remote sensing

MDPI

Article

Predicting Maize Yield at the Plot Scale of Different Fertilizer Systems by Multi-Source Data and Machine Learning Methods

Linghua Meng [1,2], Huanjun Liu [1,*], Susan L. Ustin [3] and Xinle Zhang [4]

[1] Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130102, China; mlhcrop@yeah.net
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Center for Spatial Technologies and Remote Sensing (CSTARS), Department of Land, Air, and Water Resources, University of California, Davis, CA 95616, USA; slustin@ucdavis.edu
[4] School of Information Technology, Jilin Agricultural University, Changchun 130102, China; zhangxinle@gmail.com
* Correspondence: liuhuanjun@iga.ac.cn

Abstract: Timely and reliable maize yield prediction is essential for the agricultural supply chain and food security. Previous studies using either climate or satellite data or both to build empirical or statistical models have prevailed for decades. However, to what extent climate and satellite data can improve yield prediction is still unknown. In addition, fertilizer information may also improve crop yield prediction, especially in regions with different fertilizer systems, such as cover crop, mineral fertilizer, or compost. Machine learning (ML) has been widely and successfully applied in crop yield prediction. Here, we attempted to predict maize yield from 1994 to 2007 at the plot scale by integrating multi-source data, including monthly climate data, satellite data (i.e., vegetation indices (VIs)), fertilizer data, and soil data to explore the accuracy of different inputs to yield prediction. The results show that incorporating all of the datasets using random forests (RF) and AB (adaptive boosting) can achieve better performances in yield prediction (R^2: 0.85~0.98). In addition, the combination of VIs, climate data, and soil data (VCS) can predict maize yield more effectively than other combinations (e.g., combinations of all data and combinations of VIs and soil data). Furthermore, we also found that including different fertilizer systems had different prediction accuracies. This paper aggregates data from multiple sources and distinguishes the effects of different fertilization scenarios on crop yield predictions. In addition, the effects of different data on crop yield were analyzed in this study. Our study provides a paradigm that can be used to improve yield predictions for other crops and is an important effort that combines multi-source remotely sensed and environmental data for maize yield prediction at the plot scale and develops timely and robust methods for maize yield prediction grown under different fertilizing systems.

Keywords: maize; yield prediction; fertilizer systems; machine learning

Citation: Meng, L.; Liu, H.; L. Ustin, S.; Zhang, X. Predicting Maize Yield at the Plot Scale of Different Fertilizer Systems by Multi-Source Data and Machine Learning Methods. *Remote Sens.* **2021**, *13*, 3760. https://doi.org/10.3390/rs13183760

Academic Editor: Clement Atzberger

Received: 5 August 2021
Accepted: 17 September 2021
Published: 19 September 2021

1. Introduction

Sustainable crop yield is the ultimate goal of farmland cultivation and it is also a direct indicator of farmland productivity and income. Maize (*Zea mays* L.) is the staple food for more than 4.5 billion people, and the demand is expected to double by 2050 [1]. Therefore, timely and accurate prediction of maize yield is vital for not only international policy but also for grain storage and trade. Traditional crop yield prediction primarily relies on models and statistical regression methods [2]. Remote sensing (RS) technology is objective, low cost, and rapid, and can overcome the limitations of traditional field methods for crop yield prediction. Previous studies have mostly used the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), and enhanced vegetation index 2 (EVI 2) to predict crop yield and biomass [3,4]. In recent decades, it has become a popular

technology for crop growth monitoring and yield prediction [5], but predicting yields using only RS limits their accuracy [6]. Statistical regression methods provide a simpler method for yield prediction, but such statistical models are typically localized and are unable to be extended to other areas [7]. With some explicit cause–effect relationships, statistical regression models have been increasingly replaced by crop models in recent years due to their explanatory power and spatial generalization [8]. In addition, climate variables (e.g., temperature and precipitation) and soil data are the primary inputs for crop yield prediction because they can capture important environmental information. Most statistical models predict yields by developing regression equations between climate variables (temperature, precipitation, solar radiation, etc.) and measured yields at different temporal and spatial scales [9]. Additionally, previous studies have confirmed that fertilizer significantly affects crop production [10], especially with irrigation, fertilizer application, pesticide use, farm mechanization status, among other factors. For example, adopting the best farm management practices could increase crop yield [11]. However, the yield of different fertilizer systems is different [12], and few studies have considered the contribution of different fertilizer factors (mineral fertilizer, compost, and cover crop) for crop yield prediction. Additionally, the combination of input variables from different fertilizer systems that could achieve the best results is unclear.

Crop yield is a function of the interaction between the spatial and temporal changes of variables, and crop yield prediction is affected by many variables. For example, satellite observations, climate variables, and soil properties, can be used for capturing yield variability [3]. In recent decades, many researchers have been increasingly focused on improving crop yield prediction by means of different methods, and machine learning (ML) is an immediate successor of older statistical methods that adopts important weights rather than the likelihood or probability of any forecasting information [13].

Machine learning is a sub-class of artificial intelligence. It is self-learning based on algorithms, which means that the system learns from its experience. For instance, the type of data input to the system learns the pattern and responds to it, resulting in the model learning at the output. It uses a statistical learning algorithm that automatically learns and improves without human help. Predictions based on historical data can be conducted using machine learning. Various applications include stock pricing predictions, scientific research, marketing campaigns, and many more. Generally, artificial neural networks and random forest algorithms are used for predictions [14]. Nawar et al. (2016) showed that a model based on partial least squares regression (PLSR) can perform well in predicting soil organic matter, whereas Knox et al. (2015) found that a model based on the random forest (RF) approach is better, particularly for data that do not follow a normal distribution [15,16]. Viscarra Rosseland Behrens (2010) reported that support vector machines (SVM) can achieve more accurate predictions than PLSR and RF models [17].

