

Article

Predicting China's Maize Yield Using Multi-Source Datasets and Machine Learning Algorithms

Lijuan Miao ¹, Yangfeng Zou ^{1,*}, Xuefeng Cui ², Giri Raj Kattel ^{1,3,4}, Yi Shang ⁵ and Jingwen Zhu ⁶

¹ School of Geographical Sciences, Nanjing University of Information Science & Technology, Nanjing 210044, China; 002750@nuist.edu.cn (L.M.); gkattel@unimelb.edu.au (G.R.K.)

² School of Systems Science, Beijing Normal University, Beijing 100875, China; xuefeng.cui@bnu.edu.cn

³ Department of Infrastructure Engineering, University of Melbourne, Melbourne, VIC 3010, Australia

⁴ Department of Hydraulic Engineering, Tsinghua University, Beijing 100084, China

⁵ School of Atmospheric Physics, Nanjing University of Information Science & Technology, Nanjing 210044, China; 20211103006@nuist.edu.cn

⁶ Changwang School of Honors, Nanjing University of Information Science & Technology, Nanjing 210044, China; 202183330028@nuist.edu.cn

* Correspondence: 20211210020@nuist.edu.cn

Abstract: A timely and accurately predicted grain yield can ensure regional and global food security. The scientific community is gradually advancing the prediction of regional-scale maize yield. However, the combination of various datasets while predicting the regional-scale maize yield using simple and accurate methods is still relatively rare. Here, we have used multi-source datasets (climate dataset, satellite dataset, and soil dataset), lasso algorithm, and machine learning methods (random forest, support vector, extreme gradient boosting, BP neural network, long short-term memory network, and K-nearest neighbor regression) to predict China's county-level maize yield. The use of multi-sourced datasets advanced the predicting accuracy of maize yield significantly compared to the single-sourced dataset. We found that the machine learning methods were superior to the lasso algorithm, while random forest, extreme gradient boosting, and support vector machine represented the most preferable methods for maize yield prediction in China ($R^2 \geq 0.75$, RMSE = 824–875 kg/ha, MAE = 626–651 kg/ha). The climate dataset contributed more to the prediction of maize yield, while the satellite dataset contributed to tracking the maize growth process. However, the methods' accuracies and the dominant variables affecting maize growth varied with agricultural regions across different geographic locations. Our research serves as an important effort to examine the feasibility of multi-source datasets and machine learning techniques for regional-scale maize yield prediction. In addition, the methodology we have proposed here provides guidance for reliable yield prediction of different crops.

Keywords: China; maize yield; machine learning; multi-source datasets; prediction



Citation: Miao, L.; Zou, Y.; Cui, X.; Kattel, G.R.; Shang, Y.; Zhu, J. Predicting China's Maize Yield Using Multi-Source Datasets and Machine Learning Algorithms. *Remote Sens.* **2024**, *16*, 2417. <https://doi.org/10.3390/rs16132417>

Academic Editor: João Catalão Fernandes

Received: 21 May 2024

Revised: 24 June 2024

Accepted: 27 June 2024

Published: 1 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Among major crops planted around the world, maize is an important staple food for more than 4.5 billion members of the global population [1,2]. Thus, maize plays a critical role in global food security. Currently, the population is growing at a rate of approximately 1.1% per year, with projections indicating a potential population of 9.7 billion by 2050 [3]. Such a situation will have significant implications for the relationship between grain supply and demand. Based on the food consumption in 2010, 70% more food production is needed in 2050 [4,5]. As climate change intensifies, the demand for increased food production becomes more critical [6]. A decline in maize yield poses a substantial threat in the form of global food shortages [7,8], directly impacting the achievement of the UN sustainable development goals, especially the goal of “Zero hunger” by 2030 [9]. Hence, precise crop yield prediction is urgently needed to overcome the threat.

Recently, several mathematical models, including crop models, traditional regression, and machine learning (ML) methods, have been developed for the prediction of maize yield worldwide [10–12]. Although crop process models can accurately predict yields, their reliance on fine-grained soil and climate datasets poses challenges when applying them to regional-scale yield predictions [13,14]. Statistical regression models typically predict crop yields by exploring relationships between climate variables (e.g., temperature, precipitation, solar radiation, etc.) and actual crop yields [15]. However, due to the complex nonlinear relationships among variables, regression results often exhibit low and controversial explanatory power. Moreover, the primary factors influencing crop growth can vary across different geographical locations, seasons, and crop varieties, making it challenging to extrapolate these models to larger regions [16,17]. In contrast, however, the ML methods can overcome the limitations posed by complex nonlinear relationships [18,19] inherent in many traditional regression models [20]. For example, recent studies have successfully utilized models based on high-resolution images to predict crop yield within fields at a resolution of 3 m, highlighting the enhanced practicality of machine learning methods [21,22].

Climate variables alone are used as the primary inputs for many crop yield prediction models [23–25]. However, only a few studies have considered the combination of multi-source datasets for simulating crop yield comprehensively [26,27]. For instance, plant growth is co-regulated by both abiotic factors and biotic factors [28,29], in which climate variables could be restricted to describing abiotic factors that limits maize growth, rather than detecting biotic factors that reflect maize growth conditions [30,31]. Hence, maize yield predictions solely based on climate variables are insufficient.

The combination of various spectral bands, multiple satellite datasets (e.g., the normalized difference vegetation index (NDVI), the enhanced vegetation index (EVI), the sun/solar-induced chlorophyll fluorescence (SIF), and the leaf area index (LAI)) are recommended for maize yield predictions [32,33]. Other determinants, such as soil information, are also suggested as important indicators for crop yield predictions [34]. Previous studies have successfully utilized various combinations of datasets throughout the entire maize growing season to accurately predict maize yield [35,36]. However, the optimal period for predicting maize yield using different datasets has not yet been determined. Additionally, only a few studies have considered the variations in the relative contributions of the combination of climate and satellite datasets to maize yield prediction across different maize growth stages [20]. This highlights the need for further enhancements in maize yield prediction models by combining the maize growth conditions.

China is the world's second-largest maize-producing country, accounting for 21% of the global production with less than 9% of the global maize planting areas [37]. However, limited studies on the combination of different datasets and the use of ML algorithms in China's maize planting regions have reduced the precision of county-level maize yield prediction. A timely and accurately predicted maize yield in China is therefore vital for ensuring both regional and global food security. Here, we have integrated satellite datasets, climate datasets, and soil datasets, to construct a reliable prediction model for maize yield in China. Seven algorithms were analyzed, including the random forest (RF), support vector machine (SVM), extreme gradient boosting (XGB), BP neural network (BPNN), long short-term memory network (LSTM), K-nearest neighbor (KNN), and least absolute shrinkage and selection operator (LASSO) algorithm, respectively. Overall, the main objectives of this research were as follows: (1) to develop and select optimal models for predicting county-level maize yield in China; (2) to determine the contributions of satellite datasets and climate datasets at different growth stages of maize to the prediction of maize yield; (3) to evaluate the performance of selected maize yield prediction models in China's four maize regions and identify the dominant factors influencing maize yield prediction.

2. Materials and Methods

2.1. Study Area

Based on the cultivation characteristics, management practices, and geographical environments, the planting regions of maize in China are divided into four zones: the north maize region (Zone I), the Huang-Huai-Hai summer maize region (Zone II), the southwest maize region (Zone III), and the northwest maize region (Zone IV) [38] (Figure 1). These four zones are predominantly influenced, respectively, by the continental monsoon climate, semi-humid monsoon climate, subtropical monsoon humid climate, and monsoon climate of medium latitudes. The annual mean temperature in all four zones is around 9–25 °C, which is the optimum temperature range for maize growth (9–29 °C) [39].



Figure 1. Maize area and four maize planting zones in China (Zone I: the north maize region; Zone II: the Huang-Huai-Hai maize region; Zone III: the southwest maize region; Zone IV: the northwest maize region).

2.2. Data Sources

We collected multi-source datasets with various spatial and temporal resolutions, including annual maize planting areas, county-level yield, satellite dataset, and climate dataset. An overview of these datasets is included in the Supplementary Materials (Table S1).

2.2.1. Satellite Dataset

We used the enhanced vegetation index (EVI) extracted from the MODIS Terra monthly product from 2000 to 2010, featuring a spatial resolution of 1 km (MOD13A2V061, available from <https://doi.org/10.5067/MODIS/MOD13A2.061>, accessed on 13 January 2022). The leaf area index (LAI) was extracted from the MODIS Terra monthly product from 2000 to 2010, featuring a spatial resolution of 500 m (MOD15A2V061, available from <https://doi.org/10.5067/MODIS/MOD15A2H.061>, accessed on 13 January 2022). These products are calculated based on atmospheric BRDF (bidirectional reflectance distribution function) correction for bidirectional surface reflectance. The reflectance values have been masked for the effects of water, clouds, heavy aerosols, and cloud shadows through QA quality control bands. Specifically, the correction algorithm considers the interactions between atmospheric components, solar radiation, and surface reflectance. This helps eliminate atmospheric absorption and scattering effects while considering the BRDF of the Earth's surface, ensuring that the final reflectance data more accurately reflect the true characteristics of the Earth's surface. EVI has demonstrated a strong correlation with crop yield, as evidenced by previous studies [40–42], and LAI was an important indicator of plant productivity and photosynthetic capacity [43,44].

We used a global, long-term, spatially continuous sun/solar-induced chlorophyll fluorescence (SIF) product from CSIF, covering the period from 2000 to 2010, with a spatial resolution of 0.05 degree (<https://doi.org/10.6084/m9.figshare.6387494>, accessed on 13 January 2022). The CSIF product was generated through discrete Orbiting Carbon Observatory-2 (OCO-2), SIF soundings, and meteorological reanalysis data based on a data-driven method. Specifically, these data were generated by training a neural network (NN) using surface reflectance from the MODIS and SIF from the OCO-2. SIF has been developed and increasingly applied to crop yield predictions in recent years, as it can directly reflect the dynamic changes in actual plant photosynthesis and responds to crop stress [45,46].

2.2.2. Climate and Soil Dataset

The monthly climate dataset was derived from TerraClimate datasets (<http://doi.org/10.7923/G43J3B0R>, accessed on 13 January 2022) from 2000–2010, and it was commonly used for regional maize yield prediction with a high spatial resolution ($1/24^\circ$, ~4 km) [47,48]. This dataset was created by using climatically aided interpolation, combining high-spatial-resolution data from the WorldClim version 1.4 and version 2 datasets with coarser-resolution time-varying (i.e., monthly) data from CRU Ts4.0 and JRA-55 to produce a monthly dataset. Here, we selected multiple features including monthly maximum temperature (TMAX), monthly minimum temperature (TMIN), monthly total precipitation (PRE), monthly actual evapotranspiration (AET), monthly potential evapotranspiration (PET), monthly downward surface shortwave radiation (SRAD), monthly vapor pressure deficit (VPD), and monthly Palmer drought severity index (PDSI).

The soil moisture (SM) dataset was also derived from TerraClimate datasets ($1/24^\circ$, ~4 km). In addition, seven features describing the physical and the chemical properties of the soil from the Harmonized World Soil Database (HSWD) were selected [49]. These features are subsoil gravel content (GRAVEL), sand fraction (SAND), subsoil silt fraction (SILT), subsoil clay fraction (CLAY), subsoil reference bulk density (BULK), subsoil organic carbon (OC), and subsoil pH (pH). This dataset spans from 2000 to 2010, and its spatial resolution is 1 km.

