of input data employed in previous studies.

Concerning remote sensing assimilation methods, we opted for the 4Dvar assimilation algorithm, aligning with mainstream approaches. The simulation accuracy achieved with AM-4Dvar (RMSE: 148.48 kg/ha) surpassed that of AM-Enkf (RMSE: 256.04 kg/ha). Importantly, our regional model also adopted the 4Dvar (SCE-UA) assimilation algorithm, yielding a comparable RMSE of 148.48 kg/ha, showcasing a commendable level of accuracy consistent with existing methodologies. When comparing our AM-4Dvar model to studies utilizing the 4Dvar constructor, our findings consistently exhibit superior performance in terms of both R^2 and RMSE metrics. The differentiating factor between our approach and others lies primarily in the additional step of parameter pre-optimization using the SCE-UA algorithm before 4Dvar assimilation. This highlights the profound impact of integrating SCE-UA within a yield prediction framework, notably enhancing the accuracy of

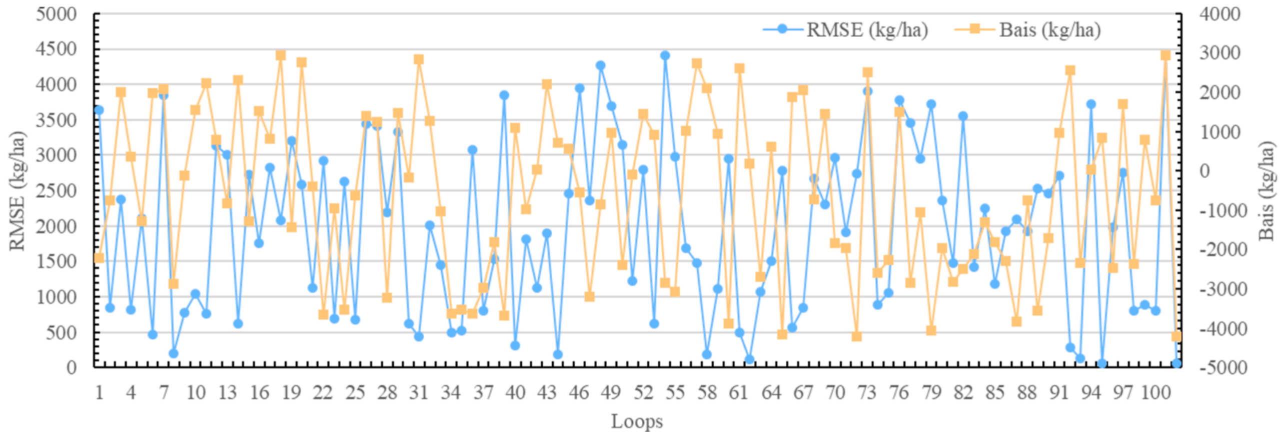

Fig. 12. SCE-UA algorithm execution efficiency.

winter wheat yield estimation. Our research stands out for its innovation in integrating diverse modern agricultural techniques to estimate winter wheat yields. We developed a spatialized crop model on a grid-based framework, augmented by a comprehensive crop parameter database. Utilizing the SCE-UA algorithm and remote sensing assimilation method, we markedly enhanced the accuracy of our grid-by-grid model.

4.3. Uncertainty analysis

4.3.1. Uncertainty is in the crop model itself

The PCSE model still has more prerequisite assumptions and empirical parameters in the simulation of crop growth processes (De Wit et al., 2019). A particular limitation pertains to the treatment of fertilizer application within the PCSE model. In its current implementation, fertilizer application is incorporated into the model through an external module utilizing empirical parameters. The PCSE lacks a comprehensive simulation of soil and crop nutrient transport conditions, particularly about crucial elements such as nitrogen, phosphorus, and potassium. A conspicuous uncertainty arises from the omission of explicit consideration for the influence of nutrient elements on crop growth within the current framework. The absence of a nuanced simulation of nutrient transport conditions undermines the model's capacity to faithfully capture the intricate interplay between soil nutrient dynamics and crop development.