Therefore, ML is more effective for noisy data and is able to interpret nonlinear relationships. Additionally, it has been widely and successfully applied in crop yield prediction. Accordingly, ML could provide powerful support for improving yield prediction models. Hunt et al. (2019) trained an RF model with high-resolution Sentinel-2 images and mapped at the field-scale wheat yield at a 10 m resolution in the UK [18]. Cai et al. (2019) accurately predicted county-scale wheat yield in Australia using three ML methods and confirmed that their performances were much better than the traditional regression model [19]. Several previous studies have proved its ability to improve crop yield prediction [5]. However, such methods have rarely been tested for crop yield prediction under different regional fertilizer systems.

In view of the current research, many of the problems indicated in previous studies should be considered, so we integrated 15 indicators derived from remote sensing data, climate data, fertilizer data, and soil properties data to build ML models to predict maize yield. We adopted six machine learning models to predict maize yield at the plot scale, including linear regression (LR), K-nearest neighbor (KNN), support vector machines (SVM), Gaussian process regression (GPR), adaptive boost (AB), and random forests (RF).

Our main objectives were (1) to construct a maize yield prediction framework and analysis with the combination of input variables could achieve the best results, (2) to explore the differences among these agricultural systems for accurate yield prediction and identify the relative importance of all variables, and (3) to identify the critical factors for maize yield prediction by considering the different fertilization systems.

2. Material and Methods

2.1. Study Area

The Russell Ranch Sustainable Agriculture Facility (RRSAF) is a 120-ha facility near the University of California, Davis (UC Davis) campus dedicated to investigating irrigated and dryland agriculture in a Mediterranean climate and is a core unit of the Agricultural Sustainability Institute at UC Davis. The RRSAF houses a 100-year study referred to as the "Century Experiment", formerly called the Long-Term Research in Agricultural Systems (LTRAS). Initiated in 1993, the Century Experiment comprised of 72 0.4-ha plots including 10 different replicated cropping systems [20]. Cropping systems were designed to compare the resource-use efficiency, productivity, environmental effects, and economic return from cropping systems that differ in crop rotation and degree of reliance on rainfall and fertilizer nitrogen. All field operations use full-scale agricultural equipment identical or similar to those used by local commercial farming operations and are either owned by RRSAF, leased from UC Davis facilities, or borrowed from local farmers. The RRSAF has a Mediterranean climate with monthly average minimum temperatures varying from 2.9 °C during the coldest month (December) to monthly average maximum temperatures of 33.7 °C during the warmest month (July). The mean annual rainfall is 440 mm. In this paper, we chose three fertilizer systems: conventional, organic tomato–maize rotations (CMT with mineral fertilizer and OMT with compost and winter cover crop (WCC), respectively), and legume–maize-tomato (LMT) rotation with WCC [12,21] shown in Figure 1.

2.2. Satellite Images

In this paper, the Landsat TM 5 from 1994 to 2007 of the study area were acquired (http://earthexplorer.usgs.gov/ (accessed on 18 September 2021)) (see Table 1). There was no cloud cover in the study area during that time. Radiometric calibration and atmospheric correction of the Landsat data were performed as a preprocessing step using the fast line-of-sight atmospheric analysis of spectral hypercubes (FLAASH) in ENVI5.1. A vegetation index can monitor dynamic changes in vegetation. Currently, many studies have shown that the VIs measured in this paper have good correlations with crop yield [4,22]. We collected three types of vegetation index data (VI), the normalized difference vegetation index (NDVI), the enhanced vegetation index (EVI), and the enhanced vegetation index 2 (EVI2). In this paper, the time-series NDVI curve is used as an example to analyze the correlation coefficient between yield and NDVI.

2.3. Soil Data and Fertilizer Data

Soil is a crucial factor affecting crop yield, as are the organic matter and the physical and chemical attributes distributed in different areas of soil. We selected seven variables that describe the physical and mineral properties of the soil. For example, organic matter content for the topsoil layer (0–25 cm and 25–50 cm), soil particle distribution for the topsoil layer, and soil bulk density data were obtained in different systems in 1992 or 1993.

