scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Mapping 1-km soybean yield across China from 2001 to 2020 based on ensemble learning

Min Zhang[1], Xinlei Xu[1], Junji Ou[1], Zengguang Zhang[1], Fangzheng Chen[1], Lijie Shi[2], Bin Wang[3], Meiqin Zhang[1], Liang He[4], Xueliang Zhang[1], Yong Chen[1], Kelin Hu[1] & Puyu Feng[1]✉

Soybean is a critical agricultural product in China, with domestic production unable to satisfy the substantial demand, leading to a huge reliance on imports. To support the scientific formulation of agricultural policies and the optimization of domestic planting structures, we developed a high-resolution annual soybean yield dataset for China (2001–2020), ChinaSoyYield1km. This dataset was generated by applying ensemble learning algorithms and spatial decomposition to a comprehensive set of multi-source data, including climate variables, remote sensing imagery, soil properties, agricultural management practices, and official yield records. The integration of these diverse datasets allows for a nuanced understanding of the factors influencing soybean yield at a 1-km resolution. The resulting dataset captures over 50% of the yield variability at the county scale, demonstrating superior accuracy compared to publicly available datasets with reductions in Root Mean Square Error (RMSE) ranging from 0.18 to 0.60 t/ha. It is anticipated that our dataset will enhance agricultural studies, planning, and policy-making related to soybean cultivation, providing a valuable resource for both the scientific community and government.

Background & Summary

Global food security is facing challenges, particularly against the backdrop of complex political and climatic conditions[1]. Among major agricultural products, soybean is not only a crucial source of protein and oil but also plays a key role in agriculture, industry, and the sustainable development of economies[2]. High-spatial-resolution and high-accuracy soybean yield dataset is a powerful tool to provide more scientific and macroscopic perspective to investigate soybean production, thereby enhancing cultivation and management techniques effectively to ensure a stable supply of soybean. However, such kind of dataset is still limited.

Currently, the primary sources of soybean yield spatialization information include yield data recorded by agricultural meteorological stations, statistical data at administrative unit scales, and rasterized yield datasets. Despite their advantages, these datasets do have their limitations either in spatial resolution or time resolution, thus limiting their use over large areas or long time periods. For instance, the yield data recorded by agricultural meteorological stations can provide real-time, location-specific records of soybean yields, which typically possess a higher degree of accuracy[3]. However, such data are limited in spatial coverage making it challenging to represent the spatial variability of soybean yield over large areas. Similarly, soybean yield data at the administrative unit scales are based on statistics and records at administrative units (such as provinces, cities, and counties). These data facilitate the analysis of differences in soybean yield levels across different regions but cannot reflect spatial differences within regions[4,5]. Compared with recorded data, rasterized soybean yield datasets should be able to provide high-resolution spatial data and enable a more detailed analysis of the distribution of soybean yield across different geographical locations. However, the existing commonly used rasterized soybean yield datasets such as Harvested Area and Yield for 4 Crops (EarthStat)[6], Spatial Production Allocation Model (MapSPAM)[7], and Global Dataset of Historical Yields (GDHY)[8] often have low temporal or spatial resolution. The highest spatial resolution of these dataset is only 10 kilometers. Moreover, some of these datasets only

[1]College of Land Science and Technology, State Key Laboratory of Efficient Utilization of Agricultural Water Resources, China Agricultural University, Beijing, 100193, China. [2]College of Hydraulic Science and Engineering, Yangzhou University, Yangzhou, Jiangsu, 225009, China. [3]New South Wales Department of Primary Industries, Wagga Wagga Agriculture Institute, Wagga Wagga, New South Wales, 2650, Australia. [4]National Meteorological Center, Beijing, 100081, China. ✉e-mail: fengpuyu@cau.edu.cn

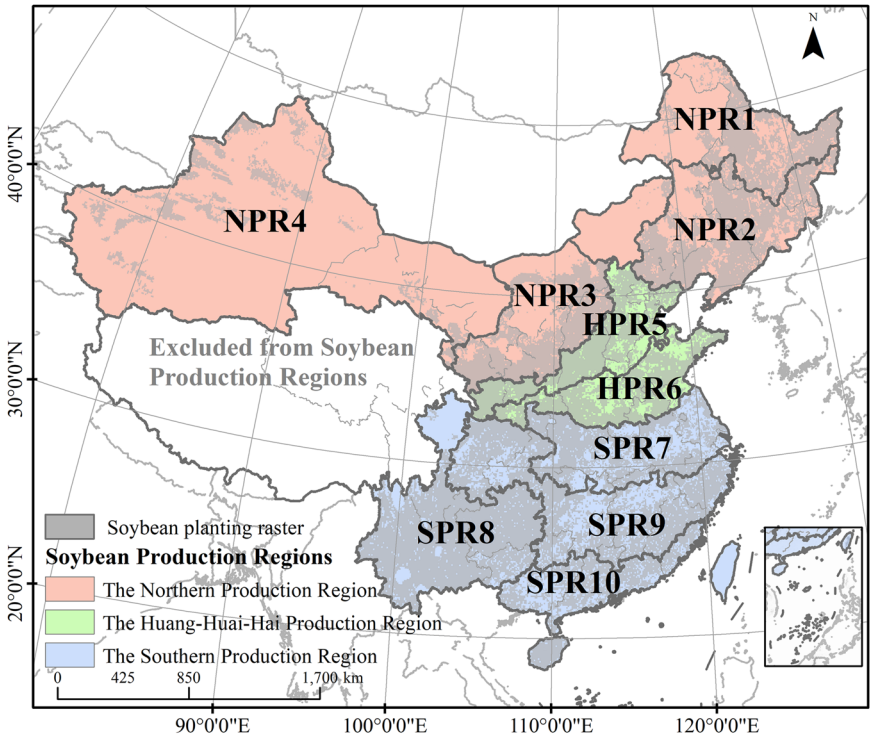

Fig. 1 Spatial distribution of 3 major soybean production regions and 10 sub-regions in China. NPR is the Northern Production Region; HPR is the Huang-Huai-Hai Production Region; SPR is the Southern Production Region.

contain data series for 2 or 3 years, which is not sufficient for establishing robust statistical analysis. In summary, the usability of the spatial distribution data for soybean yield is still inadequate, limiting their utility in precise agricultural planning and management. Thus, there is an urgent need to develop a dataset that features high temporal and spatial resolution, along with high precision, specifically a multi-year soybean yield spatialization dataset.

The primary approaches used to monitor crop yield are process-based crop models and statistic models[9,10]. However, these two types of models are not particularly suitable for generating spatially gridded yield datasets. Process-based crop models are mathematical models based on principles of crop physiology and environmental science, but they typically require high-quality ground-based observations and extensive data inputs (such as daily-scale temperature, precipitation, and solar radiation)[11], making them difficult to use in data-scarce regions[12–14]. In contrast to process-based models, statistic models link crop yield to predictor variables and calibrate empirical relationships through measurement results[15,16]. Due to their simplicity of operation, statistic models are widely used for estimating crop yield[17]. However, these models are not without issues in the study of crop yield spatialization. Statistic models are often localized, and the empirical relationships between crop yield and predictor variables cannot be easily generalized to other regions. Moreover, traditional linear statistic models are particularly limited in capturing non-linear relationships between variables[18,19].

Machine learning models is rapidly evolving and has a wide range of applications in agricultural research[20]. Compared to traditional process-based crop models and statistic models, machine learning models, e.g., random forest and support vector machine, can handle larger and more complex datasets, uncovering non-linear and intricate relationships within the data, thus improving the accuracy and reliability of model estimations[21]. In fact, more and more studies use machine learning models to predict crop yields[22–24]. Additionally, machine learning models can automatically identify and leverage key features within the data and continuously improve model performance through ongoing learning and adjustments[25]. Given its advantages in processing large datasets, it is more efficient and suitable to use machine learning models to generate rasterized crop yield maps[26–29].

Despite the above-mentioned advantages of machine learning algorithms in crop yield estimation, difference and instability were observed in their performance resulting in constrained accuracy for yield estimations[30]. To address these issues, ensemble learning methods have recently gained prominence as a powerful technique in predictive modelling[31]. Ensemble learning involves combining the predictions of multiple models to achieve improved results over those of any single model. By aggregating diverse models, ensemble methods can reduce bias, variance, or both, and capture the underlying data distribution better, thereby yielding more accurate estimation. For example, it is found that all ensemble learning models (with lower prediction bias) outperformed individual machine learning models in predicting corn yield in three US Corn Belt states[32]. In addition, numerous studies demonstrated the effectiveness of ensemble learning in various applications, highlighting its potential for advancing the precision of agricultural yield estimations[33–35]. Yet, there is a lack of research on generating rasterized soybean yield maps using ensemble learning methods.

Soybean production region		Sub-regions	
NPR	Northern Production Region	NPR1	Northern Subregion of Northeast China for Soybean Cultivation
		NPR2	Central and Southeastern Subregion of Northeast China for Soybean Cultivation
		NPR3	North China Plateau Subregion for Soybean Cultivation
		NPR4	Northwest Subregion for Soybean Cultivation
HPR	Huang-Huai-Hai Production Region	HPR5	Northern Huang-Huai-Hai Subregion for Soybean Cultivation
		HPR6	Southern Huang-Huai-Hai Subregion for Soybean Cultivation
SPR	Southern Production Region	SPR7	Middle and Lower Yangtze River Subregion for Soybean Cultivation
		SPR8	Southwest Subregion for Soybean Cultivation
		SPR9	Central and Southern Subregion
		SPR10	South China Subregion for Soybean Cultivation

Table 1. Soybean Production Regions and Sub-regions of China.

China is one of the largest soybean producers in the world. While spatialization products for wheat, rice, and corn already exist in China, such products are lacking for soybean, which is also a staple food crop. Therefore, this study aims to use ensemble learning methods and spatial decomposition techniques to generate a rasterized soybean yield map for China through the fusion of multi-source data. Specifically, our objectives are: 1) to quantify the performance of 20 machine learning models as meta-models and base models in an ensemble setting; 2) to establish the stacking models by determining the number and type of variables for base models in the ensemble; 3) to provide a rasterized soybean yield dataset for China in 2001–2020 and conduct external cross-validation to verify its accuracy.

Methods

Study area. Influencing factors on soybean cultivation showed large variation across different regions in China due to its vast territory. Our study divided China into 3 major soybean production regions, which were furtherly divided into 10 sub-regions[36], as shown in Fig. 1. The 3 main production regions were the Northern Production Region (NPR), the Huang-Huai-Hai Production Region (HPR), and the Southern Production Region (SPR) (Fig. 1 and Table 1). These sub-regions were treated as dummy variables and included in the machine learning models in the subsequent analysis. Dummy variables are commonly used in regression analysis to represent categorical variables that have more than two levels. In addition, the soybean planting raster (Fig. 1) was extracted from Tibetan Plateau Data Center (TPDC)[37], and details of this dataset can be found in Adalibieke W *et al.*'s study[38].

Data collection. Data used in this study (Table 2), mainly includes: (1) Yield data: municipal- and county-scale soybean yield data, yield data recorded by agricultural meteorological stations, and three commonly used rasterized soybean yield datasets; (2) Environmental data: climate data, remote sensing data, management data, and soil data. Detailed information at the dataset level is described in the subsequent sections.