4.3.2. Uncertainty in remote sensing raster products

The precision of remote sensing raster products is often compromised by inherent uncertainties, encompassing challenges such as low inversion accuracy, sporadic instances of abnormal and missing values, and the ubiquitous issue of mixed pixels (Tran et al., 2023). These complexities in data quality and fidelity significantly impact the reliability of information derived from remote sensing datasets. The classification accuracy of crop classification data not only affects the size and range of winter wheat planting location but also influences the input information such as soil texture and management strategy of each raster, which in turn affects the results of winter wheat yield simulation of each raster. One of particular concerns is the classification accuracy of crop data, a pivotal determinant that extends beyond the mere spatial demarcation of winter wheat plantations. The fidelity of crop classification data plays a crucial role in shaping not only the spatial distribution and extent of winter wheat planting locations but also exerts a cascading influence on essential input parameters, including soil texture and management strategies associated with each raster. The interplay between classification accuracy and input information intricately shapes the outcomes of winter wheat yield simulations. Addressing these uncertainties becomes paramount in ensuring the robustness and accuracy of remote sensing applications in the spatialized crop model.

4.3.3. Uncertainty arising from regionalisation of model input parameters

In the implementation of our grid-by-grid spatialized crop model, we endeavoured to spatialize the crop model comprehensively by integrating diverse data sources, including remote sensing and statistical data. Despite these efforts, the challenge persists due to the limited number of input parameters that can be effectively rasterized, particularly when compared to the substantial number of parameters inherent in crop models. The sheer volume of input parameters required remains a fundamental challenge, prompting reliance on default values provided by models or reference values from existing studies (O'Hagan. 2012).

Furthermore, the spatialization process introduces a layer of uncertainty to the input data. Notably, the intricacies of certain parameters, such as fertilizer application and irrigation quantities, pose challenges for accurate inversion through remote sensing. Consequently, the construction of raster data relies on county-scale values from statistical yearbooks, combined with limited information on field management practices gleaned from the study area. This amalgamation, though informed, inevitably simplifies and distorts the representation of actual field management situations.

For instance, the intricacies of fertilizer application and irrigation quantities, difficult to ascertain through remote sensing alone, necessitate reliance on county-scale values from statistical yearbooks. This data is then complemented by insights into field management habits from the study area, albeit in a simplified and distorted representation compared to the actual field management scenarios.

In navigating these challenges, it becomes imperative to acknowledge and quantify the uncertainties introduced during the spatialization of input parameters. Addressing these intricacies will be pivotal for refining the accuracy and reliability of the grid-by-grid regional crop model, facilitating more nuanced and informed insights into agricultural dynamics on a larger spatial scale.

4.3.4. Uncertainty in model parameter optimisation algorithms

The utilization of machine learning methods for the automatic optimization of crop model input parameters presents a substantial leap in terms of time and resource efficiency, vastly outperforming traditional parameter adjustment approaches (Gao et al., 2020). This advancement not only expedites the optimization process but also significantly enhances the accuracy and efficiency of parameter calibration. Despite these advantages, an inherent challenge arises in the form of parameter sets yielding identical simulation results, introducing a dilemma known as the "different parameters, same result" problem.

In grappling with this predicament, the pertinent question arises: when confronted with multiple parameter sets producing identical simulation outcomes, how does one discern the optimal set of parameters for finalizing input conditions? This conundrum necessitates thoughtful consideration, particularly given the expansive nature of the

Table 5
Comparison with relevant research.