2.2.3. Maize Yield

The county-level maize yield dataset (kg/ha) from 2000–2010 was obtained from the Agricultural Statistical Yearbook, compiled by the Ministry of Agriculture of China (<http://www.stats.gov.cn>, accessed on 19 March 2022). The maize planting area data are from a 1 km grid crop phenological dataset for three main crops from 2000–2010, with errors of the retrieved phenological date being less than 10 d [40]. Overall, we selected a total of 1021 counties for further analysis, choosing those counties that had maize yield data for over 8 years after removing outliers and that had more than 10 planting grids.

2.2.4. Data Preprocessing

First, we cleaned the data by removing counties where no yield was reported. To address missing data, we directly removed rows containing missing values. Given our focus on county-level maize yield in China, we resampled the climate, soil, leaf area index (LAI), and solar-induced fluorescence (SIF) datasets to a 1 km resolution using nearest neighbor interpolation. This ensured spatial alignment across all datasets.

Next, we used the maize planting area data to mask the climate, satellite, and soil datasets. This step eliminated the influence of other crops and vegetation types, further ensuring spatial consistency. For the satellite dataset, we employed maximum value compositing to synthesize the enhanced vegetation index (EVI) and normalized difference vegetation index (NDVI) monthly. This allowed us to temporally align the satellite dataset with the climate dataset. Considering the typical growing season of maize in China, which is generally sown in April and harvested in October [40,50], we selected satellite and climate data for the six-month period from April to September. The soil dataset, which only has one value per year, was treated as a static variable in our analysis. On the other hand, the satellite and climate datasets, which vary over time, were treated as dynamic variables.

Finally, we aggregated all datasets to the county level by calculating mean values and matched them with county-level maize yield. This comprehensive preprocessing approach ensured that our analysis was based on accurate and spatially/temporally aligned data. A total of 79 variables were considered across all datasets, as detailed in Table S2.

2.3. Methods for Predicting Maize Yield

This study primarily utilized six machine learning methods and a linear regression method for model construction, and the parameters of the machine learning models are shown in Table 1.

Table 1. Machine learning model parameters.

Models	Parameters
RF	n_estimators, max_depth, max_features, min_samples_leaf, min_samples_split
SVM	C, kernel, gamma, epsilon
XGB	n_estimators, max_depth, eta, gamma, min_child_weight, subsample
BPNN	learning rate, epochs, batch size, activation function, regularization parameters, optimizer, hidden layers
LSTM	learning rate, epochs, batch size, activation function, regularization parameters, optimizer, hidden layers
KNN	n_neighbors, weights, algorithm

2.3.1. Random Forest

Random forest (RF) is an ensemble learning algorithm that performs regressions or classifications by combining a large number of decision trees [51]. The algorithm introduces additional randomness when growing trees and searches for the best trees in a subset of random features. This situation leads to greater tree diversity and usually produces an overall better model [52]. RF is a reasonable method for variable selection, as it could quantify the relative importance of measured variables [53] and efficiently handle high-dimensional datasets [54]. Here, we employed the “GridSearchCV” to tune five hyperparameters: “n_estimators = 355”, “max_depth = 20”, “max_features = 20”, “min_samples_leaf = 1”, and “min_samples_split = 2”.

2.3.2. Support Vector Machine

Support vector machine (SVM) is a supervised non-parametric algorithm, and it is based on the usage of kernels and action on the margins [55]. The input is mapped to a high-dimensional feature space using a kernel function, and then a linear regression model is constructed in the new feature space to balance between minimizing errors and overfitting [56]. In this study, we sequentially tuned the following four hyperparameters as follows: “GridSearchCV”: “C = 0.1”, “kernel = rbf”, “gamma = 0.01”, “epsilon = 0.1”.

2.3.3. Extreme Gradient Boosting

Extreme gradient boosting (XGB) is an optimized distributed gradient-boosting library designed for efficiency, flexibility, and portability [57]. The algorithm fits the first learner to the entire input and fits the second learner to the residuals of the first learner. The algorithm also simplifies the objective function by allowing regularization terms to prevent overfitting and is very useful for large data and high-dimensional data. By comparing different kernels, we can find the optimized kernel that reflects whether the relationship between the climate dataset and maize yield are linear or non-linear. Here, we tuned the six hyperparameters as follows: “n_estimators = 261”, “max_depth = 10”, “eta = 0.05”, “gamma = 0”, “min_child_weight = 1”, “subsample = 0.8”.

2.3.4. Back Propagation Neural Network

Neural networks consist of different elements that are highly interconnected. A back propagation neural network (BPNN) is a multi-layer feed-forward network trained by an error back-propagation algorithm; it belongs to one of the most widely used artificial neural networks [58]. Its learning rule is to use the gradient descent method to continuously adjust the weights and thresholds of the network by back-propagation to minimize the sum of squared errors of the network [59]. The topology of the BPNN model includes input layers, hidden layers, and output layers. A BPNN usually has the ability to detect complex relationships between variables [60] and is suitable to the complex nonlinear problem between maize yield and variables in this study. In this study, the climate dataset, satellite dataset, and soil dataset were processed into three hidden layers, each consisting of 150 neurons and employing a “ReLU” activation function. RMSprop was used to select the optimal learning rate, and L2 regularization was applied to adjust the model’s complexity and fitting capacity.

2.3.5. Long Short-Term Memory Neural Network

A long short-term memory neural network (LSTM) is a special kind of recurrent neural network (RNN) [61]. LSTM was used to deal with the inherent problem of RNN: with the prolongation of training time and the increase of network layers, the original RNN is prone to gradient explosion or gradient disappearance during training, resulting in the inability to process longer sequences of data and thus to obtain information about long-range data [62]. In this study, the optimal parameter search was conducted with respect to the activation function, hidden size, learning rate, and batch size. As a result, the activation function, hidden size, epochs, learning rate, and batch size were determined as hard sigmoid function, 150, 0.01, and 84, respectively.

2.3.6. K-Nearest Neighbor Regression

The K-nearest neighbor (KNN) is a theoretically mature method and one of the simplest machine learning algorithms [63]. The idea of the method is very simple and intuitive: if most of the K most similar (i.e., most neighboring) samples in the feature space belong to a certain category, then the sample also belongs to that category. KNN can tolerate noise and unrelated properties, but it has higher complexity and is also more time-consuming compared to other methods. For KNN in this study, key parameters to tune included “n_neighbors = 10”, “weights = distance”, and “algorithm = auto”.

2.3.7. Least Absolute Shrinkage and Selection Operator

The least absolute shrinkage and selection operator (LASSO) was first proposed by Robert Tibshirani in 1996 [64]. It is a regularization technique based on linear models, used to address regression problems with multiple features. It compresses some coefficients by constructing a penalty function, and sets some coefficients to zero to get a more refined model [65].

2.4. Experiment Design

To reduce the complexity of the model and enhance its stability, we conducted initial feature selection using Pearson correlation analysis. Then, we removed features within the same category that exhibited high correlation and had low or insignificant correlation with maize yield. Subsequently, through multiple comparative experiments, we assessed the model's performance before and after removal to ensure that eliminating features did not adversely impact model performance. We conducted three groups of comparative experiments to achieve the objectives mentioned in the introduction: (1) To investigate model performance based on different datasets, all variables were categorized into three groups and input into seven models: the climate group (climate dataset and soil dataset), the satellite group (satellite dataset and soil dataset), and the combined group (satellite dataset, climate dataset, and soil dataset). (2) To determine the model performance for maize yield prediction for each period, variables were categorized into three stages: early—April to May (from planting to V3 (maize vegetative stages of third leaf, the first critical growth stage of maize); peak—June to July (from V3 to silking), and late—August to September (from silking to maturity), respectively. (3) To evaluate regional differences in maize yield prediction, we applied prediction models to each region, determining the order of feature importance and assessing model performance. Then, we input the three dataset groups from April to September into the model sequentially to evaluate how different categories of datasets contribute to variations in maize yield prediction across different growth stages. The flowchart of this study is shown in Figure 2.

Here, six ML methods (RF, SVM, XGB, BPNN, LSTM, KNN) and one linear regression method (LASSO) were applied to predict the maize yield, as detailed in Section 2.3.1. To eliminate the scale and unit effects of different variables and enhance the efficiency and accuracy of data analysis, the datasets were standardized using the min-max normalization method before fitting the model:

$$X_{nomal} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X_{max} is the maximum value of the dataset, and X_{min} is the minimum value of the dataset.

All models were implemented using the “sklearn” package in Python 3.7. We randomly split the dataset from 2000 to 2009 into training and test sets with a 7:3 ratio using the “train_test_split”. The dataset from 2010 was reserved for prediction purposes. Cross-validation (CV) is a widely used strategy for algorithm selections [66]. We used ten-fold cross-validation to test the performance of the above-mentioned methods.

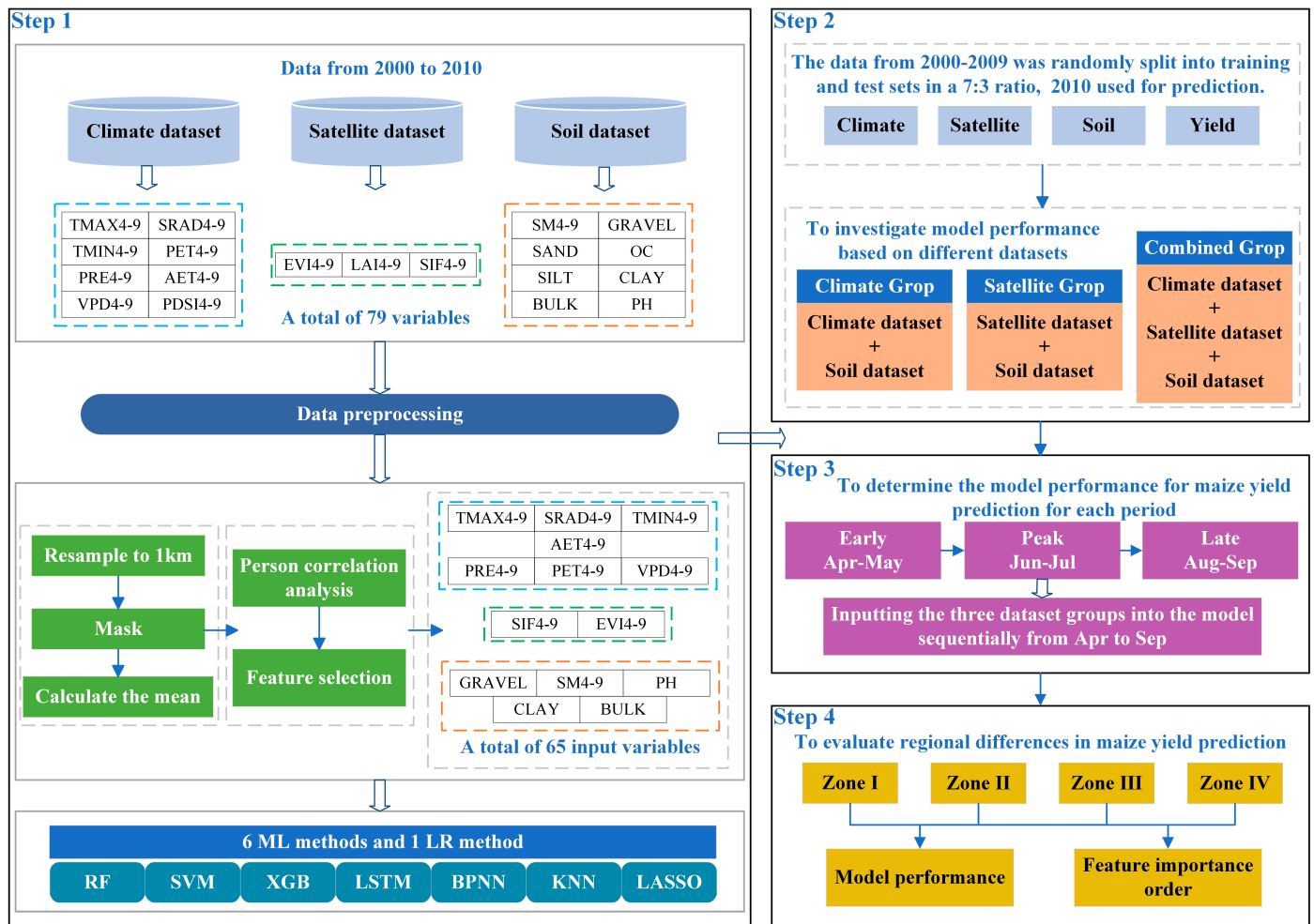


Figure 2. Flowchart of this study.