Soybean yield data. The soybean yield data at the municipal and county scales from 2001 to 2020 were all obtained from the statistical yearbooks of the cities and counties in China. These statistical yearbooks are readily available through searching the names of the cities or counties on the China Economic and Social Big Data Research Platform (https://data.cnki.net/). Unreasonable soybean yield data including those from regions with minimal soybean cultivation areas or those affected by administrative boundary adjustments were excluded during data compilation and statistical analysis. Moreover, data on soybean production from Hong Kong, Macau, Taiwan, and islands were not available. For other regions where direct yield data were not provided, it was calculated by dividing total soybean production by planting area. All yield data was standardized to units of t/ha. In total, the collected valid data comprised 3632 entries at the municipal scale and 13854 entries at the county scale. The soybean yield data recorded by agricultural meteorological stations was extracted from the National Meteorological Science Data Center (https://data.cma.cn/article/getLeft/id/251/keyIndex/6.html) upon reasonable request. Finally, we adopted all municipal-scale recorded yield data (2001–2020) to establish the model and generate the dataset. Then, we comprehensively validated our dataset and the commonly utilized rasterized soybean yield datasets (EarthStat[6], MapSPAM[7], and GDHY[8]) using both station- and county-scale recorded yield data to enhance the reliability of our results. It should be noted that we have previously used county-scale recorded yield data to establish the model and municipal-scale data for model validation and accuracy assessment. However, the model's performance was not as good as the results obtained using the current method.

Climate data. The most commonly used climate factors include temperature, precipitation, and solar radiation. These factors have impact on different stages of soybean growth and nitrogen-related processes, thereby affecting soybean yield[39,40]. In addition, drought occurs frequently in China with averaged frequency about every 2.7 years, which can significantly impact soybean yield[41]. PDSI assesses agricultural drought conditions by considering factors such as precipitation, soil moisture, and vegetation growth[42]. VPD is another factor have influence on stomatal conductance and photosynthesis, thus influencing soybean growth[43,44]. Therefore, in addition to commonly used climate factors, this study also incorporated PDSI and VPD as predictor variables to enhance the accuracy of soybean yield estimation.

Data type	Data name	Source	Time span	Spatiotemporal resolution
Yield Data	Administrative unit yield data	China economic and social big data research platform	2001–2020	Annual, municipal and county scales
	Station yield record	China Meteorological Data Service Centre	2001–2020	Annual, station scale
	Rasterized yield data	EarthStat	2005	5 years, 10 km
		MapSPAM	2005 and 2010	5 years, 10 km
		GDHY	2001–2016	Annual, 0.5°
Climate Data	Pre	NOAA PERSIANN-CDR	2001–2020	Daily, 0.25°
	Tmax, Tmin, PDSI, SRAD, VPD	TerraClimate	2001–2020	Monthly, 4 km
Remote Sensing Data	SIF	GOSIF	2001–2020	Monthly, 1 km
	NPP	MOD17A3HGF	2001–2020	Annual, 500 m
	NDVI	MOD13A3	2001–2020	Monthly, 1 km
	LSTd	MOD11A2	2001–2020	8 days, 1 km
Management Data	Soybean planting area	TPDC	2001–2020	Monthly, 10 km
	AS, MA, NPK, ONS, Urea			
	Sowing and harvest month	Dataset of Growth and Development of Major Crops in China	2001–2010	Annual, station scale
Soil Data	pH, OC, CEC_SOIL, REF_BULK, CLAY, SILT	Harmonized World Soil Database (HWSD)	2007	Static, 1 km
	SM	TerraClimate	2001–2020	Monthly, 4 km

Table 2. Data used in this study. Note: Pre: Precipitation; Tmax: Maximum temperature; Tmin: Minimum temperature; PDSI: Palmer drought severity index; SRAD: Solar radiation; VPD: Vapor pressure deficit; SIF: Solar-induced chlorophyll fluorescence; NPP: Net primary productivity; NDVI: Normalized difference vegetation index; LSTd: Daytime land surface temperature; AS: Ammonium sulfate; MA: Aanure; NPK: Nitrogen, phosphorus and potassium compounds; ONS: Other nitrogen straight; SM: Soil moisture; CEC_SOIL: Cation exchange capacity of the soil; OC: Organic carbon; REF_BULK: Reference bulk density. References for the data sources are provided in the subsequent main text.

Precipitation data were sourced from NOAA's PERSIANN-CDR dataset[45,46]. PERSIANN-CDR was created using an artificial neural network to estimate precipitation from remote sensing information, combined with bias correction using data from the Global Precipitation Climatology Project (GPCP). This dataset has a spatial resolution of 0.25° and a temporal resolution of 1 day. Other key climate data, including temperature, PDSI, SRAD, and VPD, were obtained from TerraClimate[47]. TerraClimate is a global climate dataset that integrates satellite observations, ground-based observations, and climate model simulation results. These data have a monthly temporal resolution and an approximate spatial resolution of 4 km. In addition, we did not use the precipitation data from TerraClimate because we intended to avoid potential correlations between different variables from the same data source. Such correlations might arise from similar data processing methods used within a single dataset.

Remote sensing data. Remote sensing data provide real-time monitoring and extensive spatial coverage, offering precise information on crop growth conditions and productivity. This technology has been widely employed in estimating soybean yield[48–50]. SIF (solar-induced chlorophyll fluorescence) from the GOSIF dataset shows a strong linear relationship with NPP at the ecosystem scale, significantly influencing soybean photosynthesis[51–53]. NDVI is frequently used as a variable in crop yield estimation due to its ability to assess vegetation cover, growth status, and health[26,54,55]. Numerous studies have utilized NDVI as a variable for monitoring soybean yield[56,57]. Compared to meteorological data and other factors influencing crop growth, remote sensing data can directly reflect and monitor crop growth conditions in real-time, demonstrating substantial potential for soybean yield estimation[16].

SIF data were obtained from the GOSIF dataset[51], covering the period from 2001 to 2020. GOSIF has a temporal resolution of 8 days and a spatial resolution of 1 km. This dataset is a global rasterized SIF dataset generated using machine learning models from SIF observations of the Orbiting Carbon Observatory-2 (OCO-2). Additionally, this study utilized three remote sensing data products: NPP, NDVI, and LSTd, sourced from the MOD17A3HGF[58], MOD13A3[59], and MOD11A2[60] datasets, respectively. MODIS satellite remote sensing datasets provide global coverage with medium spatial and temporal resolution, offering various remote sensing products[61].

Management data. The application of nitrogen fertilizer involves complex interactions with factors such as root activity and photosynthesis, which are crucial for crop yield[62]. Appropriate use of nitrogen fertilizer can enhance soybean yield[63]. Therefore, incorporating nitrogen fertilizer application as a predictor variable for soybean yield can contribute to more accurate yield mapping. The data on soybean harvested areas and the application of fertilizers such as AS, MA, NPK, Urea, and ONS were obtained from the global crop-specific nitrogen fertilizer dataset[38] hosted on TPDC. This dataset includes annual data on soybean planting areas and fertilizer usage from 2001 to 2020, with a spatial resolution of 10 km. Additionally, soybean growth period data at the station scale can be obtained from the Dataset of Growth and Development of Major Crops in China (https://data.cma.cn/article/getLeft/id/251/keyIndex/6.html) upon reasonable request.

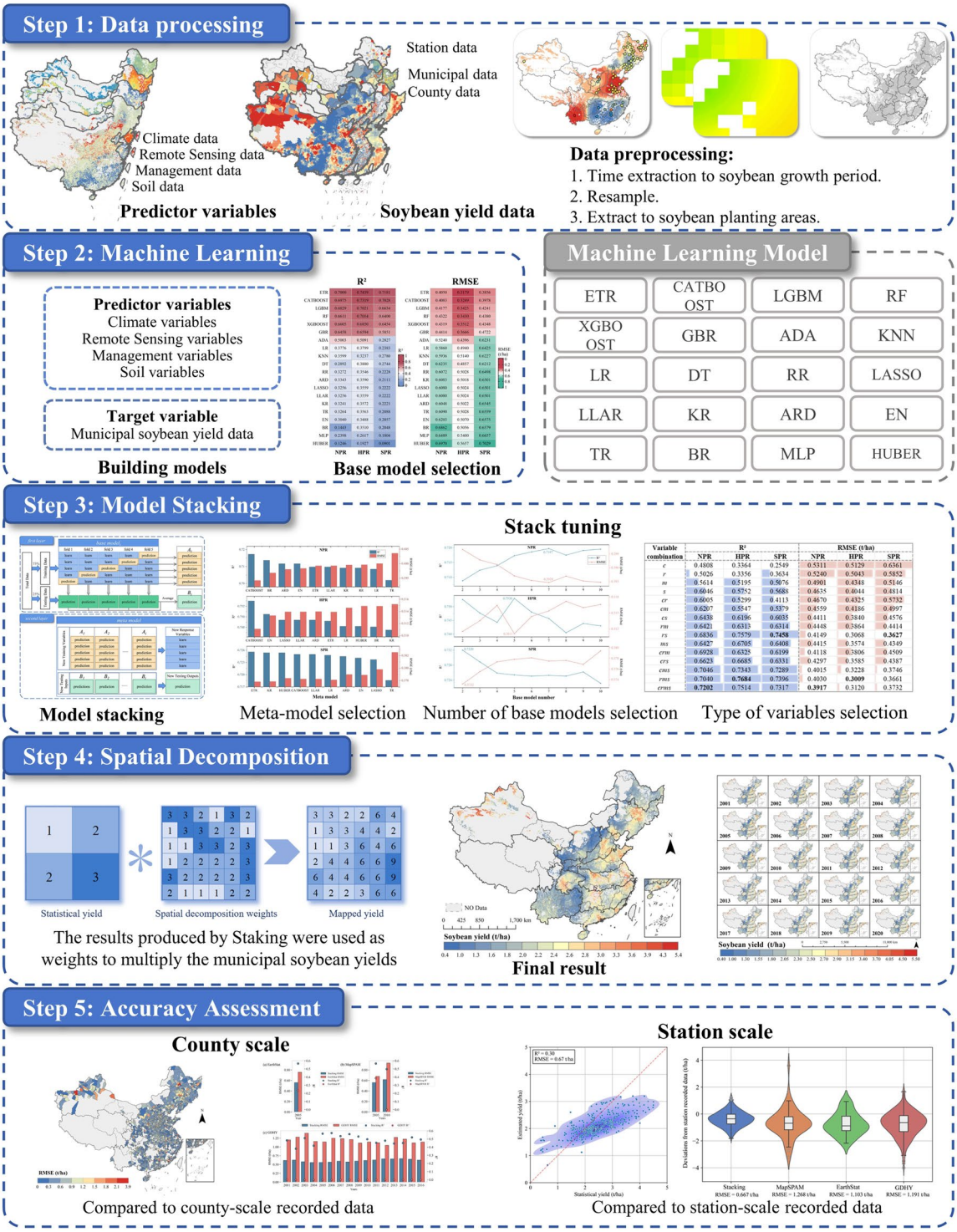

Fig. 2 Flowchart of the proposed model stacking methodology based on multiple data sources for mapping soybean yield.

Soil data. The soil in different regions of China varies significantly, e.g., black soil in the northeast and red soil in southern China. Different soil types are suitable for cultivating different soybean varieties[64] and soybean yield and quality are influenced by different soil characters. For instance, Anthony *et al.* found a negative correlation between soybean yield and soil pH across all locations and years[65]; Ferreira *et al.* confirmed that certain soil physicochemical properties, such as REF_BULK, OC, and SM, affect soybean yield[66]. Therefore, this study incorporated all these soil variables as predictors to improve the accuracy of soybean yield estimation.