ID	CGM	Input				Spatialized						Accuracy		Reference
		Crop	Soil	Fertilizer	Irrigation	Resolution	Approach	Cost function	Algorithm	Variables	RS data	R^2	RMSE(kg/ha)	
GM	WOFOST	Wheat	✓	✓	✓	0.1°	—	—	—	—	—	0.53	824.82	—
OM	WOFOST	Wheat	✓	✓	✓	0.1°	VA	RMSE	SCE-UA	—	—	0.84	477.96	—
AM-Enkf	WOFOST	Wheat	✓	✓	✓	0.1°	VA + FA	—	Enkf	LAI	MODIS	0.87	256.04	—
AM-4Dvar	WOFOST	Wheat	✓	✓	✓	0.1°	VA + VA	4Dvar	SCE-UA	LAI	MODIS	0.94	148.48	—
1	WOFOST	Wheat	✓	✓		500 *m*	VA	4Dvar	SCE-UA	LAI	MODIS	0.53	619.73	Zhuo et al., 2022
2	WOFOST	Wheat	✓			1 km	VA	4Dvar	SCE-UA	LAI	MODIS, Landsat TM	0.48	151.92	Huang et al., 2015
3	SAFY	Wheat				30 *m*	VA	RMSE	SCE-UA	LAI	MODIS, Landsat-8	0.76	176	Dong et al, 2016
4	DSSAT	Wheat	✓	✓		300 *m*	VA	WSSD	MLS	LAI	ENVISAT ASAR, MERIS	—	420	Dente et al, 2008
5	AquaCrop	Wheat				30 *m*	VA	RE	PSO	CC	HJ-1A/B, RADARSAT-2	0.42	810	Jin et al, 2017
6	WOFOST	Wheat	✓	✓	✓	1 km	FA	—	Enkf	LAI	MODIS, Landsat TM	0.37	472	Huang et al., 2016
7	WOFOST	Wheat	✓			0.5°	FA	—	Enkf	LAI	HJ-1A/B	0.42	737	Chen et al., 2018
8	DSSAT	Wheat				30 *m*	FA	—	Enkf	LAI, VTCI	Landsat TM, ETM + and OLI	0.76	549	Xie et al., 2017

* VA, FA, WSSD, MLS, RE, PSO, CC and VTCI represent the variational approaches, filtering approaches, the weighted sum of squared differences, maximum likelihood solution, relative error, particle swarm optimisation, canopy cover, vegetation temperature condition index.

grid-by-grid regional crop model deployed in this study. The study area encompasses a multitude of grids, with a spatial database containing extensive data. Within such a comprehensive dataset, the existence of multiple parameter sets generating indistinguishable simulation results is a plausible scenario. Consequently, the selection of the most appropriate set of parameters assumes paramount significance, considering the nuanced spatial intricacies inherent in the grid-by-grid regional crop model.

Addressing this uncertainty entails a meticulous examination of parameter optimization algorithms, emphasizing the need for robust methodologies capable of discerning optimal parameter sets amidst the complexity of large-scale spatial datasets (Wallach et al., 2021). By deliberating on and refining approaches to this challenge, the study endeavours to contribute to the ongoing pursuit of precision and reliability in the automatic optimization of crop model input parameters, particularly in the context of grid-based regional simulations.

4.3.5. Uncertainty introduced by remote sensing assimilation

A pivotal consideration in integrating remote sensing assimilation methods revolves around determining which source to prioritize: the outcomes derived from remote sensing inversions or the simulations generated by crop models. In the realm of remote sensing assimilation within crop models, there is a prevailing inclination to accord greater trust to the results obtained through remote sensing inversions. This preference stems from the belief that remote sensing data inherently possess authenticity and reliability, while the crop model simulation process is deemed fraught with uncertainty, leading to potential inaccuracies.

In reality, the veracity of remotely sensed data is not absolute. Despite notable enhancements brought about by ensemble Kalman filtering and four-dimensional variational assimilation algorithms in refining the precision of raster-by-raster regional crop model simulations, the extent of uncertainty introduced by remote sensing data assimilation remains a subject of ongoing debate.

The overarching question persists: to what degree does remote sensing assimilation contribute to uncertainties in regional crop model simulations? Despite notable strides in assimilation methodologies, this question remains a focal point for continued discourse and exploration. The need for a comprehensive understanding of the uncertainties introduced by remote sensing assimilation is paramount, emphasizing the importance of ongoing debate and research in this domain. The pursuit of clarity in this matter is essential for refining the accuracy and reliability of integrated remote sensing and crop model simulations in agricultural assessments.

5. Conclusions

This study introduces a novel approach by integrating remote sensing assimilation techniques and machine learning methods into a spatialized crop model. A grid-by-grid regional crop model, constructed from raster data utilizing remote sensing inputs, was implemented for winter wheat yield estimation spanning the years 2011 to 2020 in Hebei Province, China. The primary focus was to assess the impacts of two distinct remote sensing assimilation methods, namely Ensemble Kalman Filter (Enkf) and Four-Dimensional Variational Assimilation (4Dvar), coupled with a machine learning technique, Shuffled Complex Evolution Algorithm (SCE-UA), on winter wheat yield estimation.