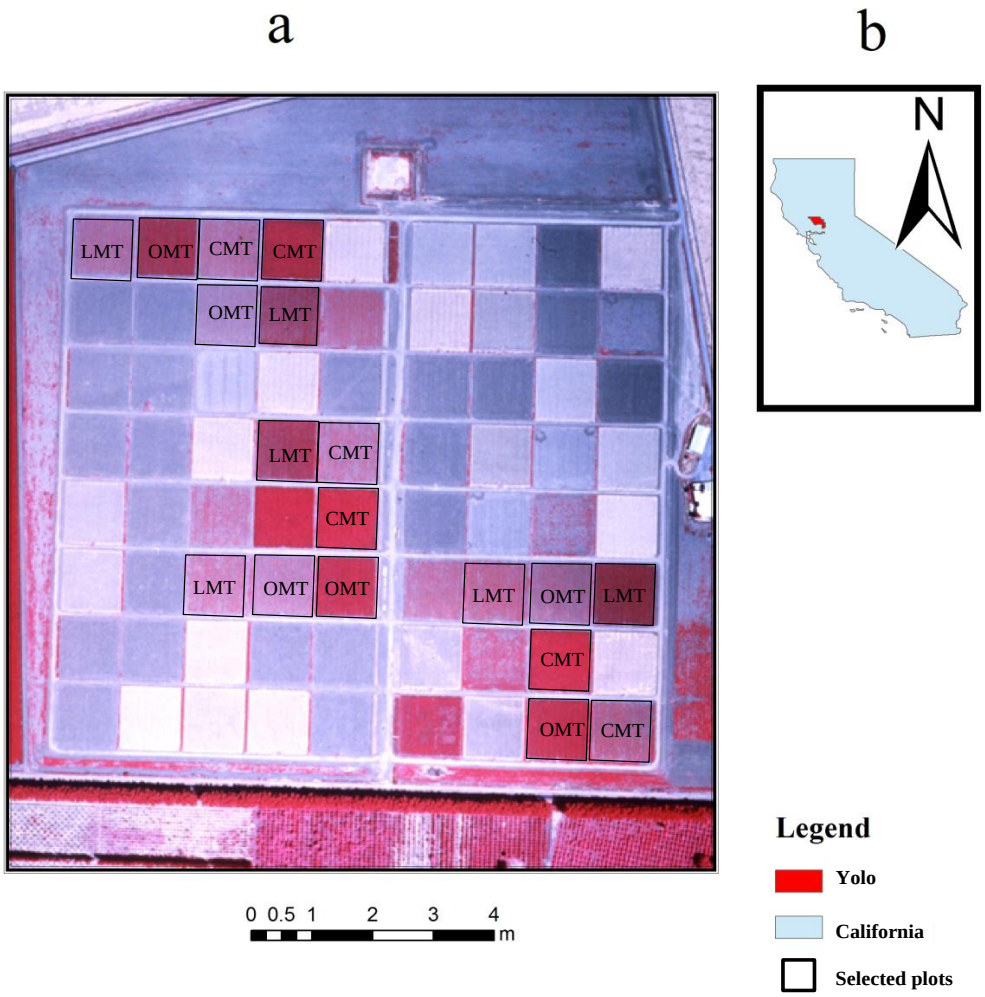

Figure 1. Study area (aerial image on 13 June 1994) ((**a**), study farm; (**b**), California and Yolo).

The CMT system in RRSAF received mineral fertilizer, and it was applied twice, including during the application of the base and topdressing. The LMT system was planted with a WCC. The OMT system received chicken manure compost from Foster Farms (Livingston, CA, USA) and was planted with a WCC.

2.4. Climate Data

Climate variables (e.g., temperature, precipitation) are important drivers for crop yield. We selected yearly maximum temperature (TMAX), yearly minimum temperature (TMIN), and yearly total precipitation (PRE)to predict the maize yield. Climate data were obtained from the meteorological observation station at Russell Ranch [21].

2.5. Machine-Learning Methods for Estimating Crop Yield

All of the variables and yields from 1994 to 2007 were normalized; therefore, all of variables in the models are at a common level and are comparable. In this paper, we had small data sets, so the final model was trained using the full data set. The five-fold cross-validated (CV) algorithm was used to evaluate model performance, which protects against overfitting [23]. This method provides a good estimate of the predictive accuracy of the final model trained using the full data set. The method requires multiple fits, but it makes efficient use of all of the data, so it works well for small data sets. Appendix A lists all of the input and output data.

Table 1. Selected RS images in this paper.

Year	Dates	Numbers	Sensors
1994	4/28, 5/30, 7/1, 7/17, 8/2, 8/18, 9/3, 9/19, 10/21	9	Landsat TM 5
1995	5/1, 5/17, 6/18, 7/4, 7/20, 8/5, 8/21, 9/6, 9/22, 10/08, 10/24	11	
1996	6/4, 6/20, 7/6, 7/22, 8/7, 9/24, 10/10, 10/26	8	
1997	5/6, 6/7, 6/23, 7/9, 7/25, 9/11, 9/27, 10/13	8	
1998	5/25, 6/26, 7/12, 7/28, 8/29, 9/14, 10/16	7	
1999	5/28, 6/13, 6/29, 7/31, 8/16, 9/17, 10/3, 10/19	8	
2000	5/30, 6/15, 7/1, 7/17, 8/2, 8/18, 9/3, 9/19, 10/5	9	
2001	5/1, 6/2, 6/18, 8/5, 8/21, 9/6, 10/24	7	
2002	6/5, 6/21, 7/7, 7/23, 8/8, 8/24, 9/9, 9/25, 10/11	9	
2003	5/23, 6/8, 6/24, 7/10, 7/26, 8/11, 8/27, 9/12, 10/14	10	
2004	5/9, 5/25, 6/10, 6/26, 7/12, 7/28, 8/13, 8/29, 9/14	9	
2005	5/12, 6/13, 6/29, 7/15, 7/31, 8/16, 9/1, 9/17, 10/3,	9	
2006	6/16, 7/2, 7/18, 8/3, 8/19, 9/4, 9/20, 10/6, 10/22	9	
2007	5/18, 6/3, 6/19, 7/5, 7/21, 8/22, 9/7	7	

2.5.1. Linear Regression (LR)

LR is the first type of regression analysis that has been well studied and widely used in practical applications [24]. This is because the linear model depends on its unknown parameters, making it easier to fit than the nonlinear model, which depends on its position parameters, and the statistical characteristics of the estimation that is produced are easier to determine. By comparing different kernel functions, the stepwise regression algorithm performed the best in this study.

2.5.2. Support Vector Machine (SVM)

SVM is a supervised non-parametric algorithm that is characterized by the use of kernels and by acting on the margin [25]. During SVM regression, 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. Kernel functions (linear, polynomial, Gaussian, etc.) are one of the most

important hyper-parameters that need to be tuned. By comparing different kernel functions, the polynomial kernel function was found to perform the best in this study.

2.5.3. Gaussian Process Regression (GPR)

GPR is a generalized Gaussian probability distribution for nonlinear regressions and a nonparametric method suitable for a variety of situations, especially for high-dimensional space problems [26]. The Gaussian process is a collection of random variables whose properties are any finite number of subsets with a joint Gaussian distribution. However, matrix inversion is a necessary challenge that needs to be handled, which increases the computational complexity and causes the model to run very slowly. The GPR used in this paper is based on the exponential kernel function.

2.5.4. KNN

KNN is a type of instance-based learning that calculates the distance of the predictor variables to the nearest training group known to the model. KNN tolerates noise and unrelated properties and has a relatively relaxed concept bias [27].

2.5.5. Adaptive Boost (AB)

AB is an iterative algorithm. Its core idea is to train different weak classifiers with the same training set and then combine these weak classifiers to form a strong classifier. AB can deal with classification and regression problems, and an advantage of AB is limiting overfitting [28].

2.5.6. Random Forests (RF)

Random forests are a combination of tree predictors and are robust with respect to noise. Each tree is built by selecting random variable sets and dataset samples, and all of the trees in the forest have the same distribution characteristics. After generating a large number of individual trees, they determine the final classes based on which trees were selected the most often. Therefore, RF has the efficiency to handle high-dimensional datasets and has avoided overfitting over the past decade of use. Additionally, RF can quantify the relative importance of measured variables and is a reasonable method for variable selection [29].