The soil physicochemical properties data except SM were sourced from the Harmonized World Soil Database (HWSD) version 2.0, which compiles soil information from around the globe into a standardized, globally

Type	Variable Name	Description	Unit
Climate Variables	Pre	Precipitation	mm
	Tmax	Maximum temperature	°C
	Tmin	Minimum temperature	°C
	PDSI	Palmer drought severity index	Dimensionless
	SRAD	Solar radiation	W/m^2
	VPD	Vapor pressure deficit	kPa
Remote Sensing Variables	SIF	Solar-induced chlorophyll fluorescence	$mW{\cdot}m^{-2}{\cdot}nm^{-1}{\cdot}sr^{-1}$
	NPP	Net primary productivity	kg C/m
	NDVI	Normalized difference vegetation index	Dimensionless
	LSTd	Daytime land surface temperature	°C
Management Variables	year	Year	year
	SBZ	Soybean sub-zones	Dimensionless
	AS	Ammonium sulfate Application	kg N/ha
	MA	Manure Application	kg N/ha
	NPK	Nitrogen, phosphorus and potassium compounds	kg N/ha
	Urea	Urea Application	kg N/ha
	ONS	Other nitrogen straight Application	kg N/ha
Soil Variables	SM	Soil moisture	mm
	pH	Potential of hydrogen	Dimensionless
	OC	Organic carbon	%
	CEC_SOIL	Cation exchange capacity of the soil	cmol/kg
	REF_BULK	Reference bulk density	g/cm^3
	CLAY	Clay content	%
	SILT	Silt content	%

Table 3. The predictor variables used in this study.

Model Name	Abbreviation	Type
AdaBoost Regressor	ADA	Ensemble Learning
Automatic Relevance Determination	ARD	Bayesian Linear
Bayesian Ridge	BR	Bayesian Regression
CatBoost Regressor	CATBOOST	Gradient Boosting
Decision Tree Regressor	DT	Decision Tree
Elastic Net	EN	Linear Model
Extra Trees Regressor	ETR	Tree Model
Gradient Boosting Regressor	GBR	Gradient Boosting
Huber Regressor	HUBER	Linear Model
K Neighbors Regressor	KNN	Nearest Neighbors
Kernel Ridge	KR	Kernel Method
Lasso Regression	LASSO	Linear Model
Light Gradient Boosting Machine	LGBM	Gradient Boosting
Lasso Least Angle Regression	LLAR	Linear Model
Linear Regression	LR	Linear Model
MLP Regressor	MLP	Neural Network
Random Forest Regressor	RF	Random Forest
Ridge Regression	RR	Linear Model
TheilSen Regressor	TR	Tree Model
Extreme Gradient Boosting	XGBOOST	Gradient Boosting

Table 4. The 20 machine learning models used in this study.

consistent soil dataset. HWSD provides soil data at a 1-km resolution from 2007[67]. As to SM data, they were obtained from TerraClimate[47], which estimates soil water content by integrating climate inputs with a hydrological model. We used soil data for a depth of 1 meter, as the root systems of soybeans are generally confined to this depth range. Additionally, soil data may be provided in multiple layers. If so, we would weight-average the data according to the thickness of each soil layer.

Model	Model parameters
ADA	base_estimator = None, n_estimators = 50, learning_rate = 1.0
ARD	n_iter = 300, alpha_1 = 1e-6, alpha_2 = 1e-6, lambda_1 = 1e-6, lambda_2 = 1e-6
BR	n_iter = 300, alpha_1 = 1e-6, alpha_2 = 1e-6, lambda_1 = 1e-6, lambda_2 = 1e-6
CATBOOST	iterations = 1000, learning_rate = 0.03, depth = 6
DT	max_depth = None, min_samples_split = 2, min_samples_leaf = 1
EN	alpha = 1.0, l1_ratio = 0.5, max_iter = 1000
ETR	n_estimators = 100, max_depth = None, max_features = 'auto'
GBR	n_estimators = 100, learning_rate = 0.1, max_depth = 3
HUBER	epsilon = 1.35, max_iter = 100, alpha = 0.0001
KNN	n_neighbors = 5, weights = 'uniform', algorithm = 'auto'
KR	alpha = 1, kernel = 'linear', gamma = None
LASSO	alpha = 1.0, max_iter = 1000, selection = 'cyclic'
LGBM	num_leaves = 31, learning_rate = 0.1, n_estimators = 100
LLAR	n_nonzero_coefs = 500, fit_intercept = True
LR	fit_intercept = True, normalize = False
MLP	hidden_layer_sizes = (100,), activation = 'relu', solver = 'adam'
RF	n_estimators = 100, max_depth = None, max_features = 'auto'
RR	alpha = 1.0, fit_intercept = True, normalize = False
TR	max_iter = 300, fit_intercept = True
XGBOOST	n_estimators = 100, learning_rate = 0.1, max_depth = 3

Table 5. Machine learning model parameters.

Modelling methods. Figure 2 delineates the entire process of estimating soybean yield in China based on a stacking model. Initially, the collected data on soybean, climate, remote sensing, management, and soil were preprocessed through synthesized by growth stages, resampling, and masking according to growth regions. Subsequently, we evaluated the fitting performance of 20 machine learning models using climate, remote sensing, management, and soil data as predictor variables and municipal soybean yield as the response variable in each soybean production region. The five machine learning models with the best fitting performance were selected as base models. We then assessed the performance of 20 meta models in the stacking ensemble, selecting the best-performing meta model for each region. By further refining the number of base models and the combination of predictor variables, we enhanced the performance of the ensemble model. Ultimately, we established an ensemble model for estimating soybean yield in each region. The simulated rasterized soybean yield was used as a weight to spatially disaggregate the municipal-scale soybean yield, developing 1-km annual rasterized soybean yield maps for China. Finally, the results of this study, along with three commonly used datasets (EarthStat, MapSPAM, and GDHY), were then aggregated to the county-scale or extracted to the station-scale to be comprehensively validated using recorded data at county or station scales.

Data preprocessing. First, we standardized the temporal scale of the data to align accurately with the soybean growth cycle by synthesizing the data based on different growth periods. By applying the kriging interpolation method to station-scale data during the soybean growth period, we generated base maps for the sowing and harvesting months. Variables with available data within the growth period were synthesized accordingly, while variables with a temporal resolution of one year or less were not processed. Next, all remote sensing images were resampled to a spatial resolution of 1-km to standardize pixel sizes and positions. Finally, using the soybean planting area as a base map, all images were masked to extract data for the soybean planting area. All data were aggregated at the municipal- and county-scale, providing the data required for model training and evaluation. The final processed predictor variables in this study can be found in Table 3.

Machine learning models. To fully leverage the capabilities of machine learning models in monitoring soybean yield, we firstly adopted a machine learning library PyCaret 3.3.1 in Python, to compare the performance of 20 machine learning models both as base models and meta models (Table 4). We tried a variety of machine learning models with the aim of screening out the better-performing ones, and the exact number of candidate models has little impact on the subsequent analysis. To ensure a fair comparison, all models were used with their default parameters (Table 5). Results showed that ETR and CATBOOST consistently exhibited outstanding performance across different regions (Fig. 5). In addition, XGBOOST, LGBM, and RF also showed excellent performance. In summary, ETR, CATBOOST, XGBOOST, LGBM, and RF were used as base models. Detailed information about these 5 models was shown in the following two paragraphs, while detailed information about the other models is not provided in this study due to space limitations.

CATBOOST[68] is a machine learning model based on gradient boosting decision trees. This model can effectively handle categorical features and minimize information loss. The ETR[69] model constructs multiple decision trees and combines their results to produce final predictions. Unlike other models, ETR uses the entire training sample at each split point rather than a random sample, and it selects split points randomly rather than

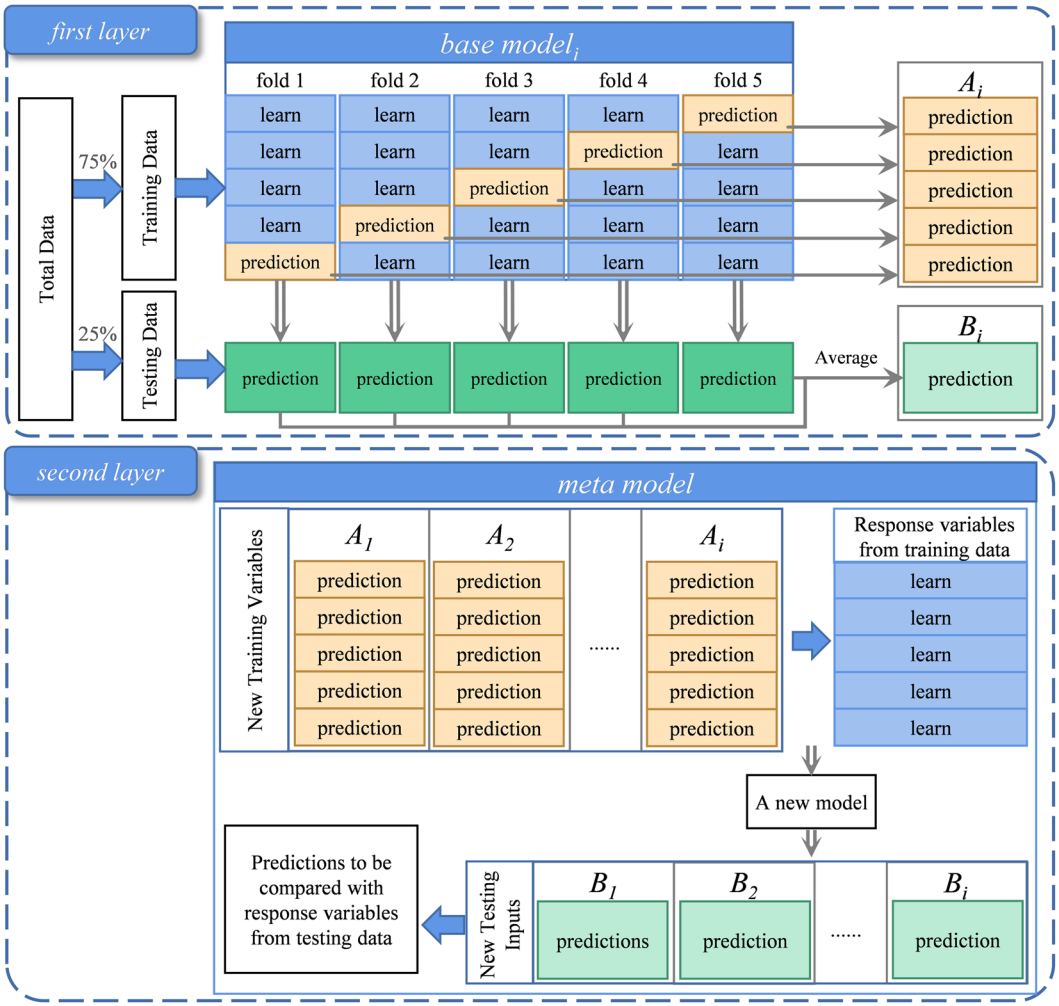

Fig. 3 Schematic representation of the proposed 5-fold cross-validation based stacking model approach.

optimally. Thus, this approach increases model diversity, reduces the risk of overfitting, and typically enhances prediction accuracy while speeding up the tree-building process due to its randomness.