2.5. Model Results Evaluations

The coefficient of determination (R^2), the root-mean-square error (RMSE), and the mean absolute error (MAE) were used to evaluate model performance. They can be calculated as follows:

$$R^2 = \frac{\left(\sum_{i=1}^n (y_i - \bar{y}_i) (f_i - \bar{f}_i) \right)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2 \sum_{i=1}^n (f_i - \bar{f}_i)^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - f_i)^2} \quad (3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - f_i| \quad (4)$$

where n ($i = 1, 2, \dots, n$) is the number of samples input in the machine learning models, y_i is the recorded maize yield, $\bar{y}_i$ is the corresponding mean value, f_i is the predicted maize yield, and $\bar{f}_i$ is the corresponding mean value. The closer R^2 is to 1, the higher the prediction performance of the model is. Small RMSE and MAE values indicate less discrepancy between the recorded yield and predicted yield.

We used percentage error to measure the accuracy error of the model in different agricultural regions. The percentage error is calculated by (predicted yield—recorded yield)/recorded yield $\times$ 100.

3. Results

3.1. Selections of the Key Features for Maize Yield Predictions

As shown in Figure 3, EVI ($R = 0.23$, $p < 0.001$) and SIF ($R = 0.16$, $p < 0.001$) were found to be highly correlated with maize yield. There were no significant correlations between PDSI and maize yield. All water-related features (SM, PRE, VPD, PET, and AET) showed higher correlations with maize yield than those of temperature-related features (TMIN and TMAX). Moreover, soil-property-related features (CLAY, GRAVEL, PH, BULK, and SILT) showed slightly stronger correlations with maize yield than that of SAND. Eventually, 15 features significantly correlated with maize yield ($p < 0.001$) were selected for further model simulations. They were, separately, EVI, SM, PRE, VPD, PH, SIF, GRAVEL, TMIN, AET, PET, CLAY, BULK, SILT, TMAX, and SRAD, by the sequence of correlation coefficient from high to low. There were six months of data for each feature in the satellite dataset and climate dataset, resulting in a total of 65 input variables (Table 2).

Pearson correlation

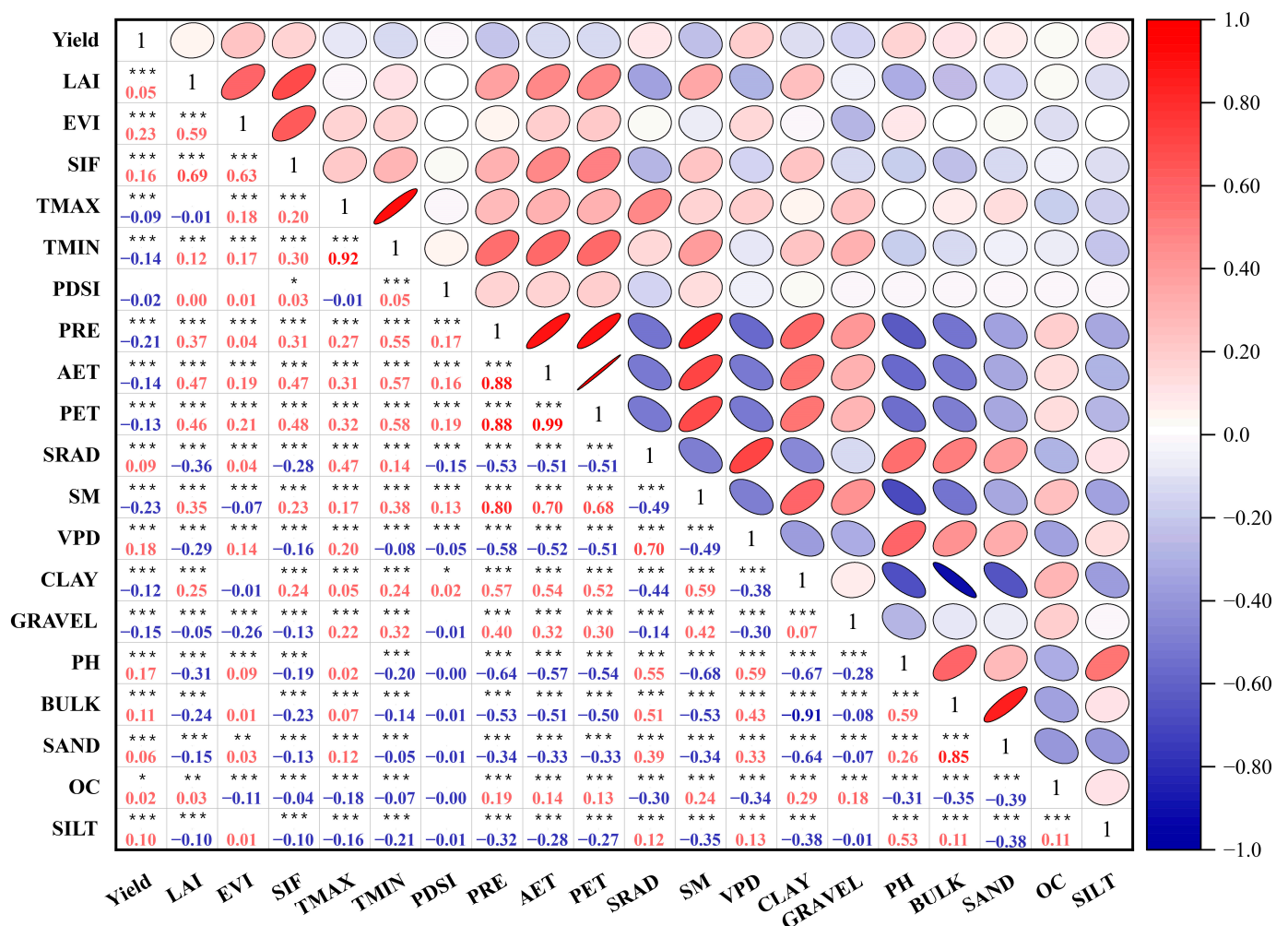


Figure 3. Pearson correlations between the observed maize yield and its related features at monthly scale. LAI: leaf area index; EVI: enhanced vegetation index; SIF: sun/solar-induced chlorophyll fluorescence; TMAX: maximum temperature; TMIN: minimum temperature; PRE: total precipitation; AET: actual evapotranspiration; PET: potential evapotranspiration; SRAD: downward surface shortwave radiation; VPD: vapor pressure deficit; PDSI: monthly Palmer drought severity index; GRAVEL: subsoil gravel content; SAND: sand fraction; SILT: the subsoil silt fraction; CLAY: subsoil clay fraction; BULK: subsoil reference bulk density; OC: the subsoil organic carbon; PH: subsoil pH; SM: soil moisture. *, **, and *** represent significance levels at $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

Table 2. Input variables for models.

Variables	Variable Descriptions
Climate variables (total of 42)	
TMAX4-TMAX9	Monthly maximum temperature for each month from Apr to Sep (°C)
TMIN4-TMIN9	Monthly minimum temperature for each month from Apr to Sep (°C)
PRE4-PRE9	Monthly total precipitation for each month from Apr to Sep (mm)
SRAD4-SRAD9	Monthly solar radiation for each month from Apr to Sep (MJ/m ²)
PET4-PET9	Monthly potential evapotranspiration for each month from Apr to Sep (mm)
AET4-AET9	Monthly actual evapotranspiration for each month from Apr to Sep (mm)
VPD4-VPD9	Monthly vapor pressure deficit for each month from Apr to Sep (h PA)
Satellite variables (total of 12)	
EVI4-EVI9	Monthly enhanced vegetation index for each month from Apr to Sep
SIF4-SIF9	Monthly solar-induced chlorophyll fluorescence for each month from Apr to Sep
Soil variables (total of 11)	
SM4-9	Monthly soil moisture for each month from Apr to Sep (mm)
GRAVEL	Gravel content (%)
SILT	Silt fraction (%)
CLAY	Clay fraction (%)
BULK	Bulk density (g/cm ³)
PH	pH

3.2. Multi-Model Performances in Estimating China's Maize Yield

3.2.1. The Performance of Predicted Maize Yield Models

We extracted the maize yield datasets from 1021 counties from 2000 to 2009 to train the model. After several trials, we found the optimal model performance when 70% of data samples were randomly chosen as the training set and the remaining 30% were used as the test set. Results (Figure 4) showed that three methods, RF, SVM, and XGB, stood out in terms of model performance in estimating China's maize yields during the period 2000–2009, with R^2 from 0.75 to 0.80, RMSE < 610 kg/ha, and MAE < 435 kg/ha. In terms of the three variable groups, the prediction performance of the combined group was constantly higher than that of the satellite group and climate group (Figure 4a–c). Finally, the three top-performing machine learning methods (i.e., RF, SVM, and XGB) were applied to predict county-level maize yield in China, leading to the following analysis. Descriptive statistics of the performance of the seven models are included in the Supplementary Materials (Table S3).