The experimentation reveals that the most favourable outcome in winter wheat yield estimation is attained through the optimized assimilation strategy of AM-4Dvar achieving an R^2 of 0.94 and reducing the Root Mean Square Error (RMSE) to 148.48 kg/ha. Subsequently, the study delves into the efficacy of the machine learning approach and addresses the inherent uncertainties associated with the regional crop model.

This research contributes significantly to the advancement of spatial crop models, offering innovative perspectives on the spatialization of crop models. Moreover, it enhances the precision of spatial crop model simulations, a critical factor in comprehending crop productivity on a large regional scale. The implications extend to macro-planning for agricultural plantation management, fostering improvements in crop yields, and ultimately contributing to the overarching goal of ensuring food security. The insights gleaned from this study open avenues for further exploration in the development of spatial crop models, reinforcing their role in informed decision-making for sustainable agricultural practices.

CRediT authorship contribution statement

Qiang Li: Writing – original draft, Methodology. **Maofang Gao:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Sibo Duan:** Investigation. **Guijun Yang:** Resources. **Zhao-Liang Li:** Writing – review & editing, Supervision.

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.

Acknowledgement

This research was supported by the National Natural Science Foundation of China (41921001)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2024.109594.

Data availability

Data will be made available on request.

References

Akkem, Y., Biswas, S.K., Varanasi, A., 2023. "Smart farming using artificial intelligence: a review. Eng. Appl. Artif. Intel. 120, 105899. https://doi.org/10.1016/j.engappai.2023.105899.

Basso, B., Liu, L., 2018. "Seasonal crop yield forecast: methods, applications, and accuracies. Adv. Agron. 154, 201–255. https://doi.org/10.1016/bs.agron.2018.11.002.

Cao, H., Gu, Y., Fang, J., Hu, Y., Ding, W., He, H., Chen, G., 2022. "Application of stacking ensemble learning model in quantitative analysis of biomaterial activity. Microchem. J. 183, 108075. https://doi.org/10.1016/j.microc.2022.108075.

Ceglar, A., Toreti, A., Prodhomme, C., Zampieri, M., Turco, M., Doblas-Reyes, F.J., 2018. Land-surface initialisation improves seasonal climate prediction skill for maize yield forecast. Scientific Reports 8 (1), 1322. https://doi.org/10.1038/s41598-018-19586-6.

Chen, Y., Zhang, Z., Tao, F., 2018. Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data. Eur. J. Agron. 101, 163–173. https://doi.org/10.1016/j.eja.2018.09.006.

De Wit, A., Boogaard, H., Fumagalli, D., Janssen, S., Knapen, R., van Kraalingen, D., Supit, I., van der Wijngaart, R., van Diepen, K., 2019. 25 years of the WOFOST cropping systems model. Agr. Syst. 168, 154–167. https://doi.org/10.1016/j.agsy.2018.06.018.

Dente, L., Satalino, G., Mattia, F., Rinaldi, M., 2008. Assimilation of leaf area index derived from ASAR and MERIS data into CERES-Wheat model to map wheat yield. Remote Sens. Environ. 112 (4), 1395–1407. https://doi.org/10.1016/j.rse.2007.05.023.

Dhakar, R., Sehgal, V.K., Chakraborty, D., Sahoo, R.N., Mukherjee, J., Ines, A., Kumar, S. N., Shirsath, P.B., Roy, S.B., 2022. Field-scale spatial wheat yield forecasting system under limited field data availability by integrating crop simulation model with weather forecast and satellite remote sensing. Agr. Syst. 195, 103299. https://doi.org/10.1016/j.agsy.2021.103299.

Dong, T., Liu, J., Qian, B., Zhao, T., Jing, Q., Geng, X., Wang, J., Huffman, T., Shang, J., 2016. Estimating winter wheat biomass by assimilating leaf area index derived from fusion of Landsat-8 and MODIS data. Int. J. Appl. Earth Obs. Geoinf. 49, 63–74. https://doi.org/10.1016/j.jag.2016.02.001.