XGBOOST[70] is an efficient gradient boosting decision tree (GBDT) algorithm that employs various optimization techniques including approximate greedy algorithms, distributed computing, and caching to accelerate the training process. Consequently, XGBOOST has its advantage in superior accuracy and speed. LGBM[71] is another GBDT algorithm that uses a histogram-based decision tree algorithm to discretize continuous features, significantly reducing memory consumption and computation time. RF[72] is a popular machine learning model widely used for crop yield prediction. It generates prediction results by multiple decision trees and aggregates them through voting or averaging. Compared to a single decision tree, RF reduces overfitting and is more robust in handling high-dimensional and missing data.

Model stacking. Stacking is an ensemble learning algorithm that utilizes the output variables of multiple base models as feature variables to train a meta-model, which predicts the final target variable[73]. In stacking, the performance of the ensemble model can be enhanced by selecting the base models and meta-model with the best evaluation results. To our knowledge, this algorithm has rarely been applied in studies on yield spatialization. Figure 3 illustrates the principle of the stacking model. In this study, a 5-fold cross-validation method was used to combine base models.

In the first layer, during the execution of each base model, 75% of the data is used as the training set, and 25% as the test set. The training set is then randomly divided into five parts. Each time, four parts are selected to train the model, and the remaining one part, along with the test set, is used for prediction. This process is repeated five times to obtain predictions for both the training and test sets. The predictions for the training set are stacked together to generate predictions consistent in length with the training set, while the test set predictions are averaged to produce predictions consistent in length with the test set.

In the second layer, the predictions from the i_{th} base models for the training set serve as input variables, and the training set serves as the response variable to train the meta-model. The test set predictions from the i_{th} base models are used as input variables, and the predictions of the meta-model are compared with the actual

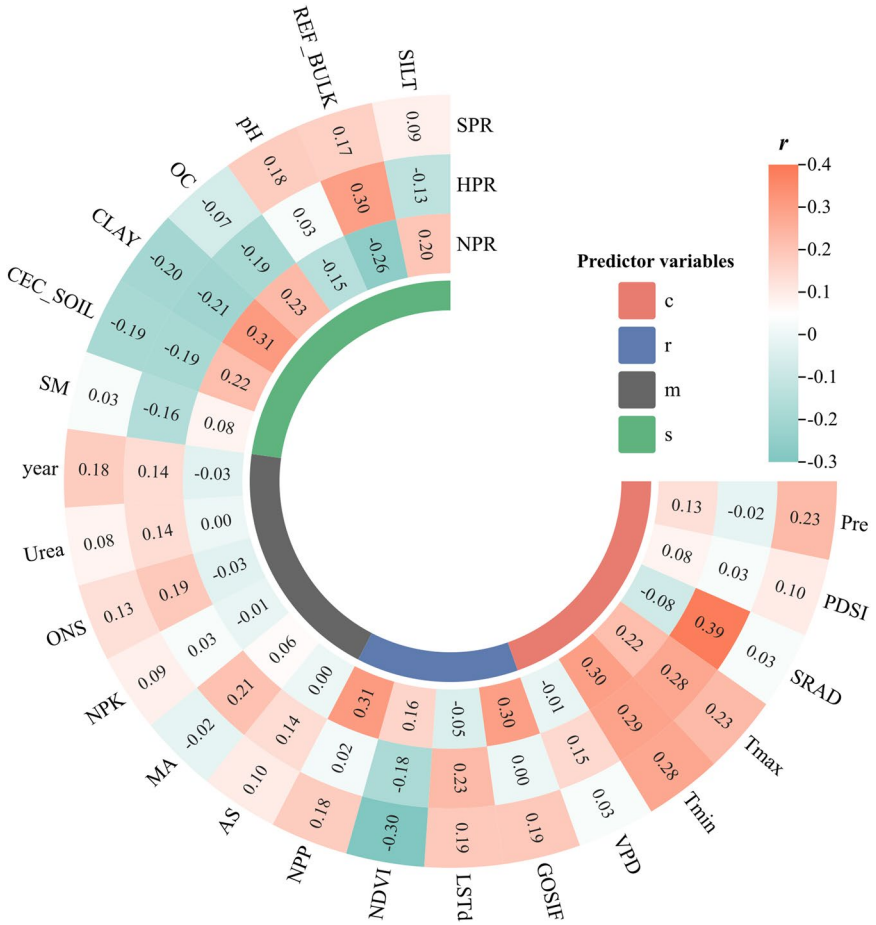

Fig. 4 The Pearson correlation between soybean yield and predictor variables in different production regions. Note: *c*: climate variables; *r*: remote sensing variables; *m*: management variables; *s*: soil variables.

test set to evaluate the performance of the stacking model. Since the number of base models can also affect the performance of the ensemble model, this study initially employs the five best-performing base models to select the most effective meta-model. Subsequently, the precision of ensemble models composed of different numbers of base models is compared to determine the optimal number of base models for constructing the soybean yield estimation ensemble model.

Spatial decomposition. After applying ensemble learning modeling, soybean yield was estimated for each production region, resulting in rasterized soybean yield simulation maps. These maps were then used as weights to spatially disaggregate the municipal soybean yield statistics. First, the annual estimated soybean yield from 2001 to 2020 were converted into spatial disaggregation weights w_{cti}:

$$w_{cti} = \frac{y_{cti}^{pred}}{\frac{1}{I}\sum_{i=1}^{I} y_{cti}^{pred}} \quad (1)$$

where y_{cti}^{pred} is the predicted soybean yield for grid cell i in city c for year t, and I is the total number of grid cells within the city. Next, the municipal-scale statistical soybean yield data Y_{ct}^{stat} for city c were disaggregated to the grid scale, generating the spatially disaggregated soybean yield for each grid cell y_{cti}:

$$y_{cti} = w_{cti} \cdot Y_{ct}^{stat} \quad (2)$$

Accuracy assessment. The results of this study, ChinaSoyYield1km[74], along with three commonly used datasets (EarthStat, MapSPAM, and GDHY), were then aggregated to the county-scale or extracted to the station-scale to be comprehensively validated using recorded data at county or station scales. Specifically, we calculated the coefficient of determination (R^2) and root mean square error (RMSE) between the recorded data and four datasets: EarthStat, MapSPAM, GDHY, and ChinaSoyYield1km. Both EarthStat and MapSPAM have a spatial resolution of 10 km but only provide data for specific years. In contrast, GDHY contains data from 2001 to 2016, but

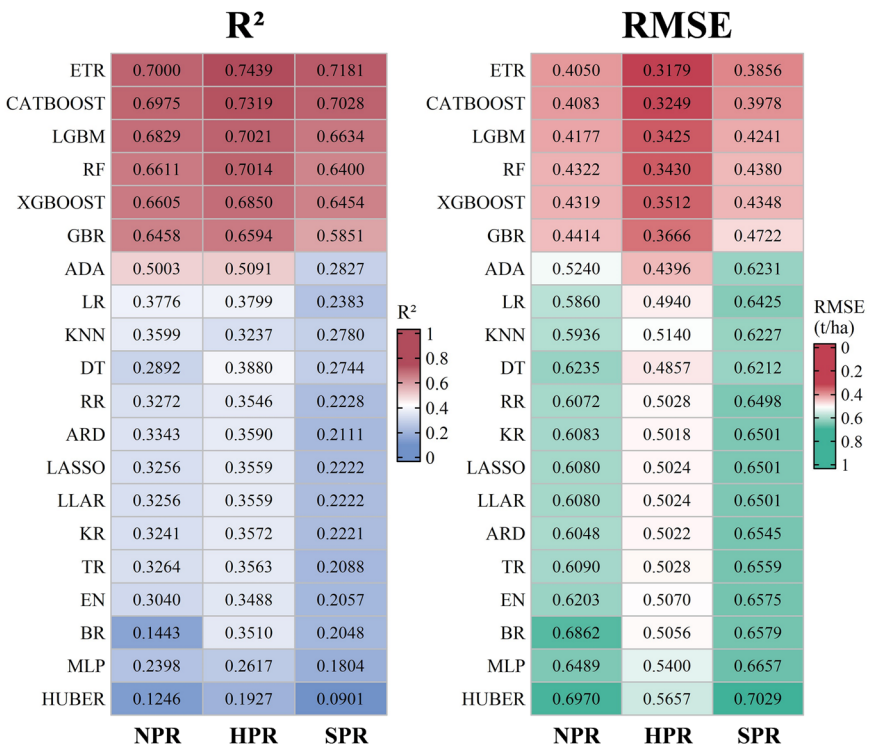

Fig. 5 The average of model performance measurements (R^2 and RMSE) with 20 machine learning models.

with a spatial resolution of only 55 km. Therefore, this study enhances the credibility of accuracy assessment by conducting a comprehensive comparison with these three commonly used datasets. The coefficient of determination (R^2) and root mean square error (RMSE) were given by the following equations:

$$R^2 = \left(\frac{\sum_{i=1}^{n}(O_i - \overline{O})(P_i - \overline{P})}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2}\sqrt{\sum_{i=1}^{n}(P_i - \overline{P})^2}} \right)^2 \quad (3)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \quad (4)$$

where n is the number of samples, O_i and P_i denote statistical and estimated soybean yield, respectively; correspondingly, $\overline{O}$ and $\overline{P}$ represents the mean of statistical and estimated soybean yield. Generally, model's performance become more accurate as RMSE approaching to 0 and R^2 approaching to 1.

Model fitting. *Correlation analysis of all variables.* Correlation analysis clearly illustrates whether there are significant relationships between each predictor variable and the target variable. Figure 4 shows the correlation between soybean yield and each predictor variable for each production region. Among all variables, climate variables show the strongest correlation with soybean yield. Specifically, Tmax and Tmin exhibit a strong positive correlation with soybean yield in all three regions, indicating that soybean growth is highly sensitive to temperature.

Among the remote sensing variables, GOSIF and NPP display a strong positive correlation with soybean yield in NPR, while LSTd shows a strong positive correlation with soybean yield in HPR. In SPR, due to the warm and humid climate, the vegetation index is generally higher. However, soybean, compared to tall crops like corn or certain grains, have smaller leaf areas and thus lower vegetation indices, resulting in a strong negative correlation between NDVI and soybean yield. In comparison to other variables, management variables show a weaker correlation with soybean yield but exhibit distinct characteristics.

All management variables have a negative correlation with soybean yield in NPR, while they show a positive correlation in both HPR and SPR. The correlations of soil variables with soybean yield vary significantly across different regions. CEC_SOIL, CLAY, and OC exhibit a strong positive correlation with soybean yield in NPR but a negative correlation in HPR and SPR. Conversely, REF_BULK and pH show a strong negative correlation with soybean yield in NPR and a strong positive correlation in HPR and SPR.

In summary, these variables show a significant correlation with soybean yield across three production regions. Therefore, they can be utilized as predictor variables in subsequent soybean yield forecasting models. In addition, we did not consider the correlations among the predictor variables or the potential issue of multicollinearity. This is because we prioritize model accuracy over the interpretability of variable impacts, so we prefer to retain more information in our subsequent model construction.