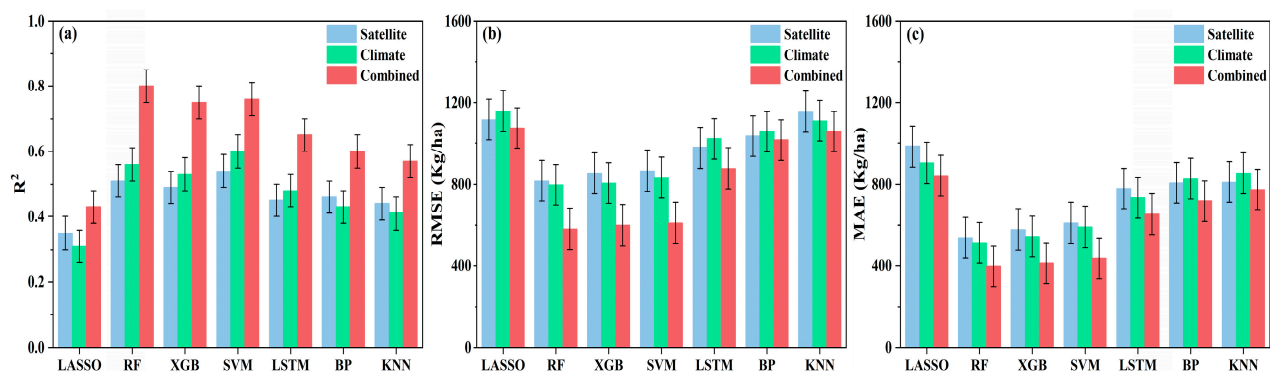


Figure 4. The performance of seven models in predicting China's maize yield from 2000 to 2009 in the test set. Inter-model comparisons were conducted among three groups (satellite, climate, and combined) based on ten-fold cross-validation. The accuracy among the seven models was assessed using (a) R^2 , (b) RMSE, and (c) MAE. Climate: climate dataset and soil dataset; satellite: satellite dataset and soil dataset; combined: satellite dataset, climate dataset, and soil dataset.

3.2.2. Comparison of Models in Predicting China's Maize Yield in 2010

We further used China's maize yield in 2010 as an independent dataset to verify the models' performance with RF, SVM, and XGB. When a single dataset group (satellite group or climate group) was fed into these models, R^2 results were lower than 0.6, and RMSE was greater than 1100 kg/ha (Figure 5d–i). The performance of the models trained with the climate group datasets ($R^2 = 0.53$ – 0.6 , RMSE = 1119–1168 kg/ha, and MAE = 790–832 kg/ha) was better than that of the models that trained with the satellite group datasets ($R^2 = 0.43$ – 0.54 , RMSE = 1151–1203 kg/ha, and MAE = 821–878 kg/ha; Figure 5d–i). The performance of models trained by the combined dataset group ($R^2 = 0.75$ – 0.81 , RMSE = 824–875 kg/ha, and MAE = 626–651 kg/ha) was much higher compared with the models trained by either of the individual dataset groups (i.e., climate or satellite group; Figure 5a–c). Overall, three machine learning models (RF, XGB, and SVM models) trained by the combined dataset group can better predict China's maize yield in 2010. The performance ranking in terms of prediction accuracy was: RF > SVM > XGB.

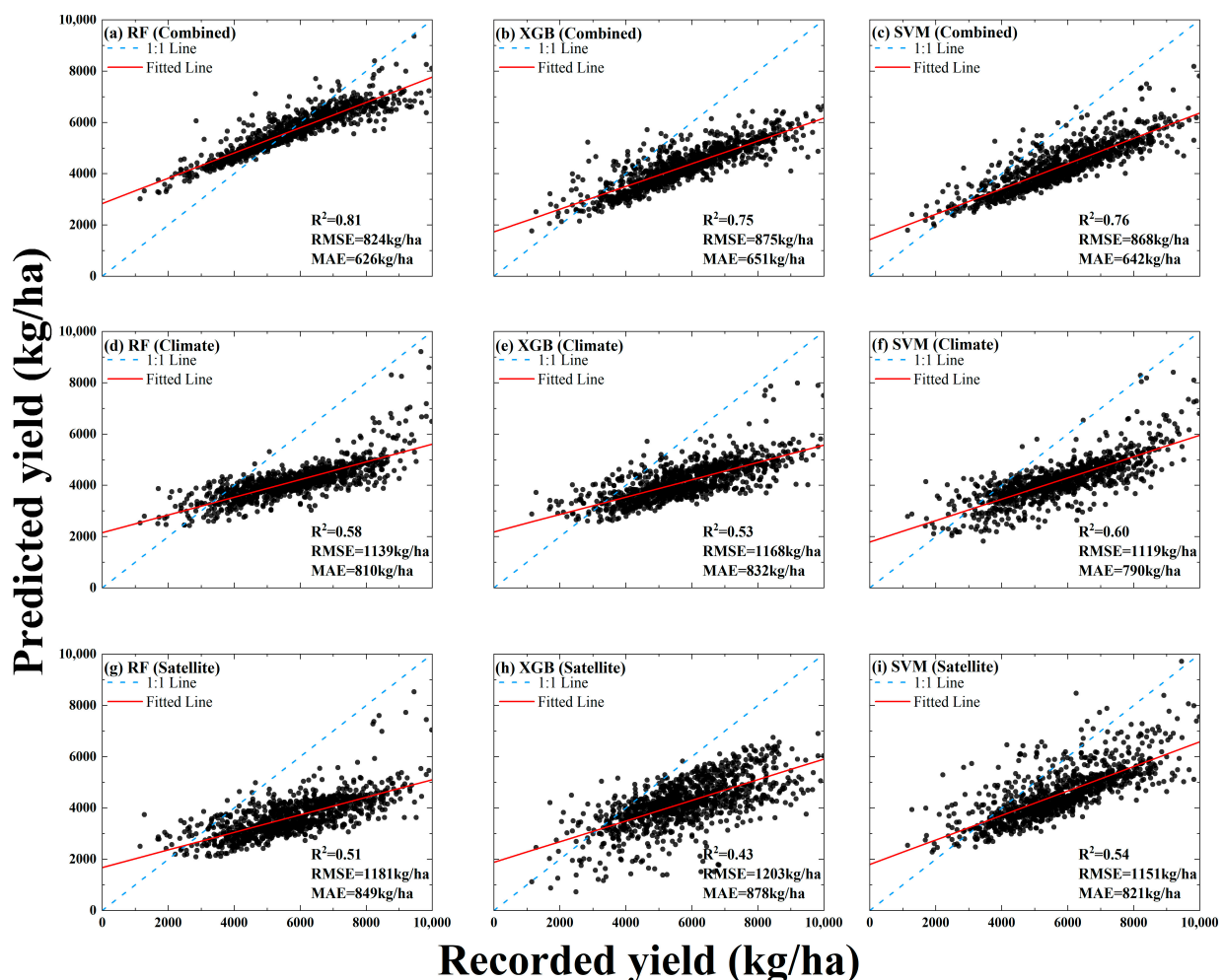


Figure 5. Scatter plots between observed maize yield and the model predictions from the RF, XGB, and SVM models. The red line represents the fit line, and the blue line represents the 1:1 line. These models were run using different dataset groups ((a–c) combined group, (d–f) climate group, (g–i) satellite group)) at the county level in 2010. RF: random forest; XGB: extreme gradient boosting; SVM: support vector machine. Climate: climate dataset and soil dataset; satellite: satellite dataset and soil dataset; combined: satellite dataset, climate dataset, and soil dataset.

3.2.3. Comparisons of Maize Yield Prediction at Different Growing Stages

To find out whether the performance of maize yield models was related to maize growth stages or not, we used all stage data from 2000 to 2009 to train the model to predict the maize yield in 2010 separately. We divided the maize growing season into three phases, as indicated in Figure 6 (the early stage: April to May; the peak stage: June to July; and the late stage: August to September). For all three models (RF, XGB, or SVM), the highest model accuracies were discovered during the peak stage ($R^2 = 0.65$, RMSE = 941 kg/ha, MAE = 686 kg/ha); it was much higher than the early-stage simulation ($R^2 = 0.45$, RMSE = 1143 kg/ha, MAE = 810 kg/ha) and slightly higher than the late-stage simulation ($R^2 = 0.56$, RMSE = 1024 kg/ha, MAE = 752 kg/ha). Thus, the satellite and climate datasets at the peak stage were essential for predicting maize yields.

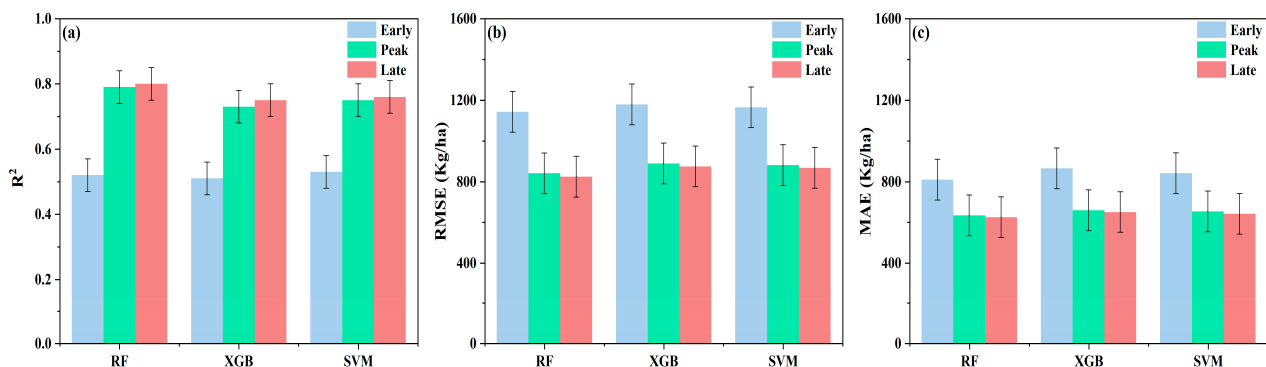


Figure 6. The performance of maize grain yield predictions ((a) R^2 , (b) RMSE, and (c) MAE) under three distinct maize growth phrases (the early stage: April to May; the peak stage: June to July; and the late stage: August to September) using the combined group. RF: random forest; XGB: extreme gradient boosting; SVM: support vector machine; combined group: satellite dataset, climate dataset, and soil dataset.

Figure 7 reveals the contributions of the climate dataset and satellite dataset from different growing stages to maize yield prediction. Figure 7a–c indicates that the R^2 of all three models consistently increased from April to September. For both the combined- and single-dataset groups, the model performance increased with the progression of time and reached the highest in September. From April to September, the models combined with the dataset group were quantified as the best models in predicting China's maize yield. Across the three models, the predictions of models based on the climate dataset group generally outperformed those of the satellite dataset group. While the predictiveness of these models continuously increased from April to September, the predictive accuracy of the satellite dataset group increased much more than that of the climate dataset group. The three models showed that the contribution of the satellite dataset (bottom panels in Figure 7) increases from the early to the peak stages (from April to July), and then stagnates. In contrast, there was an overall decreasing trend in the contribution of the climate datasets, and the gaps between the contribution of the climate dataset and satellite dataset gradually decrease. This means that as the growing season progresses, the influence of climate information could be gradually compensated by satellite information, and that the satellite dataset could closely track the growth of the maize.