2.6. Model Evaluation

In order to evaluate the six ML models, cross-validation (CV) is a widely used strategy for algorithm selection because of its simplicity, universality, and efficiency in avoiding the overfitting issue [30]. It is generally accepted that a model with the smallest estimation error is the best model.

The data from 1994–2006 in even years were used for training and testing, and the data from 1995–2007 in odd years were used to verify the predicted yield accuracy.

We adopted the root-mean-square error (*RMSE*) and the coefficient of determination (R^2) to evaluate the performance of the ML models, which can be calculated as follows:

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \overline{y_i})(x_i - \overline{x_i})\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2 \sum_{i=1}^{n}(x_i - \overline{x_i})^2} \quad (1)$$

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

where i (i = 1,2,..., n) is the number of samples used for machine learning model, y_i is the measured maize yield, $\overline{y_i}$ is the corresponding mean value, x_i is the predict maize yield, and $\overline{x_i}$ is the corresponding mean value. The closer R^2 is to 1, the higher the prediction performance of the model is. A small *RMSE* value indicates less discrepancy within the measured yield and predicted yield.

3. Results

3.1. The Key Time Selected for Vis from Correlation Coefficients between Yield and NDVI

Figure 2 is time-series NDVI and the correlation coefficient between the maize yield and the NDVI in the even years from 1994 to 2006. Figure 2 shows that the NDVI in bot h the early stage and end stage were strongly correlated with each other, and both show significant correlations with maize yield ($p < 0.05$). As for different systems, the overall trend of the time-series curves of different systems is same, showing a parabola of first growth and then shows decline. Due to the application of base fertilizer and top dressing, the NDVI of CMT is better than the other two systems in the early growth stage of maize, and OMT is better than the other two systems in the vigorous growth stage of maize. Overall, from the correlation coefficients between NDVI and yield, the dominant time windows of the VIs controlling yields are both in the early stage and end stage of maize growth. Therefore, the VIs in early maize growth from the end of May to the beginning of June were selected to predict maize yield in this study; VIs were also chosen due to the RS images obtained in this study area.

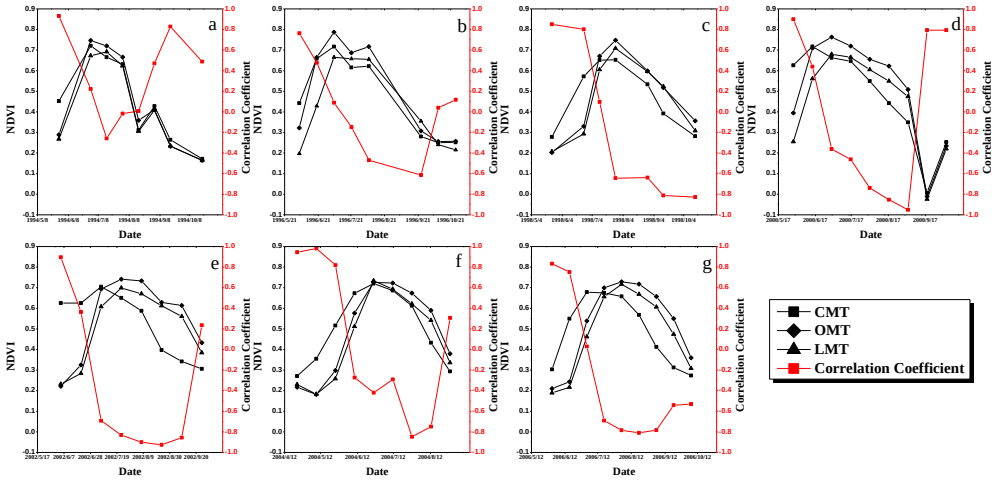

Figure 2. Time-series ((**a**–**g**), 1994–2006) for NDVI and correlation coefficient ($p < 0.05$) between yield and NDVI.

3.2. Fertilizer Factor Analysis between Yield and Fertilizer during 1994–2006

Figure 3 shows yield and fertilizer curves for different systems from 1994–2006. In the RRSAF, the high-yield plots were mostly distributed in the CMT; however, the change of yield from 1994 to 2006 was consistent with the application of top dressing mineral fertilizer (Figure 3b). As for LMT and OMT, WCC and compost were applied in RRSAF, and the yield trends were consistent with that of WCC for LMT and compost for WCC. Therefore, in this study, we selected the CMT mineral fertilizer; the WCC moisture for LMT and compost; and WCC for OMT to predict the maize yield.