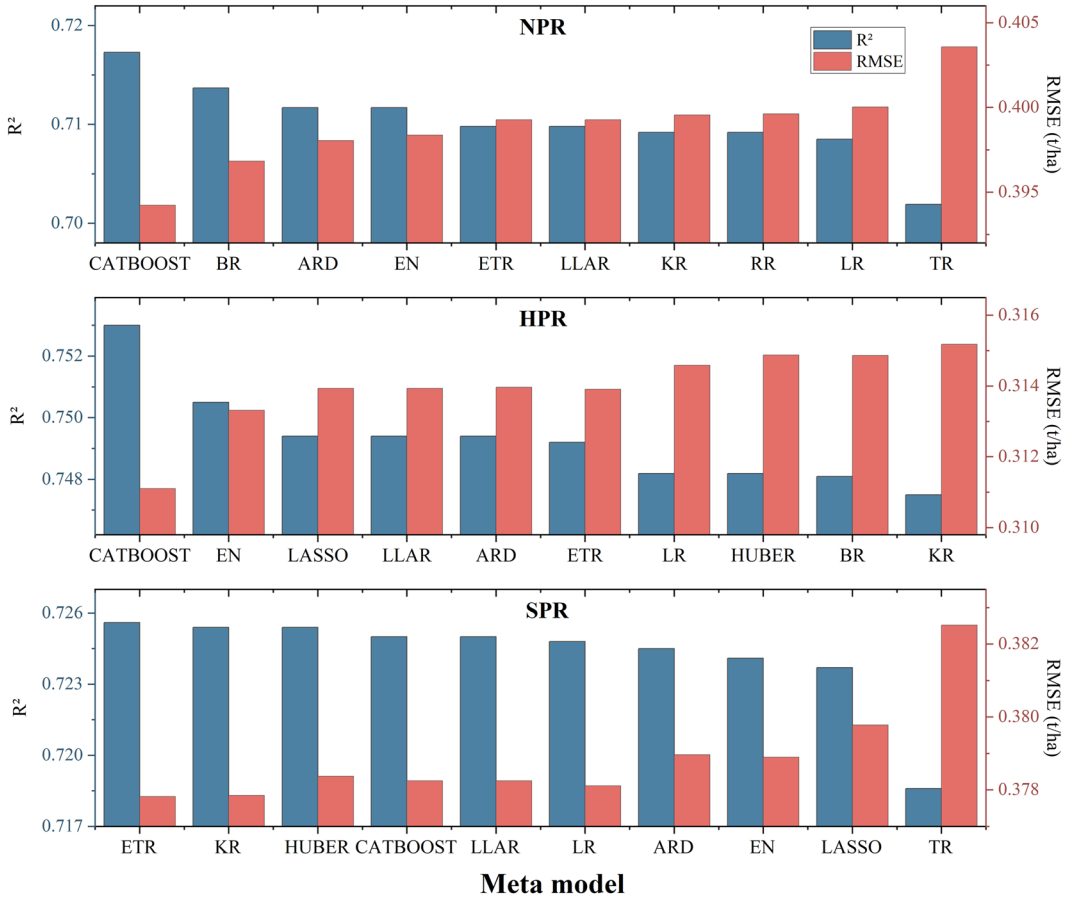

Fig. 6 The top 10 meta-models with the best performance in terms of R^2 and RMSE in different soybean production regions.

Performance evaluation of individual models. We ran 20 machine learning models separately in the three soybean production regions, evaluating their performance by comparing R^2 and RMSE, as shown in Fig. 5. The five best-performing models were ETR, CATBOOST, LGBM, RF, and XGBOOST. These models achieved R^2 values above 0.64 and RMSE values below 0.44 t/ha in all regions. Consequently, these five models were selected as the base models for subsequent meta-model selection.

Selection of meta-model in stacking. Meta-model plays a crucial role in stacking models by reducing bias among individual models and enhancing the generalization ability of the ensemble. We used the five best-performing models as base models and trained 20 meta-models separately in the three production regions. Figure 6 displays the top 10 meta-models with the best simulation performance for each soybean production region. Compared to individual base models, the stacking ensemble model achieved improved modeling accuracy in each soybean production region. CATBOOST performed the best as a meta-model in NPR, with R^2 improving to a maximum of 0.72 and RMSE decreasing to a minimum of 0.39 t/ha. Similarly, in HPR, CATBOOST also exhibited the best performance, with R^2 improving to a maximum of 0.75 and RMSE decreasing to a minimum of 0.31 t/ha. In SPR, however, ETR demonstrated the best performance as a meta-model, with R^2 improving to a maximum of 0.73 and RMSE decreasing to a minimum of 0.38 t/ha. Therefore, we selected CATBOOST as the meta-model for constructing ensemble models in NPR and HPR, and ETR as the meta-model in SPR.

Selection of the number of base models. The results of the base models serve as the variables for the meta-model. Therefore, the number of base models can significantly impact the performance of the ensemble model. We used the meta-models to select the optimal number of base models for ensemble model performance in each of the three production regions. Figure 7 illustrates the performance of the meta-model when selecting the top 2 to top 10 base models for each soybean production region, showing different trends across regions. In comparison to fixing the number of base models at 5, in NPR, increasing the number of base models to 7 resulted in the meta-model achieving the highest performance improvement, with R^2 increasing from 0.7173 to 0.7197 and RMSE decreasing from 0.3942 to 0.3928. In HPR, the best performance of the meta-model was achieved with 5 base models. In SPR, the optimal performance of the meta-model was observed with 2 base models, where R^2 improved from 0.7256 to 0.7320 and RMSE decreased from 0.3778 to 0.3732. Therefore, in this study, we selected the top 7, 5, and 2 models based on their performance as base models for NPR, HPR, and SPR, respectively.

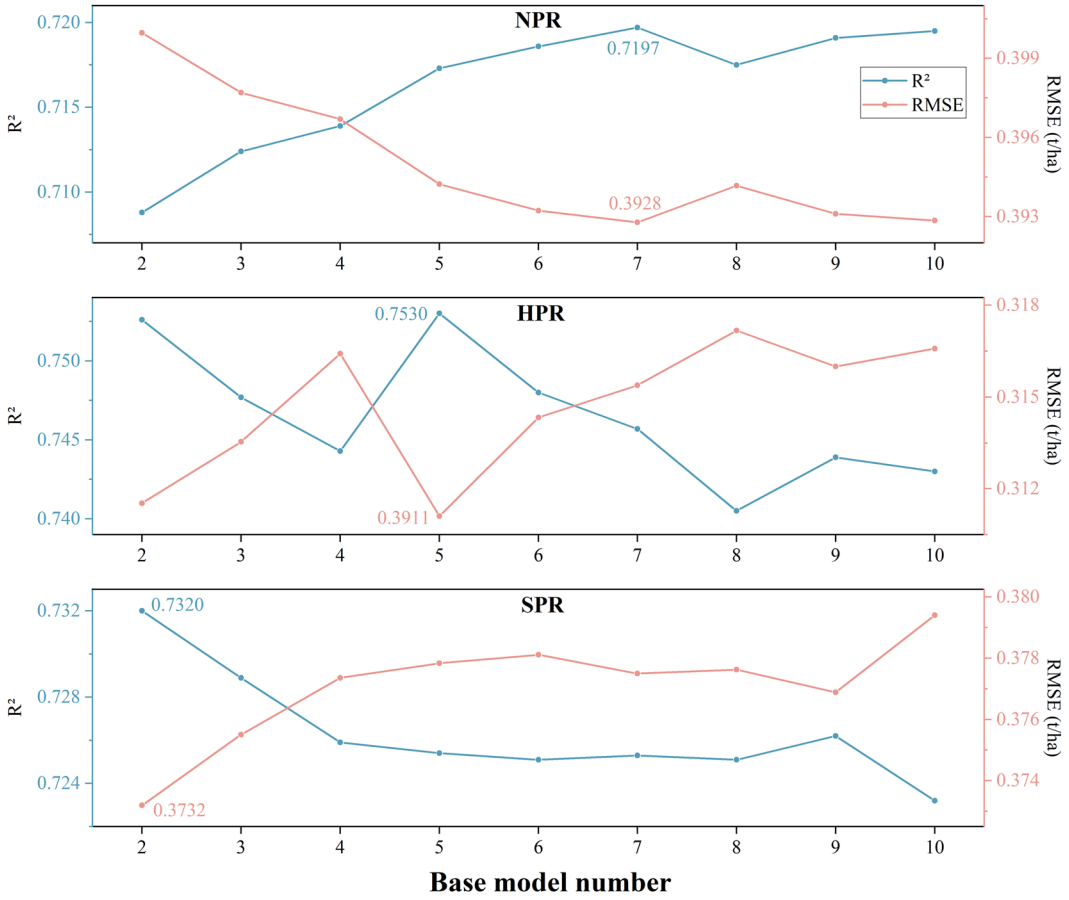

Fig. 7 The R^2 and RMSE of the ensemble model when selecting 2 to 10 base models.

Variable combination	R^2			RMSE (t/ha)		
	NPR	HPR	SPR	NPR	HPR	SPR
c	0.4808	0.3364	0.2549	0.5311	0.5129	0.6361
r	0.5026	0.3356	0.3634	0.5240	0.5043	0.5852
m	0.5614	0.5195	0.5076	0.4901	0.4348	0.5146
s	0.6046	0.5752	0.5688	0.4635	0.4044	0.4814
cr	0.6005	0.5299	0.4113	0.4670	0.4325	0.5732
cm	0.6207	0.5547	0.5379	0.4559	0.4186	0.4997
cs	0.6438	0.6196	0.6035	0.4411	0.3840	0.4576
rm	0.6421	0.6313	0.6314	0.4448	0.3864	0.4414
rs	0.6836	0.7579	**0.7458**	0.4149	0.3068	**0.3627**
ms	0.6427	0.6705	0.6408	0.4415	0.3574	0.4349
crm	0.6928	0.6325	0.6199	0.4118	0.3806	0.4509
crs	0.6623	0.6685	0.6331	0.4297	0.3585	0.4387
cms	0.7046	0.7343	0.7289	0.4015	0.3228	0.3746
rms	0.7040	**0.7684**	0.7396	0.4030	**0.3009**	0.3661
crms	**0.7202**	0.7514	0.7317	**0.3917**	0.3120	0.3732

Fig. 8 The ensemble learning fitting performance of different variable combinations in different soybean production regions. Note: *c*: climate variables; *r*: remote sensing variables; *m*: management variables; *s*: soil variables; *cr*: climate and remote sensing variables; *cm*: climate and management variables; *cs*: climate and soil variables; *rm*: remote sensing and management variables; *rs*: remote sensing and soil variables; *ms*: management and soil variables; *crm*: climate, remote sensing and management variables; *crs*: climate, remote sensing and soil variables; *cms*: climate, management and soil variables; *rms*: remote sensing, management and soil variables; *crms*: climate, remote sensing, management and soil variables.

Predictor variables selection. Different combinations of predictor variables can affect the performance of machine learning models. While multidimensional variables may enhance performance, they can also lead to

Soybean Producing Regions	Base Model	Meta Model	Use Variables
NPR	ETR CATBOOST LGBM RF XGBOOST GBR ADA	CATBOOST	*crms*
HPR	ETR CATBOOST LGBM RF XGBOOST	CATBOOST	*rms*
SPR	ETR CATBOOST	ETR	*rs*

Table 6. Basic parameters and used variables of the Stacking model for each production region.