Doraiswamy, P.C., Hatfield, J.L., Jackson, T.J., Akhmedov, B., Prueger, J., Stern, A., 2004. Crop condition and yield simulations using landsat and modis. Remote Sens. Environ. 92 (4), 548–559. https://doi.org/10.1016/j.rse.2004.05.017.

Doraiswamy, P.C., Sinclair, T.R., Hollinger, S., Akhmedov, B., Stern, A., Prueger, J., 2005. Application of modis derived parameters for regional crop yield assessment. Remote Sens. Environ. 97 (2), 192–202. https://doi.org/10.1117/12.454181.

Duan, Q., Sorooshian, S., Gupta, V.K., 1994. Optimal use of the sce-ua global optimization method for calibrating watershed models. J. Hydrol. 158 (3–4), 265–284. https://doi.org/10.1016/0022-1694(94)90057-4.

Dumont, B., Basso, B., Leemans, V., Bodson, B., Destain, J.P., Destain, M.F., 2015. A comparison of within-season yield prediction algorithms based on crop model behaviour analysis. Agric. for. Meteorol. 204, 10–21. https://doi.org/10.1016/j.agrformet.2015.01.014.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. J. Geophys. Res. Oceans 99 (C5). https://doi.org/10.1029/94JC00572.

Feng, P., Wang, B., Liu, D., Waters, C.M., Yu, Q., 2020. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agric. for. Meteorol. 285–286. https://doi.org/10.1016/j.agrformet.2020.107922.

Gao, Y., Wallach, D., Liu, B., Dingkuhn, M., Boote, K.J., Singh, U., Asseng, S., Kahvecih, T., He, J., Zhang, R., Hoogenboom, G., 2020. Comparison of three calibration methods for modeling rice phenology. Agric. for. Meteorol. 280, 107785. https://doi.org/10.1016/j.agrformet.2019.107785.

Huang, J.X., Gomez-Dans, J., Huang, H., Ma, H., Xie, X., 2019a. Assimilation of remote sensing into crop growth models: current status and perspectives. Agricultural and Forest Meteorology 276 (277), 107609. https://doi.org/10.1016/j.agrformet.2019.06.008.

Huang, J., Tian, L., Liang, S., Ma, H., Becker-Reshef, I., Huang, Y., Su, W., Zhang, X., Zhu, D., Wu, W., 2015. Improving winter wheat yield estimation by assimilation of the leaf area index from landsat tm and modis data into the wofost model. Agric. for. Meteorol. 204, 106–121. https://doi.org/10.1016/j.agrformet.2015.02.001.

Huang, J., Sedano, F., Huang, Y., Ma, H., Li, X., Liang, S., Tian, L., Zhang, X., Fang, J., Wu, W., 2016. Assimilating a synthetic kalman filter leaf area index series into the wofost model to improve regional winter wheat yield estimation. Agric. for. Meteorol. 216, 188–202. https://doi.org/10.1016/j.agrformet.2015.10.013.

Huang, J., Ma, H., Sedano, F., Lewis, P., Liang, S., Wu, Q., Su, W., Zhang, X., Zhu, D., 2019b. Evaluation of regional estimates of winter wheat yield by assimilating three remotely sensed reflectance datasets into the coupled wofost–prosail model. Eur. J. Agron. 102, 1–13. https://doi.org/10.1016/j.eja.2018.10.008.

Hwang, C., Correll, M.J., Gezan, S.A., Zhang, L., Bhakta, M.S., Vallejos, C.E., Boote, K.J., Clavijo-Michelangeli, J.A., Jones, J.W., 2017. Next generation crop models: a modular approach to model early vegetative and reproductive development of the common bean (phaseolus vulgaris l). Agr. Syst. 155, 225–239. https://doi.org/10.1016/j.agsy.2016.10.010.

Ines, A., Hansen, J.W., Robertson, A.W., 2011. Enhancing the utility of daily gcm rainfall for crop yield prediction. Int. J. Climatol. 31 (14), 2168–2182. https://doi.org/10.1002/joc.2223.