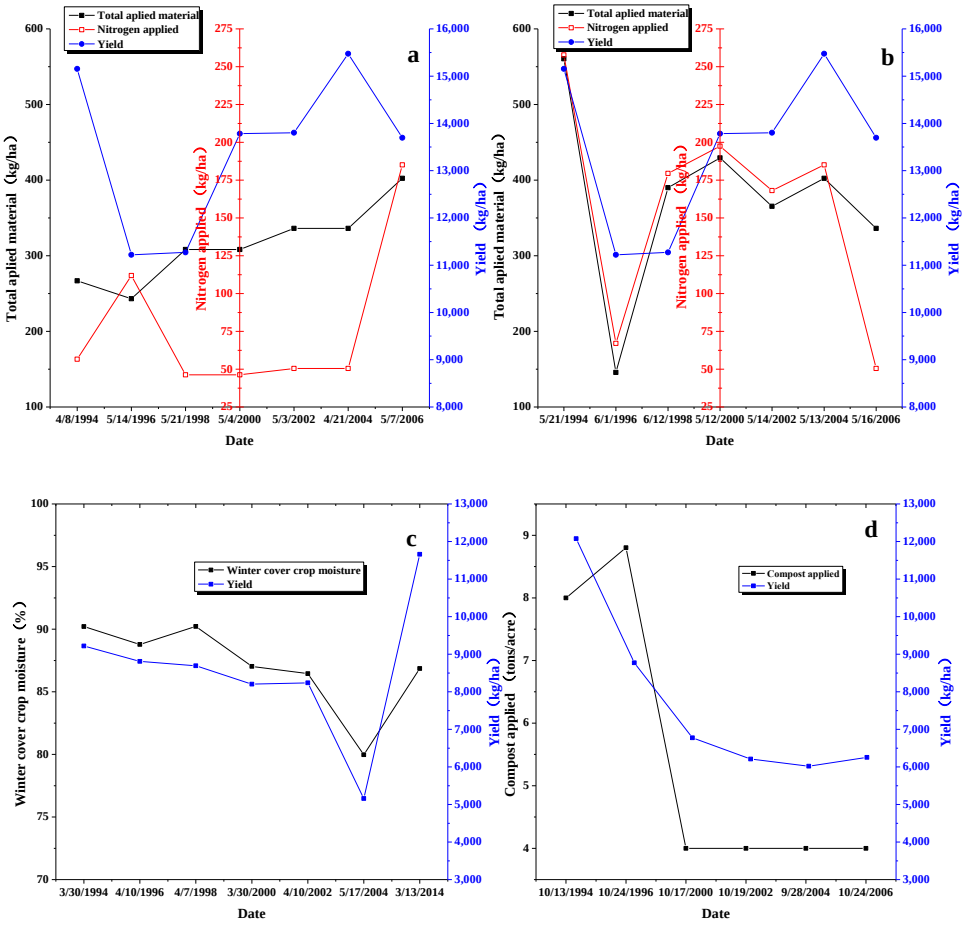

Figure 3. Yield and fertilizer curve for different systems from 1994–2006. (**a**,**b**), CMT: mineral fertilizer (total applied fertilizer and nitrogen applied); (**a**): base fertilizer; (**b**): top dressing. (**c**), LMT: winter cover crop (WCC) moisture. (**d**), OMT: compost applied.

3.3. The Performances of Multi-Models for Predicting Maize Yield

In this study, six ML models were trained with the measured yields, and 15 maize variables were measured at the plot level. The results were evaluated based on cross-validation and were summarized according to different models and different data combinations (Figure 4). Comprehensively considering the evaluation indicators (R^2, *RMSE*), RF, and AB models showed the highest accuracy, with higher R^2 (0.85~0.98) and lower *RMSE* (<1000 kg/ha). Although, the R^2 of the other models are above 0.65, all of their *RMSE*s were over 1100 kg/ha and even over 1900 kg/ha in some cases (VS for SVM), indicating an insignificant relationship between the predicted and the measured yields and larger errors. Thus, RF and AB are more suitable for maize yield prediction than the other algorithms in RRSAF. Moreover, we found that the accuracy varied by different data combinations even with the same machine learning algorithm, and VCS and VCSF achieved the highest accuracy. Finally, two algorithms (RF and AB) and two data combinations (VCS and VCFS) in this paper were selected to establish prediction models for maize yield at the plot level (Figure 5).

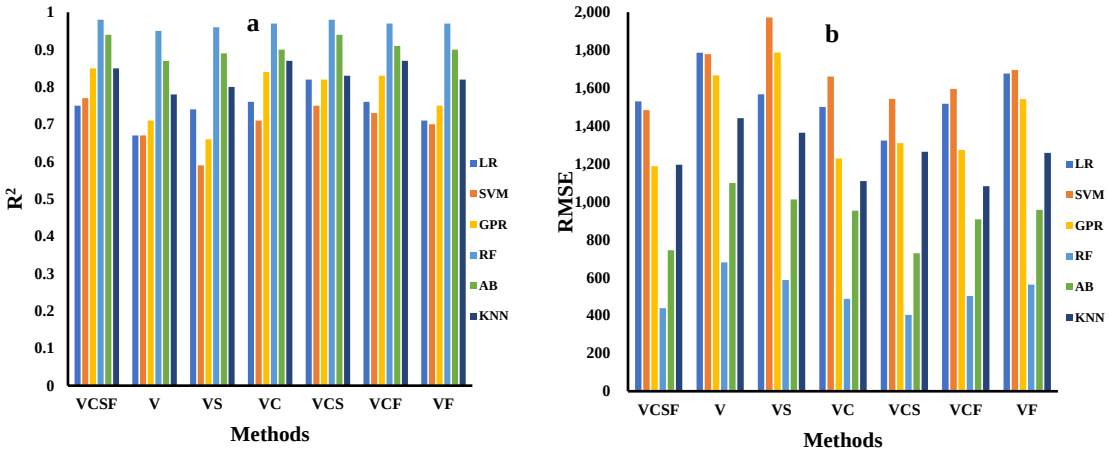

Figure 4. R^2 (**a**), *RMSE* (**b**) of six models for maize at the plot scale in different data combinations (V: VIs; VS: VIs + soil; VC: VIs + climate; VIs + climate + soil; VF: VIs + fertilizer; VCS: VIs + climate + soil; VCSF: VIs + climate + soil + fertilizer).