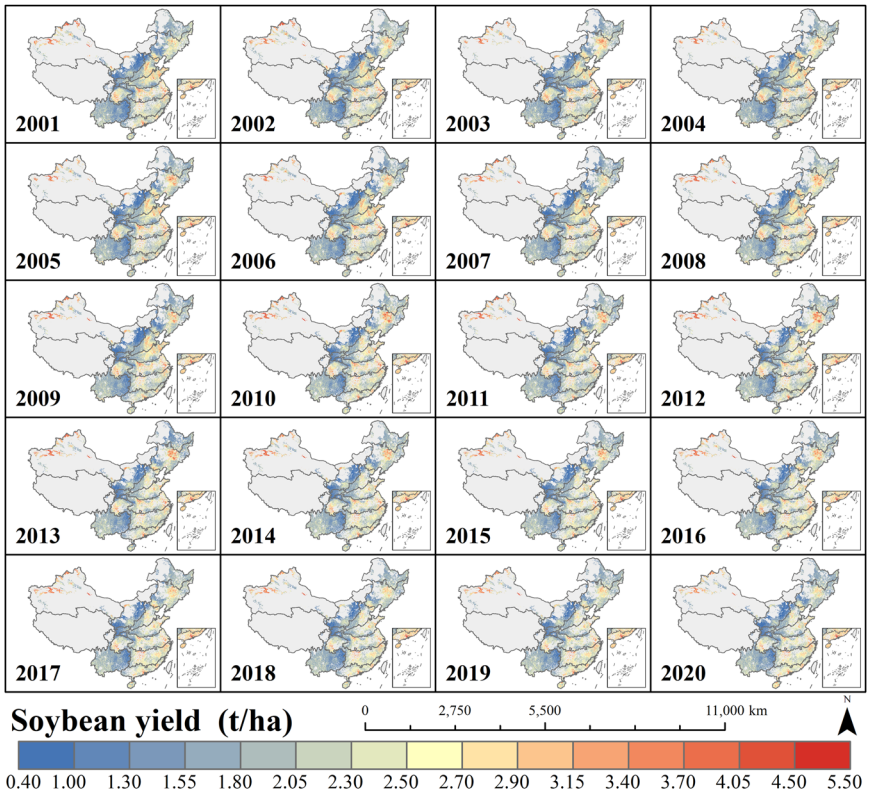

Fig. 9 Visualization of the ChinaSoyYield1km dataset.

overfitting. We evaluated the performance of the meta-model across 15 different combinations of variables in Fig. 8, based on the optimal meta-models and base models. In NPR, the meta-model achieves optimal performance when all four predictive variables are inputted. In HPR, the meta-model's performance peaks when the input variable is *rms*, with R^2 improving from 0.7514 to 0.7684 and RMSE decreasing from 0.3120 to 0.3009. Similarly, in SPR, further performance enhancement is observed when the input variable is *rs*, with R^2 increasing from 0.7317 to 0.7458 and RMSE decreasing from 0.3732 to 0.3627. Therefore, in this study, we selected *crsm*, *rms*, and *rs* as the input variables for the ensemble model in NPR, HPR, and SPR, respectively. Finally, by selecting meta-models, base models, and input variables, we constructed soybean yield estimation ensemble models separately for the three production regions. Details are presented in Table 6.

Data Records

The derived yield dataset for ChinaSoyYield1km[74] during 2001–2020 is available at https://doi.org/10.57760/sciencedb.18390. The dataset is stored in GeoTiff format under the EPSG: 4326 (GCS_WGS_1984) spatial reference, with the unit of kg/ha. We did not use the unit of kg/ha as presented in this study because using the unit of kg/ha can reduce the data file size by half, making it more convenient for users to download and store. The maps can be visualized and analyzed using software such as ArcGIS, QGIS, or similar applications (Fig. 9).

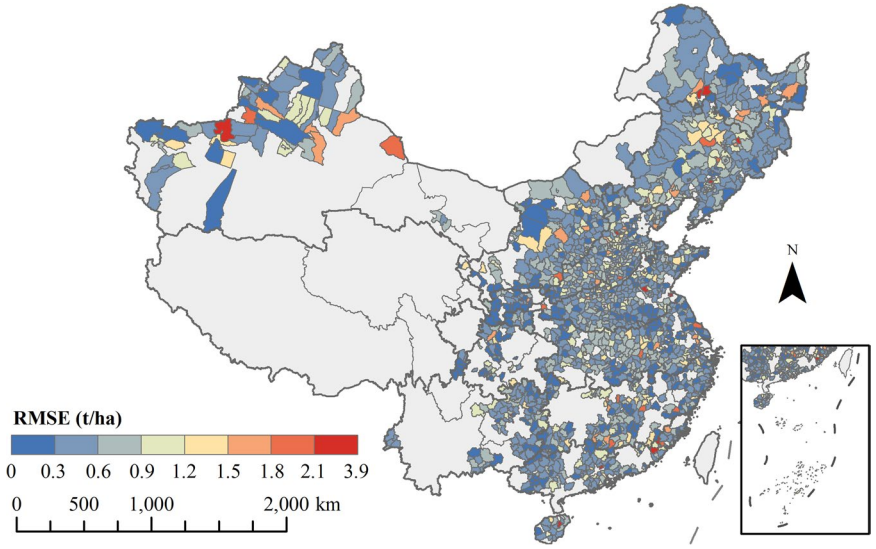

Fig. 10 The RMSE between the results of our stacking model and recorded data at the county scale.

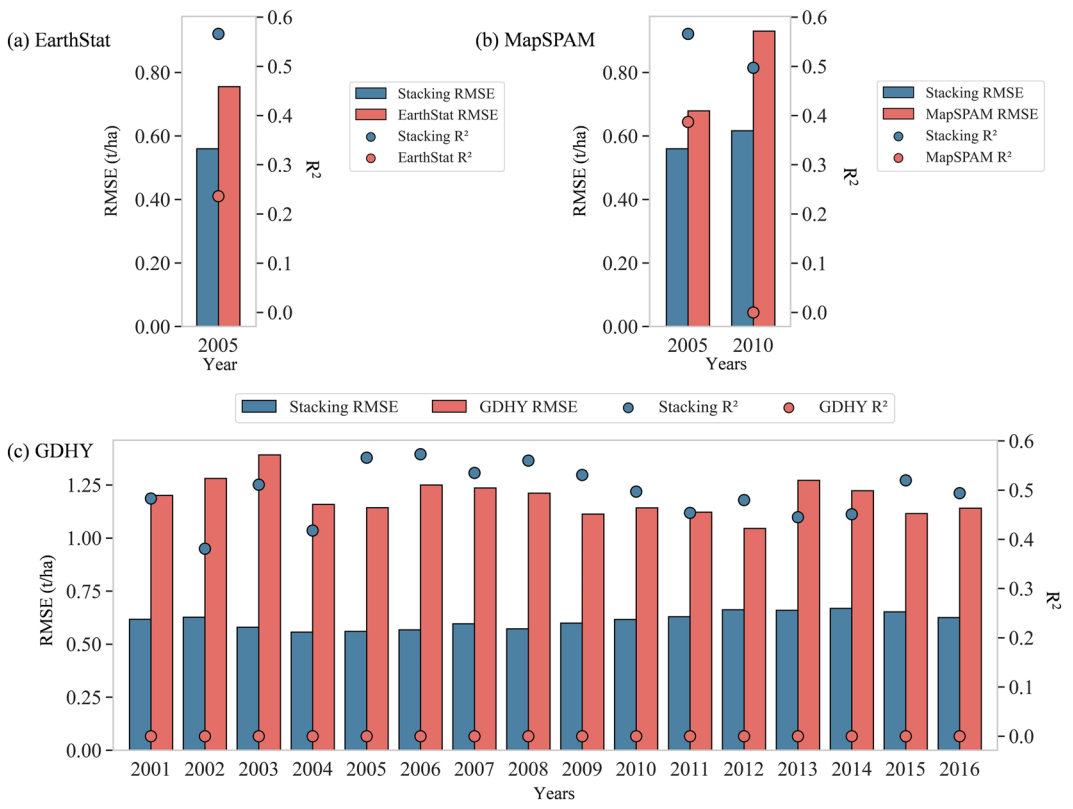

Fig. 11 The RMSE and R^2 between the recorded data at the county scale and the results of our stacking model as well as three existing rasterized yield datasets (EarthStat, MapSPAM, and GDHY).

Technical Validation

We compared four datasets, including the ChinaSoyYield1km, EarthStat, MapSPAM, and GDHY, with recorded yield data at both county and station scales (data not used in our model development process). Overall, both at the county and station scales, the soybean yield estimates in this study demonstrate higher accuracy compared to the three commonly used datasets.

We first aggregated the 2001–2020 ChinaSoyYield1km estimates and the EarthStat, MapSPAM, and GDHY rasterized yield estimates to the county scale and compared them with recorded data. Figure 10 shows the root mean square error (RMSE) between the ChinaSoyYield1km and recorded soybean yield. At the county scale,

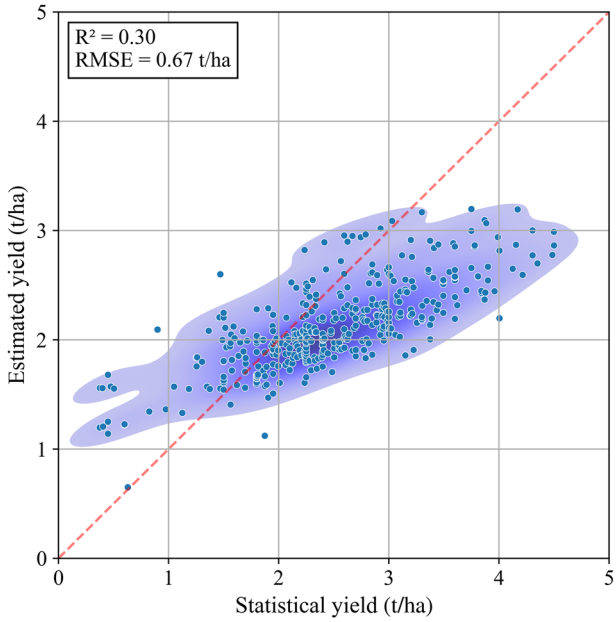

Fig. 12 The comparison between station recorded soybean yield and the results of our stacking model.

the RMSE between the results of this study and the recorded data is generally within the range of 3.90 t/ha. In over 90% of the regions, the RMSE for soybean yield is within 2.10 t/ha, with only a few counties having RMSE outside the reasonable range, indicating the high accuracy of the soybean yield data generated in this study.

Figure 11 presents the R^2 and RMSE for four datasets, each compared to the recorded data. Due to the temporal resolution limitations of the EarthStat and MapSPAM data, comparisons could only be made for specific years. It can be seen that in 2005 and 2010 (Fig. 11a and b), the R^2 values for our dataset were over 0.50, and when compared to GDHY (Fig. 11c), they were also above 0.50 in most years. This indicates that our resulting dataset captures over 50% of the yield variability at the county scale, demonstrating superior accuracy compared to publicly available datasets. The RMSE results also illustrate the same fact. The RMSE of our dataset compared to the county-scale recorded data was lower than that of existing datasets in all years, with reductions ranging from 0.18 to 0.60 t/ha.

We extracted the station-specific yield estimates from the ChinaSoyYield1km dataset, as well as from the EarthStat, MapSPAM, and GDHY datasets, and compared them with station-recorded data. Figure 12 shows that the R^2 between the results of this study and the station-recorded data is 0.30, with an RMSE of 0.67 t/ha. The station-recorded data mainly range between 1.50–3.00 t/ha, within which the soybean yield estimates from this study show high consistency with the station-recorded data. Figure 13 compares the RMSE of the deviations between the estimated yield from the four datasets and the station-recorded yield. The data from this study show the smallest deviations, all within 2.00 t/ha, whereas the deviations for MapSPAM and GDHY exceed 3.00 t/ha. The average and peak deviations of the data from this study are closest to 0 t/ha when compared to station-recorded yield. Although EarthStat shows fewer extreme values, its average deviation from station-recorded yield is the largest. Due to its lower spatial resolution, GDHY shows relatively lower accuracy at the station scale and contains numerous extreme values. The RMSE between the results of this study and the station-recorded yield is 0.67 t/ha, which is smaller than the RMSE of other raster yield datasets when compared to station-recorded data.