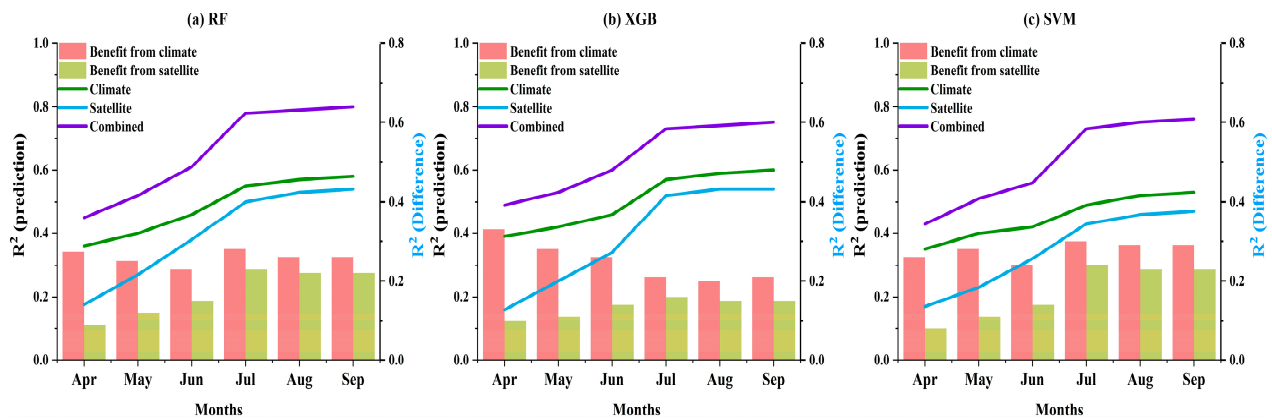


Figure 7. The temporal variations of the model performance of (a) RF, (b) XGB, and (c) SVM (the prediction at any specific month contains input data covering from the beginning of the growing season

to that specific month, thus the later period contains more inputs and usually has a higher performance). R^2 (prediction) presented in the solid lines indicates the performance of maize yield prediction models, while R^2 (difference) presented in the bars shows the contribution of satellite data and climate data to maize yield prediction. Climate: climate dataset and soil dataset; satellite: satellite dataset and soil dataset; combined: satellite dataset, climate dataset, and soil dataset.

3.3. Regional Differences in Predicting Maize Yield in China

3.3.1. Spatial Patterns of Predicted Maize Yield in China

The high-maize-yield counties in China were mainly observed in the Huang-Huai-Hai maize planting area and in northeastern and northwestern China (Figure 8a). Counties with relatively low maize yield were concentrated in the farming–grazing ecotone of northern China and the mountainous areas of southwestern China (Figure 8a). The predicted maize yield showed a similar spatial pattern as the recorded maize yield (Figure 8b–d). This finding indicated that the three models (RF, XGB, and SVM) were suitable for county-level maize yield predictions in China. Nonetheless, the difference between predictions and observations suggested that maize yields were underestimated in high-yielding counties and overestimated in low-yielding counties (Figure 8a–d).

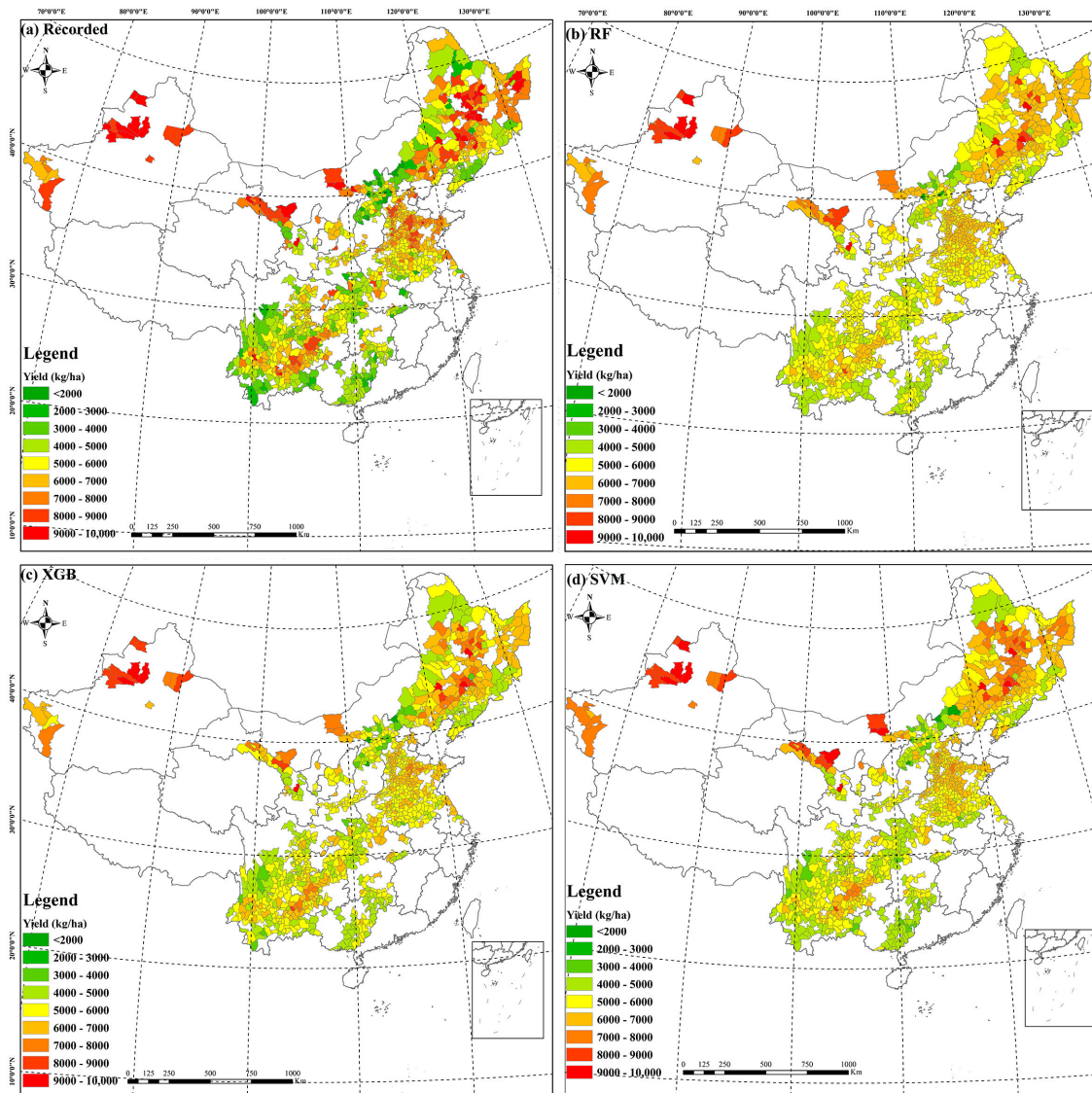


Figure 8. The spatial patterns of the recorded (a) and the predicted maize yield in 2010. The maize yield was predicted by RF (b), XGB (c), and SVM (d) models (RF: random forest; XGB: extreme gradient

boosting; SVM: support vector machine) using combined group (satellite dataset, climate dataset, and soil dataset).

3.3.2. Model Comparisons in Different Agricultural Zones

Across the four agricultural zones, the ranking of the prediction accuracy was RF > SVM > XGB (Figure 6). RF was the best in simulating China's maize yield, with a relatively lower simulating error of -1.69% (Figure 9). The prediction accuracies of the machine learning methods varied by the maize planting zones. SVM was the optimal model in Zones I and III, with percentage errors of -2.58% and 1.60% , while RF performed best in Zones II and IV, with percentage errors of -4.23% , -3.63% . All three models overestimated maize yield in Zone III, which featured a relatively lower maize yield (Figure 6).

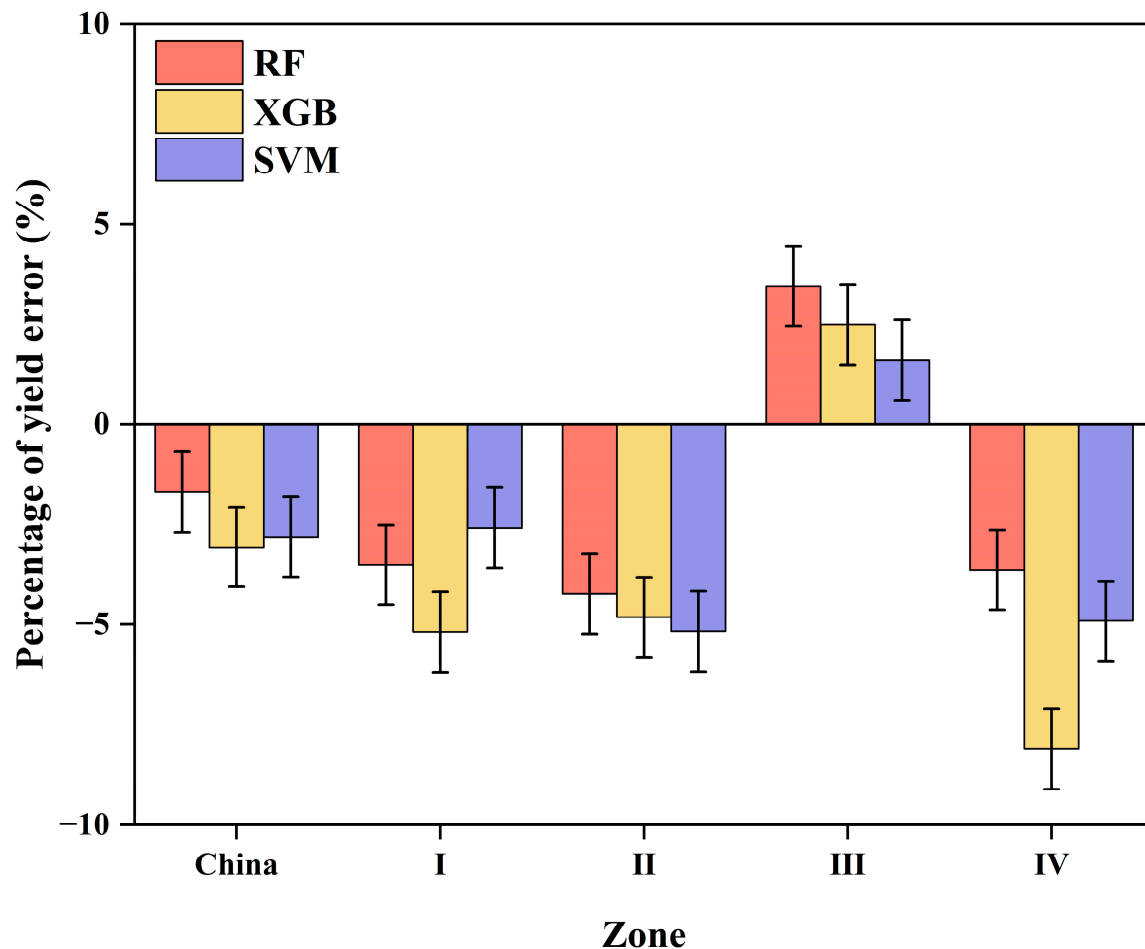


Figure 9. The percentage errors of maize yield in China and four maize planting zones (Zone I: the north maize planting zone; Zone II: the Huang-Huai-Hai maize planting zone; Zone III: the southwest maize planting zone; and Zone IV: the northwest maize planting zone).

3.3.3. The Relative Importance of Individual Variables in Maize Yield Prediction

Given the good performance, we used RF to identify the important influencing variables in the four maize planting zones and their relative importance among the top 14 variables (Figure 10). As for China and its four maize planting zones, some satellite variables (EVI7, SIF7) in July were consistently important in influencing maize yield forecasting accuracy (Figure 10a–e), and it was consistent with previous conclusions that satellite variables contribute the most to maize yield prediction at the peak stage. Although EVI7 was the most important variable throughout the study area, the order of variable importance varies across regions, as shown in Figure 10a. In Zone I, the temperature-related variable (TMIN7) was identified as the most important variable in maize yield prediction (Figure 10b). Satellite variables were relatively more important than climate variables in Zone II and Zone III (Figure 10c–d). In Zone IV, the water-related variable (VPD6) was the most important influencing variable (Figure 10e). In addition, soil variables (PH, CLAY, SM, GRAVEL, and SLIT) were also important influencing factors in maize yield predictions (Figure 10).