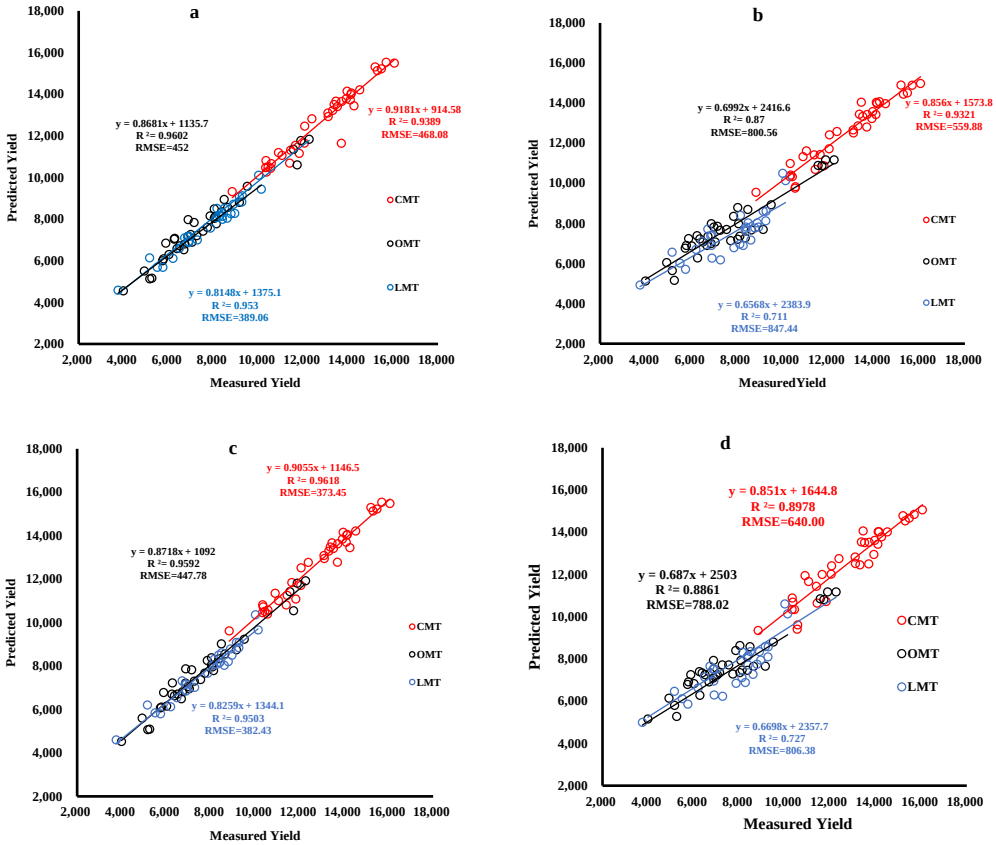

Figure 5. Scatter plots of measured yield and predicted yield of RF and AB models for different maize systems at the plot scale from 1994 to 2006 in even years. ((**a**), RF_VCSF; (**b**), AB_VCSF; (**c**), RF_VCS; (**d**), AB_VCS).

Based on the trained RF and AB models in last section, the RRSAF yields were predicted, and we forecasted the different systems separately. The scatter diagrams of the predicted and measured yields of the models in different systems from 1994 to 2006 in even years are shown in Figure 5. We found that the predicted and measured yields showed a good linear fit, with a R^2 of above 0.87. Such results indicated that the two ML models can predict the maize yield at the plot level with higher accuracy and RF > AB. Although all of the predicted yields were close to the measured yields, consistent underestimations were found for LMT and OMT and both VCS and VCSF for the two models, especially for the

AB model. In the next section analyzing the different systems, we used the RF model to predicted the maize yield for different systems in odd years.

Figure 6 shows the scatter plots of the measured yield and predicted yield of RF for different maize systems at the plot scale from 1995 to 2007 in odd years. It was found that the accuracy of the RF model is still very high when using the data from odd years.

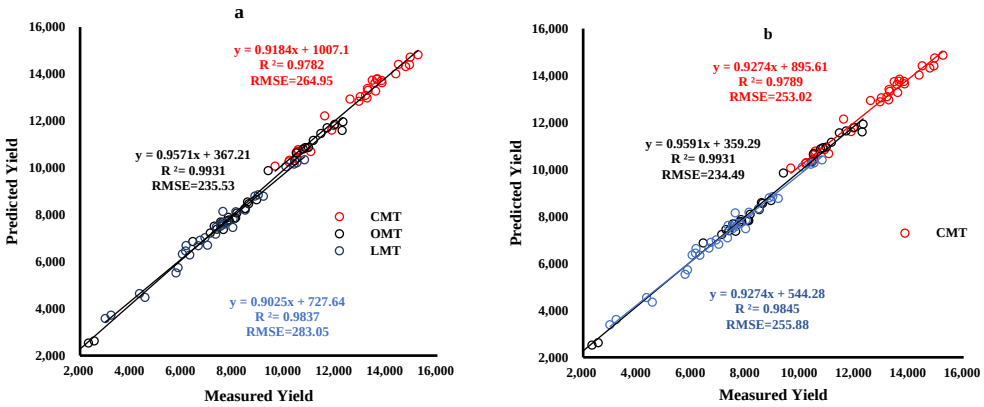

Figure 6. Scatter plots of measured yield and predicted yield of RF for different maize systems at the plot scale from 1995 to 2007 in odd years. ((**a**), RF_VCSF; (**b**), RF_VCS).

3.4. Comparison of Forecast Errors in Different Crop Systems

Crop yield is driven by the interaction between fertilizer management, soil, and weather conditions. Yield variation is not only from soil and climate data but also from fertilizer systems. To investigate the prediction errors comprehensively, we summarized them according to the systems and the RF model. The results showed that the errors of the RF models vary by the systems (CMT, LMT, OMT). The accuracy of the three fertilizer systems in the VCS is better than that in the VCSF at different degrees (Figures 5 and 6), and there is no significant difference between the VCS and VCSF in OMT. However, the LMT and CMT in VCS had better accuracy than the VCFS.

3.5. The Important Factors for Maize Yield Prediction in RRSAF

To identify the critical factors for maize yield prediction in RRSAF, we further analyzed the important orders of the 15 variables from the RF model. The importance of the prediction variables is as ordered: EVI > NDVI > EVI2> fertilizer > soil bulk density (0–25 cm) > soil bulk density (25–50 cm) > PRE > TMAX > SOM (0–25 cm) > TMIN > SOM (25–50 cm)> soil particle.

Distribution (Figure 7a). EVI is more important for maize yield prediction than NDVI in RRSAF, which is consistent with the study by Bolton et al. [31]. For the two variables related to temperature, TMAX contributed more significantly than TMIN to the accurate prediction of maize yield, implying that TMAX is of greater importance on maize yield. As for the three systems in RRSAF, CMT was the most affected by TMAX followed by VIs and fertilizer; LMT was the most affected by WCC followed by TMAX and VIs; and OMT was more affected by compost + WCC followed by precipitation and TMIN.