Ines, A., Das, N.N., Hansen, J.W., Njoku, E.G., 2013. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. Remote Sens. Environ. 138 (6), 149–164. https://doi.org/10.1016/j.rse.2013.07.018.

Iniyan, S., Jebakumar, R., 2022. Mutual information feature selection (mifs) based crop yield prediction on corn and soybean crops using multilayer stacked ensemble regression (mser). Wireless Personal Communications 126 (3), 1935–1964. https://doi.org/10.1007/s11277-021-08712-9.

Jin, X., Li, Z., Yang, G., Yang, H., Feng, H., Xu, X., Wang, J., Li, X., Luo, J., 2017. Winter wheat yield estimation based on multi-source medium resolution optical and radar imaging data and the aquacrop model using the particle swarm optimization algorithm. ISPRS J. Photogramm. Remote Sens. 126 (APR.), 24–37. https://doi.org/10.1016/j.isprsjprs.2017.02.001.

Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., Wang, J., 2018. A review of data assimilation of remote sensing and crop models. Eur. J. Agron. 92, 141–152. https://doi.org/10.1016/j.eja.2017.11.002.

Laar, H. H. V., J. Goudriaan, and H. van Keulen. 1992. Simulation of crop growth for potential and water-limited production situations (as applied to spring wheat). *Simulation Reports Cabo-Tt 27. Cabo-Dlo, Wau-Tpe, Wageningen.*

Lecerf, R., Ceglar, A., López-Lozano, R., Marijn, V., Baruth, B., 2018. Assessing the information in crop model and meteorological indicators to forecast crop yield over europe. Agr. Syst. https://doi.org/10.1016/j.agsy.2018.03.002. S0308521X17310223.

Liao, Q., Leng, P., Li, Z., Ren, C., Sun, Y., Gao, M., Duan, S., Shang, G., 2020. A method for deriving relative humidity from modis data under all-sky conditions. IEEE Trans. Geosci. Remote Sens. 1–15. https://doi.org/10.1109/TGRS.2020.3036248.

Ma, G., Huang, J., Wu, W., Fan, J., Zou, J., Wu, S., 2013. Assimilation of modis-lai into the wofost model for forecasting regional winter wheat yield. Math. Comput. Model. 58 (3), 634–643. https://doi.org/10.1016/j.mcm.2011.10.038.

Machwitz, M., Giustarini, L., Bossung, C., Frantz, D., Schlerf, M., Lilienthal, H., Wandera, L., Matgen, P., Hoffmann, L., Udelhoven, T., 2014. Enhanced biomass prediction by assimilating satellite data into a crop growth model. Environ. Model. Softw. 62, 437–453. https://doi.org/10.1016/j.envsoft.2014.08.010.

O'Hagan, A., 2012. Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux. Environ. Model. Softw. 36, 35–48. https://doi.org/10.1016/j.envsoft.2011.03.003.

Olauson, J., 2018. Era5: the new champion of wind power modelling? Renew. Energy 126, 322–331. https://doi.org/10.1016/j.renene.2018.03.056.

Paola, A.D., Valentini, R., Santini, M., 2016. An overview of available crop growth and yield models for studies and assessments in agriculture. J. Sci. Food Agric. 96 (3), 709–714. https://doi.org/10.1002/jsfa.7359.

Paudel, D., Boogaard, H., Wit, A.D., Janssen, S., Athanasiadis, I.N., 2020. Machine learning for large-scale crop yield forecasting. Agr. Syst. 187, 103016. https://doi.org/10.1016/j.agsy.2020.103016.

Philippe, Courtier, Olivier, and Talagrand. 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation. Ii: numerical results. *Quarterly Journal of the Royal Meteorological Society*. doi:10.1002/qj.49711347813.

Praveen, B., Sharma, P., 2020. Climate variability and its impacts on agriculture production and future prediction using autoregressive integrated moving average method (arima). Journal of Public Affairs 20 (2). https://doi.org/10.1002/pa.2016.

Qian, B., Jong, R.D., Warren, R., Chipanshi, A., Hill, H., 2009. Statistical spring wheat yield forecasting for the canadian prairie provinces. Agric. for. Meteorol. 149 (6–7), 1022–1031. https://doi.org/10.1016/j.agrformet.2008.12.006.