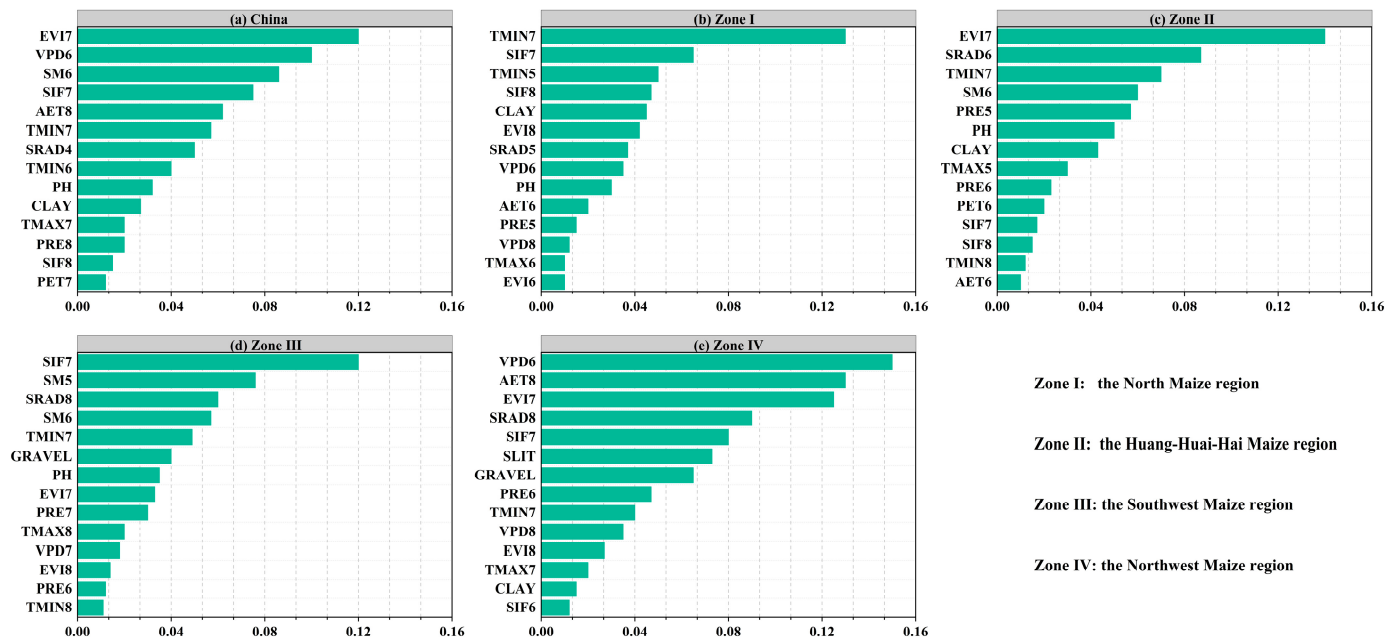


Figure 10. Feature importance values for the top 14 variables via RF models in the whole of (a) China and the four agricultural zones ((b) Zone I, (c) Zone II, (d) Zone III, (e) Zone IV). EVI: enhanced vegetation index; SIF: sun/solar-induced chlorophyll fluorescence; TMAX: maximum temperature; TMIN: minimum temperature; PRE: total precipitation; AET: actual evapotranspiration; PET: potential evapotranspiration; SRAD: downward surface shortwave radiation; VPD: vapor pressure deficit; GRAVEL: subsoil gravel content; SAND: sand fraction; SILT: subsoil silt fraction; CLAY: subsoil clay fraction; BULK: subsoil reference bulk density; OC: subsoil organic carbon; PH: subsoil pH; SM: soil moisture (numbers represent the number of months).

4. Discussion

Based on analysis of the seven methods, the machine learning methods were found to be superior to the LASSO algorithm. This can be primarily attributed to the capacity of machine learning methods to unravel intricate relationships between maize yield and variables [67]. Machine learning methods exhibit significant superiority over traditional linear statistical regression methods, boasting higher computational efficiency and enhanced spatial generalization capabilities. RF, XGB, and SVM showed better advantages in predicting China county maize yield. The sequence of the prediction accuracy was RF > SVM > XGB (Table S3), and RF showed the greatest improvement over linear regression models (R^2 increased by 0.37, RMSE decreased by 494 kg/ha, and MAE decreased by 441 kg/ha). These improvements highlight the reliability and effectiveness of the RF model for maize yield prediction in China. The data used in this study are highly complex in terms of dimensions (a total of 65 input variables), with intricate nonlinear relationships among the variables. RF, known for its robustness, exhibits adaptability to complex nonlinear relationships and high-dimensional data, enabling it to capture specific patterns or relationships within the data [54]. Traditionally, county-level maize yield estimation was conducted through surveys, which were both time-consuming and costly. Furthermore, traditional methods could only estimate maize yield after harvest, making it impractical for making policy decisions before harvest. In contrast, the machine learning models used in this study accurately predicted county-level maize yield and provided forecasts 1–2 months before harvest. Hence, these models can be widely employed for pre-harvest yield predictions, significantly saving resources, and enabling timely maize yield estimation based on the current year's climate conditions and environmental factors, thus facilitating policymaking processes.

Many studies suggest that the inclusion of multi-source datasets in yield prediction models usually increases the model accuracy when predicting maize yield. The satellite datasets provide mostly the biotic information related to crop growth [68,69], while the

climate datasets contain the abiotic information that also influences the crop yield [70–72]. Yet studies have suggested that various factors related to maize yield showed disparate sensitivities to the growing season [73,74]. For instance, the silking stage of maize in China can happen in July or August. During this stage, maize exhibits heightened sensitivity to environmental conditions. Satellite datasets possess the capability to capture the physiological activities of maize, encompassing its growth and photosynthetic intensity. These factors were closely linked to the yield performance of maize during critical growth stages [75]. Drought and high temperatures during critical growth stages, particularly silking, significantly impact maize yield. Soil moisture reduction limits the availability of water for plant growth, while decreased photosynthetic efficiency reduces the production of energy necessary for plant development. Accelerated growth processes due to high temperatures can also alter the accuracy of yield predictions [76]. These factors explain why the random forest model achieves higher accuracy at the peak (silking) stage of maize in our study, as it captures the complex interactions between these critical environmental factors and plant growth.

The satellite dataset at early growth stages of maize, when used independently in the model, was weakly associated with the yield. Between the two, the climate dataset presented higher performance than the satellite dataset at this stage, probably because the early seasonal climatic information should have captured some spatial patterns of the maize yield [20]. For example, temperature shows a clear spatial pattern from the early to the late growing season, with significant differences between different maize growing areas [38]. If maize yield is correlated with the temperature, the early season temperature could capture the spatial maize yield patterns, consequently leading to a better performance in model accuracy. On the contrary, the distribution of EVI or SIF in space is uniformly low in the early stage of maize growth, with little variation between different maize growing areas, and provides little information about spatial patterns [77]. Hence, the relatively higher predictive performance obtained in the early stage based on climate-related dataset groups (e.g., climate dataset and combined dataset) is found to be largely attributable to the spatial patterns of climate variables and could capture the more spatial maize yield patterns [20]. Additionally, we recognized that the satellite dataset can provide supplementary information for yield prediction models, further enhancing crop yield forecasts.

The maize yield would vary not only from season to season but also from location to location [16], and the regional differences in model performance are mostly caused by regional differences in environmental characteristics such as temperature and precipitation [78,79]. In Zone I, the colder climate prolongs the growth cycle, affecting pollination and grain formation. This underscores the importance of temperature-related features in predicting maize yield, as colder temperatures can restrict nutrient uptake and exacerbate pest and disease impacts [80]. In Zones IV and III, water-related variables emerge as crucial factors (Figure 10d,e). In the Northwest (Zone IV), the importance of VPD6 highlights the region's dependence on water availability for optimal yield [81]. In the Southwest (Zone III), excessive moisture can lead to rapid nitrogen loss in the soil, affecting maize yield [38,82]. The high ranking of soil moisture variables (SM5, SM6) reflects this region's sensitivity to soil moisture conditions (Figure 10d). We argue that for an accurate maize yield prediction in different areas of China, the models should be adapted to appropriate methods and variables. For example, in Zones II and IV, we recommend using RF for maize yield prediction, while in Zones I and III, the SVM is used. Additionally, the observed impact of excessive moisture on maize yield suggests that cumulative precipitation and cumulative temperature may significantly influence maize growth [82]. Therefore, incorporating variables such as cumulative precipitation and cumulative temperature for the growing season into the model could potentially enhance its accuracy.

Yet there are some limitations in our study. Ground truth data may have inherent biases and limitations. The methods of data collection and environmental conditions can impact results, and the representativeness of samples and data gaps can affect the

comprehensiveness of the analysis. In our study, despite implementing rigorous quality control measures, differences in data collection across different years and regions may lead to inconsistencies in the data. Future research could improve data collection methods, enhance sample diversity, utilize advanced technologies to increase data accuracy and comprehensiveness, and employ more robust uncertainty analysis methods to enhance the credibility and utility of the study. In addition, human management practices, such as irrigation, fertilization, and the selection of correct maize varieties as per season and location, are critical for maize growth and yield [83]. Their inclusion in the model could significantly improve predictions. Moreover, machine learning inherently operates as a black box, where model inputs and operations are not visible, posing challenges in comprehending internal decision-making processes and interpreting results [74]. This limitation restricts the informativeness of the conclusions drawn from our analysis in terms of plant physiology. Hence, it is still challenging to provide testable hypotheses giving biological explanations for crop growth and the final yield formation. Employing interpretability techniques to enhance understanding of model decisions, such as local interpretability methods and integrating crop process models, could improve the model's credibility, making it more acceptable in practical applications. Future applications of process-based models may help elaborate the mechanisms behind the relationships between variables and yield. Therefore, the issue of combining machine learning with crop process models is an intriguing approach that could be explored in further research [84].

5. Conclusions

In the present study, we improved the prediction accuracy of China's maize yield, specifically by employing a combination of multi-source datasets through various models. The performance of maize yield prediction was not sufficient when only the single satellite dataset or climate dataset was inputted into these models. As expected, machine learning methods outperformed traditional regression models. For example, RF, XGB, and SVM showed prominent advantages in predicting China's maize yield, while RF was the best method. Nonetheless, it is worth noting that the climate dataset contributed more to the maize yield prediction than that of the satellite dataset. The satellite dataset was able to closely track the growth cycle of crops; its contribution to maize yield prediction, however, generally reached a maximum at the peaking growth stage. We also found that the model accuracy and maize yield prediction varied across China's maize planting regions. The ML methods clearly showed advantages of spatial generalizations in providing new insights into crop yield prediction over a larger scale, and even a global scale. Our research could advance our understanding of the contributions of different growth stages of crops and multi-source dataset combinations in predicting maize yield at a larger spatial scale.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/rs16132417/s1>, Table S1: Data sources used in this study; Table S2: A total of 79 variables in the dataset; Table S3. The RMSE, MAE, and R^2 of county-level maize yield prediction model performance from 2000 to 2009.