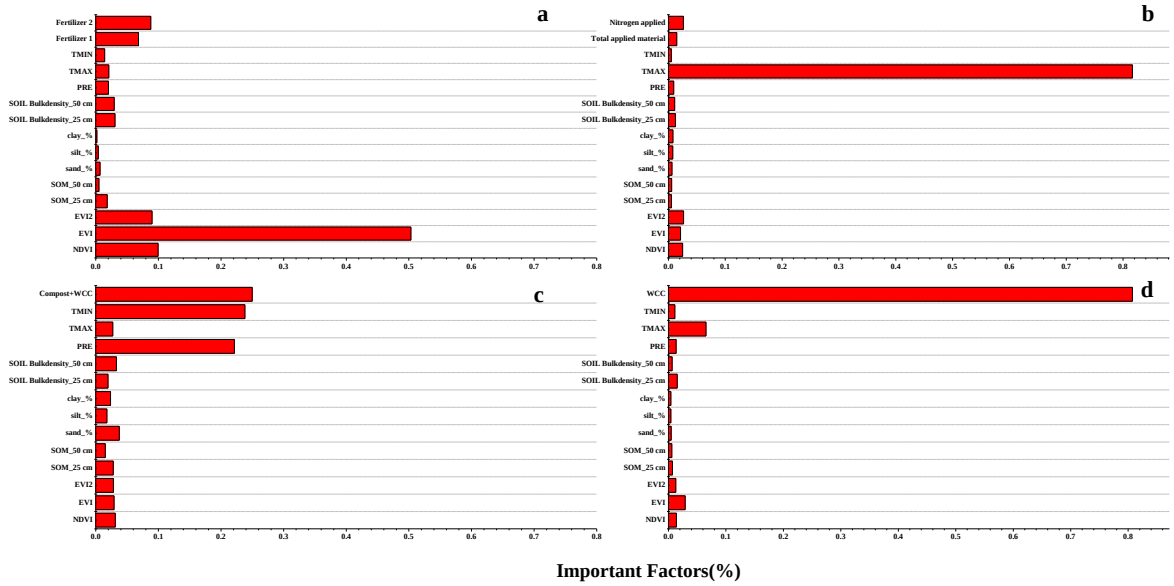

Figure 7. Feature importance values for the 15 variables from the random forest model in each system. ((**a**), all systems together: fertilizer 1 for total applied material + WCC+ Compost, fertilizer 2 nitrogen + WCC + compost; (**b**), CMT; (**c**), OMT; (**d**), LMT).

4. Discussion

4.1. Comparing the Performances of ML Models in Predicting Maize Yield

We proposed a framework for yield prediction through the development of a comprehensive yield prediction model driven by weather, remote sensing data, soil datasets, and fertilizer data. Additionally, we found that the non-linear ML methods (RF and AB) perform better overall than the traditional yield production method (Figure 4), and similar results have been reported in previous studies [32]. This could be explained by the non-linear relationships between variables and crop yields [19]. Moreover, ML methods have the advantages of computational efficiency and spatial generalizations relative to the deep learning network, which opens new prospects for crop yield prediction at a plot or even a larger scale. Additionally, the RF and AB models performed much better for VCS (VIs + climate + soil) and all combinations including fertilizer data (VCSF: VIs + climate + soil + fertilizer). It was found that including fertilizer information in the model does not produce outstandingly better predictions after modeling the three systems together, and the accuracy of the three systems is different for the different systems (Figures 5 and 6).

4.2. Model Performance for Different Systems and Feature Importance in Yield Prediction

The relative impact of a single variable cannot be quantified independently of other variables, but the RF method provides a measure for assessing the relative importance of variables to the prediction results [33]. The results in this study show that all variables with the exception of soil particle distribution and SOM (25–50 cm) are crucial for yield prediction (Figure 7a). Additionally, previous studies have also emphasized the importance of using a vegetation index, precipitation, and temperature in predicting crop yield [34]. However, the use of Vis in this paper to predict the maize yield alone limited yield accuracy (Figure 4), which is inconsistent with a previous study that showed that the VIs achieved a lower R^2, which was below 0.5, for yield estimation [10]. Moreover, the contributions to maize yield estimation of other critical factors, such as soil factors, were identified and isolated. Those factors contribute to better explanations of yield variability, proving the hypothesis from Guan et al. [34].

In this study, three systems in RRSAF were separated, namely "CMT", "LMT", and "OMT". We also found that the accuracy of yield prediction varied across different systems (Figure 6). The importance of VIs is weaker than putting the three data systems together to

evaluate the different farming systems (Figure 7a). Fertilizer and climate factors can provide more information for more accurate yield prediction (Figure 7b–d). Additionally, we found that the relative importance of TMAX is greater than other variables (Figure 7b) in CMT. Within a certain fertilizer and range of temperatures, an increase of temperature enhances respiration in crops, and the energy that is generated increases, as is the absorption of nutrients and the synthesis and accumulation of organic nutrients. It was also found that crop yields increase with an increase of TMIN and PRE in OMT, which is mainly due to the increasing ground humidity, which benefits the effect of organic fertilizer. LMT was the most affected by WCC in terms of soil nutrients and microorganisms; the decomposition of WCC can not only increase the content of organic matter and activate soil nutrients but can also provide nutrients for later crops to absorb [13].

4.3. Uncertainties in the Study

This study faced several uncertainties. The first limitation is that using the Mediterranean climate region as a study area could lead to errors when it is applied to other climatic regions due to its winter dominated precipitation regime. Another concern is that VIs suffers from the coarse spatiotemporal resolution of Landsat TM 5 at the plot scale in the RRSAF. With the more recent availability of the Sentinel-2 satellites, higher spatial and temporal resolution data are expected to be available in upcoming studies. In addition, the current study focused on predicting plot-level crop yield because the yield data were recorded and are only available for the whole plot or for two yield data points per plot. Additionally, we can obtain newer data each year. In future research, new data will be added to study and yield model correction. Large-scale maize yield prediction needs further verification when we extract the area for the input variables. In contrast, we should note that no mechanistic processes for crop growth were included in the ML models due to their internal black box structures, which the prevented examination of their physiological processes. Although ML is able to identify RRSAF in this paper efficiently, including unknown processes will inevitably increase the uncertainty of model performance. Alternatively, crop growth (DSSAT, WOFOST) models are developed and validated by many experts over decades of research [35]. Crop models have characterized the internal growth and development mechanism of major crops to some degree, and these have been widely applied with higher accuracy in many regions. Thus, combining ML with crop growth models is an idea for future studies on yield prediction.