Roudier, P., Alhassane, A., Baron, C., Louvet, S., Sultan, B., 2016. Assessing the benefits of weather and seasonal forecasts to millet growers in niger. Agric. for. Meteorol. 223, 168–180. https://doi.org/10.1016/j.agrformet.2016.04.010.

Shi, X.Z., Yu, D.S., Warner, E.D., Pan, X.Z., Petersen, G.W., Gong, Z.G., Weindorf, D.C., 2004. Soil database of 1:1,000,000 digital soil survey and reference system of the chinese genetic soil classification system. Soil Survey Horizons 45, 129–136. https://doi.org/10.2136/sh2004.4.0129.

Sobol, I.M., Kucherenko, S.S., 2005. Global sensitivity indices for nonlinear mathematical models. Review. Wilmott 2005 (1), 56–61. https://doi.org/10.1002/wilm.42820050114.

Tran, B.N., Van Der Kwast, J., Seyoum, S., Uijlenhoet, R., Jewitt, G., Mul, M., 2023. Uncertainty assessment of satellite remote-sensing-based evapotranspiration estimates: a systematic review of methods and gaps. Hydrol. Earth Syst. Sci. 27 (24), 4505–4528. https://doi.org/10.5194/hess-27-4505-2023.

Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Gourdain, E., Andrianasolo, F., Seidel, S.J., 2021. 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.

Wit, A., Diepen, C., 2007. Crop model data assimilation with the ensemble kalman filter for improving regional crop yield forecasts. Agric. for. Meteorol. 146 (1–2), 38–56. https://doi.org/10.1016/j.agrformet.2007.05.004.

Wit, A., Duveiller, G., Defourny, P. Estimating regional winter wheat yield with WOFOST through the assimilation of green area index retrieved from MODIS observations. Agricultural and Forest Meteorology, 2012, 320 164:39-52.321. doi:10.1016/j.agrformet.2012.04.011.

Wu, S., Yang, P., Chen, Z., Ren, J., Sun, L., 2021. Estimating winter wheat yield by assimilation of remote sensing data with a four-dimensional variation algorithm considering anisotropic background error and time window. Agric. for. Meteorol. 301–302 (1), 108345. https://doi.org/10.1016/j.agrformet.2021.108345.

Xie, Y., Wang, P., Bai, X., Khan, J., Zhang, S., Li, L., Wang, L., 2017. Assimilation of the leaf area index and vegetation temperature condition index for winter wheat yield estimation using Landsat imagery and the CERES-Wheat model. Agricultural and Forest Meteorology 246, 194–206. https://doi.org/10.1016/j.agrformet.2017.06.015.

Yang, J., 2011. Convergence and uncertainty analyses in monte-carlo based sensitivity analysis. Environ. Model. Softw. 26 (4), 444–457. https://doi.org/10.1016/j.envsoft.2010.10.007.

Yanyan, D., Qing, H., Jianqiang, R., 2020. Aerosol variation and its effect on winter wheat productivity during the covid-19. Chin. J. Eco-Agric. https://doi.org/10.13930/j.cnki.cjea.200717.

Zambrano, F., Vrieling, A., Nelson, A., Meroni, M., Tadesse, T., 2018. Prediction of drought-induced reduction of agricultural productivity in chile from modis, rainfall estimates, and climate oscillation indices. Remote Sens. Environ. 219, 15–30. https://doi.org/10.1016/j.rse.2018.10.006.

Zhuo, W., Huang, J., Li, L., Zhang, X., Xiao, X., 2019. Remote sensing assimilating soil moisture retrieved from sentinel-1 and sentinel-2 data into wofost model to improve winter wheat yield estimation. Remote Sens. (basel) 11 (11), 1618. https://doi.org/10.3390/rs11131618.

Zhuo, W., Fang, S., Gao, X., Wang, L., Wu, D., Fu, S., Wu, Q., Huang, J., 2022. Crop yield prediction using modis lai, tigge weather forecasts and wofost model: a case study for winter wheat in hebei, china during 2009–2013. Itc Journal 106. https://doi.org/10.1016/j.jag.2021.102668.