Author Contributions: Conceptualization: L.M., Y.Z. and X.C.; investigation: L.M. and Y.Z.; data collection: L.M. and Y.Z.; software and formal analysis: L.M. and Y.Z.; writing—original draft: L.M. and Y.Z.; writing—review and editing: L.M., X.C., G.R.K., Y.S. and J.Z.; funding acquisition: L.M., Y.Z. and G.R.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (42101295), the Natural Science Foundation of Jiangsu Province (BK20210657), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX23_1294), and the Longshan Professorship and Talent Grant (1511582101011).

Data Availability Statement: All data resources are provided in the manuscript with links.

Acknowledgments: Giri Raj Kattel would like to acknowledge the Longshan Professorship and the Talent Grant at Nanjing University of Information Science & Technology.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Curtis, T.; Halford, N.G. Food security: The challenge of increasing wheat yield and the importance of not compromising food safety. *Ann. Appl. Biol.* **2014**, *164*, 354–372. [CrossRef] [PubMed]
2. Cole, M.B.; Augustin, M.A.; Robertson, M.J.; Manners, J.M. The science of food security. *NPJ Sci. Food* **2018**, *2*, 14. [CrossRef] [PubMed]
3. Hunter, M.C.; Smith, R.G.; Schipanski, M.E.; Atwood, L.W.; Mortensen, D.A. Agriculture in 2050: Recalibrating Targets for Sustainable Intensification. *BioScience* **2017**, *67*, 386–391. [CrossRef]
4. Keating, B.A.; Herrero, M.; Carberry, P.S.; Gardner, J.; Cole, M.B. Food wedges: Framing the global food demand and supply challenge towards 2050. *Glob. Food Secur.* **2014**, *3*, 125–132. [CrossRef]
5. Van Dijk, M.; Morley, T.; Rau, M.L.; Saghai, Y. A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nat. Food* **2021**, *2*, 494–501. [CrossRef] [PubMed]
6. IPCC. Climate Change and Land: An IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems. 2019. Available online: <https://www.ipcc.ch/srccl-report-download-page/> (accessed on 13 June 2022).
7. Gomez-Zavaglia, A.; Mejuto, J.C.; Simal-Gandara, J. Mitigation of emerging implications of climate change on food production systems. *Food Res. Int.* **2020**, *134*, 109256. [CrossRef] [PubMed]
8. Gonzalez, C.G. CLIMATE CHANGE, FOOD SECURITY, AND AGROBIODIVERSITY: TOWARD A JUST, RESILIENT, AND SUSTAINABLE FOOD SYSTEM. *Fordham Environ. Law Rev.* **2011**, *22*, 493–522. Available online: <http://www.jstor.org/stable/44175833> (accessed on 1 June 2022).
9. Godfray, H.C.J.; Beddington, J.R.; Crute, I.R.; Haddad, L.; Lawrence, D.; Muir, J.F.; Pretty, J.; Robinson, S.; Thomas, S.M.; Toulmin, C. Food Security: The Challenge of Feeding 9 Billion People. *Science* **2010**, *327*, 812–818. [CrossRef] [PubMed]
10. Zhang, Z.; Song, X.; Tao, F.; Zhang, S.; Shi, W. Climate trends and crop production in China at county scale, 1980 to 2008. *Theor. Appl. Climatol.* **2016**, *123*, 291–302. [CrossRef]
11. Paudel, D.; Boogaard, H.; de Wit, A.; Janssen, S.; Osinga, S.; Pylianidis, C.; Athanasiadis, I.N. Machine learning for large-scale crop yield forecasting. *Agric. Syst.* **2021**, *187*, 103016. [CrossRef]
12. Hao, S.; Ryu, D.; Western, A.; Perry, E.; Bogen, H.; Franssen, H.J.H. Performance of a wheat yield prediction model and factors influencing the performance: A review and meta-analysis. *Agric. Syst.* **2021**, *194*, 103278. [CrossRef]
13. Kang, Y.; Özdoğan, M. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. *Remote Sens. Environ.* **2019**, *228*, 144–163. [CrossRef]
14. Folberth, C.; Baklanov, A.; Balkovič, J.; Skalský, R.; Khabarov, N.; Obersteiner, M. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agric. For. Meteorol.* **2019**, *264*, 1–15. [CrossRef]
15. Asseng, S.; Cammarano, D.; Basso, B.; Chung, U.; Alderman, P.D.; Sonder, K.; Reynolds, M.; Lobell, D.B. Hot spots of wheat yield decline with rising temperatures. *Glob. Chang. Biol.* **2017**, *23*, 2464–2472. [CrossRef] [PubMed]
16. Filippi, P.; Jones, E.J.; Wimalathunge, N.S.; Somarathna, P.D.S.N.; Pozza, L.E.; Ugbaje, S.U.; Jephcott, T.G.; Paterson, S.E.; Whelan, B.M.; Bishop, T.F.A. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precis. Agric.* **2019**, *20*, 1015–1029. [CrossRef]
17. Liu, Y.; Heuvelink, G.B.M.; Bai, Z.; He, P.; Xu, X.; Ding, W.; Huang, S. Analysis of spatio-temporal variation of crop yield in China using stepwise multiple linear regression. *Field Crops Res.* **2021**, *264*, 108098. [CrossRef]
18. Cai, Y.; Guan, K.; Peng, J.; Wang, S.; Seifert, C.; Wardlow, B.; Li, Z. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote Sens. Environ.* **2018**, *210*, 35–47. [CrossRef]
19. Paudel, D.; Boogaard, H.; de Wit, A.; van der Velde, M.; Claverie, M.; Nisini, L.; Janssen, S.; Osinga, S.; Athanasiadis, I.N. Machine learning for regional crop yield forecasting in Europe. *Field Crops Res.* **2022**, *276*, 108377. [CrossRef]
20. Cai, Y.; Guan, K.; Lobell, D.; Potgieter, A.B.; Wang, S.; Peng, J.; Xu, T.; Asseng, S.; Zhang, Y.; You, L.; et al. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agric. For. Meteorol.* **2019**, *274*, 144–159. [CrossRef]
21. Joshi, D.R.; Clay, S.A.; Sharma, P.; Rekabdarkolaee, H.M.; Kharel, T.; Rizzo, D.M.; Thapa, R.; Clay, D.E. Artificial intelligence and satellite-based remote sensing can be used to predict soybean (*Glycine max*) yield. *Agron. J.* **2023**, *116*, 917–930. [CrossRef]
22. Ziliani, M.G.; Altaf, M.U.; Aragon, B.; Houborg, R.; Franz, T.E.; Lu, Y.; Sheffield, J.; Hoteit, I.; McCabe, M.F. Early season prediction of within-field crop yield variability by assimilating CubeSat data into a crop model. *Agric. For. Meteorol.* **2022**, *313*, 108736. [CrossRef]
23. Lobell, D.B.; Burke, M.B. On the use of statistical models to predict crop yield responses to climate change. *Agric. For. Meteorol.* **2010**, *150*, 1443–1452. [CrossRef]
24. Brown, J.N.; Hochman, Z.; Holzworth, D.; Horan, H. Seasonal climate forecasts provide more definitive and accurate crop yield predictions. *Agric. For. Meteorol.* **2018**, *260–261*, 247–254. [CrossRef]
25. Mayer, D.G.; Chandra, K.A.; Burnett, J.R. Improved crop forecasts for the Australian macadamia industry from ensemble models. *Agric. Syst.* **2019**, *173*, 519–523. [CrossRef]