Usage Notes

Advantages of ChinaSoyYield1km. This study has generated a 1-km rasterized soybean yield dataset for China. To our knowledge, previous research has produced distribution maps of major soybean production areas in China, but these maps were at lower spatial resolutions. High-resolution yield datasets offer more precise spatial information on crop production, enabling rapid monitoring and analysis of large agricultural regions. This, in turn, facilitates the timely implementation of appropriate measures to enhance agricultural productivity.

This study has generated an annual soybean yield dataset for China spanning the period from 2001 to 2020. Understanding the long-term trends in soybean production over the past two decades is highly significant. Analyzing these temporal dynamics can assist agricultural decision-makers, researchers, and policymakers in comprehending the changing patterns of soybean yield. This understanding can inform the development of agricultural policies, resource allocation strategies, and management practices aimed at enhancing the efficiency and stability of soybean production. Ultimately, these efforts contribute to better meeting food demand.

Limitations of ChinaSoyYield1km. The data used in this study, including remote sensing, climate, management, and soil data, are subject to uncertainties. During data preprocessing, all data were resampled to a 1-km resolution. Additionally, not all areas within each 1 km × 1 km grid may be planted with soybeans; in some cases,

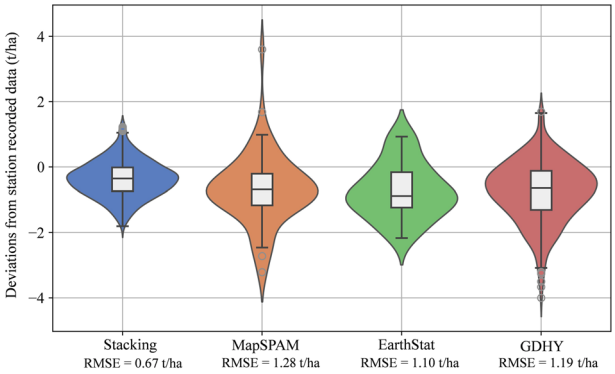

Fig. 13 The comparison between station recorded soybean yield and the results of our stacking model as well three existing rasterized yield datasets.

only a small portion of the land within the grid may be cultivated with soybeans. These issues may lead to some loss of surface information and introduce uncertainties in yield estimation. However, the purpose of yield spatialization is to inform potential data users about the expected yield level if soybeans are planted within a grid. Since the study is conducted at a 1-km resolution, the uncertainties arising from the aforementioned issues should be tolerable. Moreover, we believe that as higher-resolution planting area and environmental variable datasets become available in the future, such uncertainties will continue to decrease. Future research could benefit from using higher accuracy and resolution soybean planting area data, such as the 10-meter spatial resolution dataset[75], or carry out experiments to quantify the uncertainty caused by input data. This would enhance the precision of analyses and improve the reliability of yield estimations.

The uncertainty in statistical data is acknowledged in this study. The source and statistical methodologies for municipal-scale soybean yield used in modeling, as well as county-scale soybean yield used in model accuracy assessment, differ. While these data are collected and compiled by professional institutions, discrepancies in data collection methods, definitions, and classifications can introduce uncertainties in statistical data. It is important to recognize these potential sources of uncertainty when interpreting and applying statistical findings in agricultural research and policymaking.

The selection of predictor variables in this study lacks granularity. Variables were chosen based on four broad categories: remote sensing, climate, management, and soil, without considering finer subdivisions within each category. This approach may overlook the potential influence of specific sub-variables that could significantly impact soybean yield modeling. Sub-variables within these categories might not have been included in the model if their parent categories were filtered out during selection, thereby reducing the estimation accuracy of the models. Moreover, the precision of the models could also benefit from better predictor variables, such as vegetation indices like EVI (enhanced vegetation index) and GCVI (green chlorophyll vegetation index), or indices refined according to crop phenological stages. Future analyses could selectively screen individual sub-variables within each category, introduce new predictor variables, or attempt new spatialization approaches (e.g., state-of-the-art deep learning), to enhance model precision.

Code availability

All source scripts used for model calibration, validation, data visualization and generation are publicly available on GitHub at https://github.com/PuyuFeng/ChinaSoyYield1km.git. Please be so kind to contact the authors for more detailed information.

Received: 13 November 2024; Accepted: 28 February 2025;
Published online: 08 March 2025

References

1. Wheeler, T. & von Braun, J. Climate Change Impacts on Global Food Security. *Science* **341**, 508–513, https://doi.org/10.1126/science.1239402 (2013).
2. Islam, M. S. *et al*. Soybean and sustainable agriculture for food security. In Soybean-Recent Advances in Research and Applications. IntechOpen, https://doi.org/10.5772/intechopen.104129 (2022).
3. Thomasz, E. O., Vilker, A. S., Pérez-Franco, I. & García-García, A. Impact valuation of droughts in soybean and maize production: the case of Argentina. *Int. J. Environ. Clim.* **16**, 63–90, https://doi.org/10.1108/IJCCSM-11-2022-0139 (2023).
4. Tripathy, R. *et al*. Towards Fine-Scale Yield Prediction of Three Major Crops of India Using Data from Multiple Satellite. *J. Indian Soc. Remote Sens.* **50**, 271–284, https://doi.org/10.1007/s12524-021-01361-2 (2022).
5. Ojeda, J. J. *et al*. Implications of data aggregation method on crop model outputs – The case of irrigated potato systems in Tasmania, Australia. *Eur. J. Agron.* **126**, 126276, https://doi.org/10.1016/j.eja.2021.126276 (2021).
6. Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C. & Foley, J. A. Recent patterns of crop yield growth and stagnation. *Nat. Commun.* **3**, 1293, https://doi.org/10.1038/ncomms2296 (2012).
7. Yu, Q. *et al*. A cultivated planet in 2010 – Part 2: The global gridded agricultural-production maps. *Earth Syst. Sci. Data* **12**, 3545–3572, https://doi.org/10.5194/essd-12-3545-2020 (2020).
8. Iizumi, T. & Sakai, T. The global dataset of historical yields for major crops 1981–2016. *Sci. Data* **7**, 97, https://doi.org/10.1038/s41597-020-0433-7 (2020).

9. Feng, L., Wang, Y., Zhang, Z. & Du, Q. Geographically and temporally weighted neural network for winter wheat yield prediction. *Remote Sens. Environ.* **262**, 112514, https://doi.org/10.1016/j.rse.2021.112514 (2021).
10. Jin, Z., Azzari, G. & Lobell, D. B. Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches. *Agric. For. Meteorol.* **247**, 207–220, https://doi.org/10.1016/j.agrformet.2017.08.001 (2017).
11. Moriondo, M., Maselli, F. & Bindi, M. A simple model of regional wheat yield based on NDVI data. *Eur. J. Agron.* **26**, 266–274, https://doi.org/10.1016/j.eja.2006.10.007 (2007).
12. Wu, B. *et al*. Challenges and opportunities in remote sensing-based crop monitoring: a review. *Natl. Sci. Rev.* **10**, nwac290, https://doi.org/10.1093/nsr/nwac290 (2023).
13. Blanc, É. Statistical emulators of maize, rice, soybean and wheat yields from global gridded crop models. *Agric. For. Meteorol.* **236**, 145–161, https://doi.org/10.1016/j.agrformet.2016.12.022 (2017).
14. Kasampalis, D. A. *et al*. Contribution of Remote Sensing on Crop Models: A Review. *J. Imaging* 4, https://doi.org/10.3390/jimaging4040052 (2018).
15. Kern, A. *et al*. Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices. *Agric. For. Meteorol.* **260**, 300–320, https://doi.org/10.1016/j.agrformet.2018.06.009 (2018).
16. Ren, P. *et al*. Estimation of Soybean Yield by Combining Maturity Group Information and Unmanned Aerial Vehicle Multi-Sensor Data Using Machine Learning. *Remote Sens* 15, https://doi.org/10.3390/rs15174286 (2023).
17. Qader, S. H., Dash, J. & Atkinson, P. M. Forecasting wheat and barley crop production in arid and semi-arid regions using remotely sensed primary productivity and crop phenology: A case study in Iraq. *Sci. Total Environ.* **613**, 250–262, https://doi.org/10.1016/j.scitotenv.2017.09.057 (2018).
18. Schauberger, B., Jägermeyr, J. & Gornott, C. A systematic review of local to regional yield forecasting approaches and frequently used data resources. *Eur. J. Agron.* **120**, 126153, https://doi.org/10.1016/j.eja.2020.126153 (2020).
19. Paudel, D. *et al*. Machine learning for regional crop yield forecasting in Europe. *Field Crops Res.* **276**, 108377, https://doi.org/10.1016/j.fcr.2021.108377 (2022).
20. Chlingaryan, A., Sukkarieh, S. & Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Comput. Electron Agr.* **151**, 61–69, https://doi.org/10.1016/j.compag.2018.05.012 (2018).
21. Khaki, S. & Wang, L. Crop Yield Prediction Using Deep Neural Networks. *Front. Plant Sci*. 10, https://doi.org/10.3389/fpls.2019.00621 (2019).
22. Sun, J., Di, L., Sun, Z., Shen, Y. & Lai, Z. County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model. *Sensors* 19, https://doi.org/10.3390/s19204363 (2019).
23. Herrero-Huerta, M., Rodriguez-Gonzalvez, P. & Rainey, K. M. Yield prediction by machine learning from UAS-based multi-sensor data fusion in soybean. *Plant Methods* **16**, 78, https://doi.org/10.1186/s13007-020-00620-6 (2020).
24. Teodoro, P. E. *et al*. Predicting Days to Maturity, Plant Height, and Grain Yield in Soybean: A Machine and Deep Learning Approach Using Multispectral Data. *Remote Sens*. 13, https://doi.org/10.3390/rs13224632 (2021).
25. Kuradusenge, M. *et al*. Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *Agriculture* 13 https://doi.org/10.3390/agriculture13010225 (2023).
26. Chen, S. *et al*. Improving Spatial Disaggregation of Crop Yield by Incorporating Machine Learning with Multisource Data: A Case Study of Chinese Maize Yield. *Remote Sens*. 14, https://doi.org/10.3390/rs14102340 (2022).
27. Torsoni, G. B. *et al*. Soybean yield prediction by machine learning and climate. *Theor. Appl. Climatol.* **151**, 1727–1727, https://doi.org/10.1007/s00704-023-04389-1 (2023).
28. Song, X.-P., Li, H., Potapov, P. & Hansen, M. C. Annual 30 m soybean yield mapping in Brazil using long-term satellite observations, climate data and machine learning. *Agric. For. Meteorol.* **326**, 109186, https://doi.org/10.1016/j.agrformet.2022.109186 (2022).
29. Zou, Y., Kattel, G. R. & Miao, L. Enhancing Maize Yield Simulations in Regional China Using Machine Learning and Multi-Data Resources. *Remote Sens*. 16, https://doi.org/10.3390/rs16040701 (2024).
30. Attri, I., Awasthi, L. K. & Sharma, T. P. Machine learning in agriculture: a review of crop management applications. *Multimed. Tools Appl.* **83**, 12875–12915, https://doi.org/10.1007/s11042-023-16105-2 (2024).
31. Tao, S. *et al*. Retrieving soil moisture from grape growing areas using multi-feature and stacking-based ensemble learning modeling. *Comput. Electron. Agr.* **204**, 107537, https://doi.org/10.1016/j.compag.2022.107537 (2023).
32. Shahhosseini, M., Hu, G. & Archontoulis, S. V. Forecasting corn yield with machine learning ensembles. *Front. Plant Sci.* **11**, 1120, https://doi.org/10.3389/fpls.2020.01120 (2020).
33. Dietterich, T.G. Ensemble Methods in Machine Learning. In: *Multiple Classifier Systems*, MCS 2000, Lecture Notes in Computer Science, vol 1857, Springer, Berlin, Heidelberg, https://doi.org/10.1007/3-540-45014-9_1 (2000).
34. Pham, H. & Olafsson, S. Bagged ensembles with tunable parameters. *Comput. Intell.* **35**, 184–203, https://doi.org/10.1111/coin.12198 (2019).
35. Shahhosseini, M., Hu, G. & Pham, H. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications* **7**, 100251, https://doi.org/10.1016/j.mlwa.2022.100251 (2022).
36. Song, W. *et al*. Geographic distributions and the regionalization of soybean seed compositions across China. *Int. Food Res.* **164**, 112364, https://doi.org/10.1016/j.foodres.2022.112364 (2023).
37. Adalibieke, W., Cui, X., Cai, H., You, L., Zhou, F. Global crop-specific nitrogen fertilization dataset in 1961-2020. National Tibetan Plateau / Third Pole Environment Data Center, org/ https://doi.org/10.11888/Terre.tpdc.300446 (2023).
38. Adalibieke, W., Cui, X., Cai, H., You, L. & Zhou, F. Global crop-specific nitrogen fertilization dataset in 1961–2020. *Sci. Data* **10**(1), 617, https://doi.org/10.1038/s41597-023-02526-z (2023).
39. Ning, X., Dong, P., Wu, C., Wang, Y. & Zhang, Y. Influence Mechanisms of Dynamic Changes in Temperature, Precipitation, Sunshine Duration and Active Accumulated Temperature on Soybean Resources: A Case Study of Hulunbuir, China, from 1951 to 2019. *Energies* 15, https://doi.org/10.3390/en15228347 (2022).
40. Elli, E. F. *et al*. Climate Change and Management Impacts on Soybean N Fixation, Soil N Mineralization, N_2O Emissions, and Seed Yield. *Front. Plant Sci.* 13, https://doi.org/10.3389/fpls.2022.849896 (2022).
41. Xu, X. *et al*. Unleashing the power of machine learning and remote sensing for robust seasonal drought monitoring: A stacking ensemble approach. *J. Hydrol.* **634**, 131102, https://doi.org/10.1016/j.jhydrol.2024.131102 (2024).
42. Wells, N., Goddard, S. & Hayes, M. J. A Self-Calibrating Palmer Drought Severity Index. *J. Clim.* **17**, 2335–2351 10.1175/1520-0442(2004)017<2335:ASPDSI>2.0.CO;2 (2004).
43. Inoue, T. *et al*. Minimizing VPD Fluctuations Maintains Higher Stomatal Conductance and Photosynthesis, Resulting in Improvement of Plant Growth in Lettuce. *Front. Plant Sci*. 12, https://doi.org/10.3389/fpls.2021.646144 (2021).
44. Sun, W. *et al*. Projected long-term climate trends reveal the critical role of vapor pressure deficit for soybean yields in the US Midwest. *Sci. Total Environ.* **878**, 162960, https://doi.org/10.1016/j.scitotenv.2023.162960 (2023).
45. Sorooshian, S., Hsu, K., Braithwaite, D., Ashouri, H. & NOAA CDR Program. NOAA Climate Data Record (CDR) of Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN-CDR), Version 1 Revision 1. [For scientific research]. NOAA National Centers for Environmental Information. https://doi.org/10.7289/V51V5BWQ [15-Feb-2024] (2014).
46. Ashouri H. *et al*. PERSIANN-CDR: Daily Precipitation Climate Data Record from Multi-Satellite Observations for Hydrological and Climate Studies. *Bull. Amer. Meteor. Soc*., https://doi.org/10.1175/BAMS-D-13-00068.1 (2015).

47. Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A. & Hegewisch, K. C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. *Sci. Data* **5**, 170191, https://doi.org/10.1038/sdata.2017.191 (2018).
48. Wu, X. *et al*. Spatial-temporal dynamics of maize and soybean planted area, harvested area, gross primary production, and grain production in the Contiguous United States during 2008-2018. *Agric. For. Meteorol.* **297**, 108240, https://doi.org/10.1016/j.agrformet.2020.108240 (2021).
49. Marshall, M., Tu, K. & Brown, J. Optimizing a remote sensing production efficiency model for macro-scale GPP and yield estimation in agroecosystems. *Remote Sens. Environ.* **217**, 258–271, https://doi.org/10.1016/j.rse.2018.08.001 (2018).
50. Xin, Q. *et al*. A Production Efficiency Model-Based Method for Satellite Estimates of Corn and Soybean Yields in the Midwestern US. *Remote Sens.* **5**, 5926–5943, https://doi.org/10.3390/rs5115926 (2013).
51. Li, X. & Xiao, J. A Global, 0.05-Degree Product of Solar-Induced Chlorophyll Fluorescence Derived from OCO-2, MODIS, and Reanalysis Data. *Remote Sens.* 11, https://doi.org/10.3390/rs11050517 (2019).
52. Guo, M., Li, J., Li, J., Zhong, C. & Zhou, F. Solar-Induced Chlorophyll Fluorescence Trends and Mechanisms in Different Ecosystems in Northeastern China. *Remote Sens*. 14, https://doi.org/10.3390/rs14061329 (2022).
53. Qiu, R. *et al*. Monitoring drought impacts on crop productivity of the U.S. Midwest with solar-induced fluorescence: GOSIF outperforms GOME-2 SIF and MODIS NDVI, EVI, and NIRv. *Agric. For. Meteorol.* **323**, 109038, https://doi.org/10.1016/j.agrformet.2022.109038 (2022).
54. He, M. *et al*. Regional Crop Gross Primary Productivity and Yield Estimation Using Fused Landsat-MODIS Data. *Remote Sens*. 10, https://doi.org/10.3390/rs10030372 (2018).
55. Shammi, S. A. & Meng, Q. Use time series NDVI and EVI to develop dynamic crop growth metrics for yield modeling. *Ecol. Indic.* **121**, 107124, https://doi.org/10.1016/j.ecolind.2020.107124 (2021).
56. de la Casa, A. *et al*. Soybean crop coverage estimation from NDVI images with different spatial resolution to evaluate yield variability in a plot. *ISPRS J. Photogramm. Remote Sens.* **146**, 531–547, https://doi.org/10.1016/j.isprsjprs.2018.10.018 (2018).
57. Wardlow, B. D. & Egbert, S. L. Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains. *Remote Sens. Environ.* **112**, 1096–1116, https://doi.org/10.1016/j.rse.2007.07.019 (2008).
58. Running, S. & Zhao, M. MOD17A3HGF MODIS/Terra Net Primary Production Gap-Filled Yearly L4 Global 500 m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. Accessed 2024-02-08 from https://doi.org/10.5067/MODIS/MOD17A3HGF.006 (2019).
59. Didan, K. MOD13A3 MODIS/Terra vegetation Indices Monthly L3 Global 1km SIN Grid V006, distributed by NASA EOSDIS Land Processes Distributed Active Archive Center, Accessed 2024-02-08 from https://doi.org/10.5067/MODIS/MOD13A3.006 (2015).
60. Wan, Z., Hook, S., Hulley, G. MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. Accessed 2024-02-08 from https://doi.org/10.5067/MODIS/MOD11A2.006 (2015).
61. Zheng, J., Huang, X., Sangondimath, S., Wang, J. & Zhang, Z. Efficient and Flexible Aggregation and Distribution of MODIS Atmospheric Products Based on Climate Analytics as a Service Framework. *Remote Sens*. 13, https://doi.org/10.3390/rs13173541 (2021).
62. Coulibali, Z., Cambouris, A. N. & Parent, S.-É. Site-specific machine learning predictive fertilization models for potato crops in Eastern Canada. *PLoS one* **15**, e0230888, https://doi.org/10.1371/journal.pone.0230888 (2020).
63. Assefa, Y. *et al*. Assessing Variation in US Soybean Seed Composition (Protein and Oil). *Front. Plant Sci*. 10, https://doi.org/10.3389/fpls.2019.00298 (2019).
64. Xu, Y. *et al*. Bacterial communities in soybean rhizosphere in response to soil type, soybean genotype, and their growth stage. *Soil Biol. Biochem*. **41**, 919–925, https://doi.org/10.1016/j.soilbio.2008.10.027 (2009).
65. Anthony, P., Malzer, G., Sparrow, S. & Zhang, M. Soybean Yield and Quality in Relation to Soil Properties. *Agron. J.* **104**, 1443–1458, https://doi.org/10.2134/agronj2012.0095 (2012).
66. Ferreira, C. J. B. *et al*. Effectiveness of narrow tyne and double-discs openers to overcome shallow compaction and improve soybean yield in long-term no-tillage soil. *Soil Till. Res.* **227**, 105622, https://doi.org/10.1016/j.still.2022.105622 (2023).
67. Nachtergaele F. *et al*. Harmonized world soil database version 2.0. *FAO*, https://doi.org/10.4060/cc3823en (2023).
68. Hancock, J. T. & Khoshgoftaar, T. M. CatBoost for big data: an interdisciplinary review. *J. Big Data* **7**(1), 94, https://doi.org/10.1186/s40537-020-00369-8 (2020).
69. Mastelini, S. M., Nakano, F. K., Vens, C. & de Leon Ferreira, A. C. P. Online extra trees regressor. *IEEE Trans. Neural Netw. Learn. Syst.* **34**(10), 6755–6767, https://doi.org/10.1109/TNNLS.2022.3212859 (2022).
70. Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J. & Gifford, E. M. Extreme gradient boosting as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.* **56**(12), 2353–2360, https://doi.org/10.1021/acs.jcim.6b00591 (2016).
71. Fan, J. *et al*. Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agric. Water Manag.* **225**, 105758, https://doi.org/10.1016/j.agwat.2019.105758 (2019).
72. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32, https://doi.org/10.1023/A:1010933404324 (2001).
73. Wolpert, D. H. Stacked generalization. *Neural Networks* **5**, 241–259, https://doi.org/10.1016/S0893-6080(05)80023-1 (1992).
74. Zhang, M. *et al*. ChinaSoybeanYield1km: a 1-km annual soybean yield dataset from 2001 to 2020 in China. V3. Science Data Bank https://doi.org/10.57760/sciencedb.18390 (2024).
75. Mei, Q. *et al*. ChinaSoyArea10m: a dataset of soybean planting areas with a spatial resolution of 10 m across China from 2017 to 2021. *Earth Syst. Sci. Data Discuss*., 1-27, https://doi.org/10.5194/essd-2023-467 (2023).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant No. 42301112), the National Key Research & Development Program of China (Grant No. 2023YFD1701804), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA28060200), and the Pinduoduo-China Agricultural University Research Fund (Grant No. PC2024B01007).

Author contributions

Min Zhang, Xinlei Xu, and Puyu Feng contributed to the design of this research. Min Zhang and Xinlei Xu collectively prepared the manuscript with contributions from all coauthors. Junji Ou, Zengguang Zhang, Fangzheng Chen, Lijie Shi, Meiqin Zhang, Bin Wang, Liang He, Xueliang Zhang, Yong Chen, Kelin Hu and Puyu Feng revised the manuscript. Xinlei Xu and Puyu Feng developed the model code.

Competing interests

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.

Additional information

Correspondence and requests for materials should be addressed to P.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.