26. Li, L.; Wang, B.; Feng, P.; Li Liu, D.; He, Q.; Zhang, Y.; Wang, Y.; Li, S.; Lu, X.; Yue, C.; et al. Developing machine learning models with multi-source environmental data to predict wheat yield in China. *Comput. Electron. Agric.* **2022**, *194*, 106790. [\[CrossRef\]](#)
27. Liu, Y.; Wang, S.; Wang, X.; Chen, B.; Chen, J.; Wang, J.; Huang, M.; Wang, Z.; Ma, L.; Wang, P.; et al. Exploring the superiority of solar-induced chlorophyll fluorescence data in predicting wheat yield using machine learning and deep learning methods. *Comput. Electron. Agric.* **2022**, *192*, 106612. [\[CrossRef\]](#)
28. Hatfield, J.L.; Gitelson, A.A.; Schepers, J.S.; Walthall, C.L. Application of Spectral Remote Sensing for Agronomic Decisions. *Agron. J.* **2008**, *100*, S-117–S-131. [\[CrossRef\]](#)
29. Mahlein, A.-K.; Oerke, E.-C.; Steiner, U.; Dehne, H.-W. Recent advances in sensing plant diseases for precision crop protection. *Eur. J. Plant Pathol.* **2012**, *133*, 197–209. [\[CrossRef\]](#)
30. Yu, Z.; Cao, Z.; Wu, X.; Bai, X.; Qin, Y.; Zhuo, W.; Xiao, Y.; Zhang, X.; Xue, H. Automatic image-based detection technology for two critical growth stages of maize: Emergence and three-leaf stage. *Agric. For. Meteorol.* **2013**, *174*–*175*, 65–84. [\[CrossRef\]](#)
31. Wang, R.; Cherkauer, K.; Bowling, L. Corn Response to Climate Stress Detected with Satellite-Based NDVI Time Series. *Remote Sens.* **2016**, *8*, 269. [\[CrossRef\]](#)
32. Gao, Y.; Wang, S.; Guan, K.; Wolanin, A.; You, L.; Ju, W.; Zhang, Y. The Ability of Sun-Induced Chlorophyll Fluorescence from OCO-2 and MODIS-EVI to Monitor Spatial Variations of Soybean and Maize Yields in the Midwestern USA. *Remote Sens.* **2020**, *12*, 1111. [\[CrossRef\]](#)
33. Leroux, L.; Falconnier, G.N.; Diouf, A.A.; Ndao, B.; Gbodjo, J.E.; Tall, L.; Balde, A.A.; Clermont-Dauphin, C.; Bégué, A.; Affholder, F.; et al. Using remote sensing to assess the effect of trees on millet yield in complex parklands of Central Senegal. *Agric. Syst.* **2020**, *184*, 102918. [\[CrossRef\]](#)
34. Pourmohammadali, B.; Hosseini-fard, S.J.; Hassan Salehi, M.; Shirani, H.; Esfandiarpour Boroujeni, I. Effects of soil properties, water quality and management practices on pistachio yield in Rafsanjan region, southeast of Iran. *Agric. Water Manag.* **2019**, *213*, 894–902. [\[CrossRef\]](#)
35. Jiang, Z.; Liu, C.; Ganapathysubramanian, B.; Hayes, D.J.; Sarkar, S. Predicting county-scale maize yields with publicly available data. *Sci. Rep.* **2020**, *10*, 14957. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Maitah, M.; Malec, K.; Ge, Y.; Gebeltová, Z.; Smutka, L.; Blažek, V.; Pánková, L.; Maitah, K.; Mach, J. Assessment and Prediction of Maize Production Considering Climate Change by Extreme Learning Machine in Czechia. *Agronomy* **2021**, *11*, 2344. [\[CrossRef\]](#)
37. Yang, Y.; Xu, W.; Hou, P.; Liu, G.; Liu, W.; Wang, Y.; Zhao, R.; Ming, B.; Xie, R.; Wang, K.; et al. Improving maize grain yield by matching maize growth and solar radiation. *Sci. Rep.* **2019**, *9*, 3635. [\[CrossRef\]](#)
38. Liu, W.; Li, Z.; Li, Y.; Ye, T.; Chen, S.; Liu, Y. Heterogeneous impacts of excessive wetness on maize yields in China: Evidence from statistical yields and process-based crop models. *Agric. For. Meteorol.* **2022**, *327*, 109205. [\[CrossRef\]](#)
39. Butler, E.E.; Huybers, P. Adaptation of US maize to temperature variations. *Nat. Clim. Chang.* **2013**, *3*, 68–72. [\[CrossRef\]](#)
40. Luo, Y.; Zhang, Z.; Chen, Y.; Li, Z.; Tao, F. ChinaCropPhen1km: A high-resolution crop phenological dataset for three staple crops in China during 2000–2015 based on leaf area index (LAI) products. *Earth Syst. Sci. Data* **2020**, *12*, 197–214. [\[CrossRef\]](#)
41. Mkhabela, M.S.; Bullock, P.; Raj, S.; Wang, S.; Yang, Y. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agric. For. Meteorol.* **2011**, *151*, 385–393. [\[CrossRef\]](#)
42. Dong, J.; Xiao, X.; Wagle, P.; Zhang, G.; Zhou, Y.; Jin, C.; Torn, M.S.; Meyers, T.P.; Suyker, A.E.; Wang, J.; et al. Comparison of four EVI-based models for estimating gross primary production of maize and soybean croplands and tallgrass prairie under severe drought. *Remote Sens. Environ.* **2015**, *162*, 154–168. [\[CrossRef\]](#)
43. Pu, R.; Gong, P. Wavelet transform applied to EO-1 hyperspectral data for forest LAI and crown closure mapping. *Remote Sens. Environ.* **2004**, *91*, 212–224. [\[CrossRef\]](#)
44. Haboudane, D.; Miller, J.R.; Pattey, E.; Zarco-Tejada, P.J.; Strachan, I.B. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sens. Environ.* **2004**, *90*, 337–352. [\[CrossRef\]](#)
45. He, M.; Kimball, J.S.; Yi, Y.; Running, S.; Guan, K.; Jensco, K.; Maxwell, B.; Maneta, M. Impacts of the 2017 flash drought in the US Northern plains informed by satellite-based evapotranspiration and solar-induced fluorescence. *Environ. Res. Lett.* **2019**, *14*, 074019. [\[CrossRef\]](#)
46. Guan, K.; Wu, J.; Kimball, J.S.; Anderson, M.C.; Frolking, S.; Li, B.; Hain, C.R.; Lobell, D.B. The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. *Remote Sens. Environ.* **2017**, *199*, 333–349. [\[CrossRef\]](#)
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* **2018**, *5*, 170191. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Zhang, L.; Zhang, Z.; Luo, Y.; Cao, J.; Tao, F. Combining Optical, Fluorescence, Thermal Satellite, and Environmental Data to Predict County-Level Maize Yield in China Using Machine Learning Approaches. *Remote Sens.* **2020**, *12*, 21. [\[CrossRef\]](#)
49. Wieder, W. *Regridded Harmonized World Soil Database v1.2, Version 1*; DAAC: West Melbourne, Australia, 2014. [\[CrossRef\]](#)
50. Liu, Y.; Qin, Y.; Ge, Q. Spatiotemporal differentiation of changes in maize phenology in China from 1981 to 2010. *J. Geogr. Sci.* **2019**, *29*, 351–362. [\[CrossRef\]](#)
51. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
52. Rhee, J.; Im, J. Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data. *Agric. For. Meteorol.* **2017**, *237*–*238*, 105–122. [\[CrossRef\]](#)

53. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 25. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Vincenzi, S.; Zucchetta, M.; Franzoi, P.; Pellizzato, M.; Pranovi, F.; De Leo, G.A.; Torricelli, P. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of *Ruditapes philippinarum* in the Venice lagoon, Italy. *Ecol. Model.* **2011**, *222*, 1471–1478. [\[CrossRef\]](#)
55. Brereton, R.G.; Lloyd, G.R. Support Vector Machines for classification and regression. *Analyst* **2010**, *135*, 230–267. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28. [\[CrossRef\]](#)
57. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
58. Wang, B.; Gu, X.; Ma, L.; Yan, S. Temperature error correction based on BP neural network in meteorological wireless sensor network. *Int. J. Sens. Netw.* **2017**, *23*, 265–278. [\[CrossRef\]](#)
59. Li, J.; Cheng, J.-h.; Shi, J.-y.; Huang, F. Brief Introduction of Back Propagation (BP) Neural Network Algorithm and Its Improvement. In Proceedings of the Advances in Computer Science and Information Engineering, Berlin/Heidelberg, Germany, 19–20 May 2012; pp. 553–558. [\[CrossRef\]](#)
60. Bélisle, E.; Huang, Z.; Le Digabel, S.; Gheribi, A.E. Evaluation of machine learning interpolation techniques for prediction of physical properties. *Comput. Mater. Sci.* **2015**, *98*, 170–177. [\[CrossRef\]](#)
61. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#)
62. Graves, A. Long Short-Term Memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Graves, A., Ed.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45. [\[CrossRef\]](#)
63. Appelhans, T.; Mwangomo, E.; Hardy, D.R.; Hemp, A.; Naus, T. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. *Spat. Stat.* **2015**, *14*, 91–113. [\[CrossRef\]](#)
64. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [\[CrossRef\]](#)
65. Ransam, J.; Cook, J.A. LASSO regression. *Br. J. Surg.* **2018**, *105*, 1348. [\[CrossRef\]](#)
66. Sylvain, A.; Alain, C. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [\[CrossRef\]](#)
67. Crane-Droesch, A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ. Res. Lett.* **2018**, *13*, 114003. [\[CrossRef\]](#)
68. Lambert, M.-J.; Traoré, P.C.S.; Blaes, X.; Baret, P.; Defourny, P. Estimating smallholder crops production at village level from Sentinel-2 time series in Mali's cotton belt. *Remote Sens. Environ.* **2018**, *216*, 647–657. [\[CrossRef\]](#)
69. Yang, K.; Ryu, Y.; Dechant, B.; Berry, J.A.; Hwang, Y.; Jiang, C.; Kang, M.; Kim, J.; Kimm, H.; Kornfeld, A.; et al. Sun-induced chlorophyll fluorescence is more strongly related to absorbed light than to photosynthesis at half-hourly resolution in a rice paddy. *Remote Sens. Environ.* **2018**, *216*, 658–673. [\[CrossRef\]](#)
70. Azzari, G.; Jain, M.; Lobell, D.B. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. *Remote Sens. Environ.* **2017**, *202*, 129–141. [\[CrossRef\]](#)
71. Lobell, D.B.; Thau, D.; Seifert, C.; Engle, E.; Little, B. A scalable satellite-based crop yield mapper. *Remote Sens. Environ.* **2015**, *164*, 324–333. [\[CrossRef\]](#)
72. Jin, Z.; Azzari, G.; You, C.; Di Tommaso, S.; Aston, S.; Burke, M.; Lobell, D.B. Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sens. Environ.* **2019**, *228*, 115–128. [\[CrossRef\]](#)
73. Kim, N.; Lee, Y.-W. Machine Learning Approaches to Corn Yield Estimation Using Satellite Images and Climate Data: A Case of Iowa State. *J. Korean Soc. Surv. Geod. Photogramm. Cartogr.* **2016**, *34*, 383–390. [\[CrossRef\]](#)
74. Zhao, Y.; Lobell, D.B. Assessing the heterogeneity and persistence of farmers' maize yield performance across the North China Plain. *Field Crops Res.* **2017**, *205*, 55–66. [\[CrossRef\]](#)
75. Shanahan, J.F.; Schepers, J.S.; Francis, D.D.; Varvel, G.E.; Wilhelm, W.W.; Tringe, J.M.; Schlemmer, M.R.; Major, D.J. Use of Remote-Sensing Imagery to Estimate Corn Grain Yield. *Agron. J.* **2001**, *93*, 583–589. [\[CrossRef\]](#)
76. Benincasa, P.; Reale, L.; Tedeschini, E.; Ferri, V.; Cerri, M.; Ghitarrini, S.; Falcinelli, B.; Frenguelli, G.; Ferranti, F.; Ayano, B.E.; et al. The relationship between grain and ovary size in wheat: An analysis of contrasting grain weight cultivars under different growing conditions. *Field Crops Res.* **2017**, *210*, 175–182. [\[CrossRef\]](#)
77. Zhou, W.; Liu, Y.; Ata-Ul-Karim, S.T.; Ge, Q.; Li, X.; Xiao, J. Integrating climate and satellite remote sensing data for predicting county-level wheat yield in China using machine learning methods. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *111*, 102861. [\[CrossRef\]](#)
78. Ji, B.; Sun, Y.; Yang, S.; Wan, J. Artificial neural networks for rice yield prediction in mountainous regions. *J. Agric. Sci.* **2007**, *145*, 249–261. [\[CrossRef\]](#)
79. Chen, X.; Feng, L.; Yao, R.; Wu, X.; Sun, J.; Gong, W. Prediction of Maize Yield at the City Level in China Using Multi-Source Data. *Remote Sens.* **2021**, *13*, 146. [\[CrossRef\]](#)
80. Chen, C.; Lei, C.; Deng, A.; Qian, C.; Hoogmoed, W.; Zhang, W. Will higher minimum temperatures increase corn production in Northeast China? An analysis of historical data over 1965–2008. *Agric. For. Meteorol.* **2011**, *151*, 1580–1588. [\[CrossRef\]](#)
81. Wang, Y.; Zhao, W.; Zhang, Q.; Yao, Y.-b. Characteristics of drought vulnerability for maize in the eastern part of Northwest China. *Sci. Rep.* **2019**, *9*, 964. [\[CrossRef\]](#) [\[PubMed\]](#)

82. Folberth, C.; Elliott, J.; Müller, C.; Balkovic, J.; Chryssanthacopoulos, J.; Izaurrealde, R.C.; Jones, C.D.; Khabarov, N.; Liu, W.; Reddy, A.; et al. Uncertainties in global crop model frameworks: Effects of cultivar distribution, crop management and soil handling on crop yield estimates. *Biogeosci. Discuss.* **2016**, *2016*, 1–30. [[CrossRef](#)]
83. Jin, Z.; Azzari, G.; Burke, M.; Aston, S.; Lobell, D.B. Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. *Remote Sens.* **2017**, *9*, 931. [[CrossRef](#)]
84. Pierre Pott, L.; Jorge Carneiro Amado, T.; Augusto Schwalbert, R.; Mateus Corassa, G.; Antonio Ciampitti, I. Crop type classification in Southern Brazil: Integrating remote sensing, crop modeling and machine learning. *Comput. Electron. Agric.* **2022**, *201*, 107320. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.