5. Conclusions

We predicted maize yield at the plot scale based on multi-source data and multiple machine learning models at the plot scale in RRSAF. It was found that RF and AB predicted maize yields with the highest accuracy, and RF demonstrated the best generalization ability among the methods. The RF model can estimate maize yields accurately in advance (before the harvesting dates) in RRSAF. It was found that fertilizer information does not produce outstandingly better predictions, and the accuracy is different for different systems. Additionally, this paper has its limitations, including the restricted access and locality of the study region. However, our study also highlights the necessity of integrating multi-spectral satellite data and environmental variables for predicting crop yield.

Author Contributions: Data curation, L.M. and X.Z.; formal analysis, L.M. and X.Z.; funding acquisition, H.L.; methodology, L.M.; resources, S.L.U.; software, L.M.; writing—original draft, L.M.; writing—review and editing, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (2017YFD0201803) and the National Natural Science Foundation of China (41671438).

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

Appendix A

Input Data 1				Input Data 2	Out Data
All Data	**CMT**	**LMT**	**OMT**		
NDVI EVI EVI2 SOM_25 SOM_50 sand_% silt_% clay_% SOIL Bulk-density_25 SOIL Bulk-density_50 Precipition Tmax Tmin	NDVI EVI EVI2 SOM_25 SOM_50 sand_% silt_% clay_% SOIL Bulk-density_25 SOIL Bulk-density_50 Precipition Tmax Tmin	NDVI EVI EVI2 SOM_25 SOM_50 sand_% silt_% clay_% SOIL Bulk-density_25 SOIL Bulk-density_50 Precipition Tmax Tmin	NDVI EVI EVI2 SOM_25 SOM_50 sand_% silt_% clay_% SOIL Bulk-density_25 SOIL Bulk-density_50 Precipition Tmax Tmin	Measured Yield	Predicted Yield
Fertilizer 1: for total applied material+ WCC + Compost	Total applied material	WCC	WCC + Compost		
Fertilizer 2: WCC + Compost	Nitrogen				

References

1. Cole, M.B.; Augustin, M.A.; Robertson, M.; Manners, J.M. The science of food security. *NPJ Sci. Food* **2018**, *2*, 1–8. [CrossRef]
2. 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]
3. Azzari, G.; Jain, M.; Lobell, D. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. *Remote. Sens. Environ.* **2017**, *202*, 129–141. [CrossRef]
4. Meng, L.; Liu, H.; Zhang, X.; Ren, C.; Ustin, S.; Qiu, Z.; Xu, M.; Guo, D. Assessment of the effectiveness of spatiotemporal fusion of multi-source satellite images for cotton yield estimation. *Comput. Electron. Agric.* **2019**, *162*, 44–52. [CrossRef]
5. Johnson, D.M. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote. Sens. Environ.* **2014**, *141*, 116–128. [CrossRef]
6. Liu, H.; Meng, L.; Zhang, X.; Susan, U.; Ning, D.; Sun, S. Estimation model of cotton yield with time series Landsat images. *Trans. Chin. Soc. Agric. Eng.* **2015**, *31*, 215–220.
7. Pede, T.; Mountrakis, G.; Shaw, S.B. Improving corn yield prediction across the US Corn Belt by replacing air temperature with daily MODIS land surface temperature. *Agric. For. Meteorol.* **2019**, *276*, 107615. [CrossRef]
8. Chen, Y.; Zhang, Z.; Tao, F.; Wang, P.; Wei, X. Spatio-Temporal patterns of winter wheat yield potential and yield gap during the past three decades in North China. *Field Crop. Res.* **2017**, *206*, 11–20. [CrossRef]
9. 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. Clim.* **2015**, *123*, 291–302. [CrossRef]
10. Cao, J.; Zhang, Z.; Tao, F.; Zhang, L.; Luo, Y.; Han, J.; Li, Z. Identifying the Contributions of Multi-Source Data for Winter Wheat Yield Prediction in China. *Remote. Sens.* **2020**, *12*, 750. [CrossRef]
11. Mueller, N.D.; Gerber, J.; Johnston, M.; Ray, D.; Ramankutty, N.; Foley, J.A. Closing yield gaps through nutrient and water management. *Nature* **2012**, *490*, 254–257. [CrossRef] [PubMed]
12. Huanjun, L.; Danqian, W.; Linghua, M.; Ustin, S.; Yang, C.; Haoxuan, Y.; Xinle, Z. Remote sensing recognition method of different fertilization methods in NDVI time series. *Trans. Chin. Soc. Agric. Eng.* **2019**, *35*, 162–168.
13. Lee, J.H.; Shin, J.; Realff, M.J. Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Comput. Chem. Eng.* **2018**, *114*, 111–121. [CrossRef]
14. Sharma, N.; Sharma, R.; Jindal, N. Machine Learning and Deep Learning Applications-A Vision. *Glob. Transit. Proc.* **2021**, *2*, 24–28. [CrossRef]

15. Nawar, S.; Buddenbaum, H.; Hill, J.; Kozak, J.; Mouazen, A.M. Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. *Soil Tillage Res.* **2016**, *155*, 510–522. [CrossRef]
16. Knox, N.M.; Grunwald, S.; McDowell, M.L.; Bruland, G.L.; Myers, D.B.; Harris, W.G. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma* **2015**, *239*, 229–239. [CrossRef]
17. Rossel, R.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [CrossRef]
18. Hunt, M.L.; Blackburn, G.A.; Carrasco, L.; Redhead, J.W.; Rowland, C.S. High resolution wheat yield mapping using Sentinel-2. *Remote. Sens. Environ.* **2019**, *233*, 111410. [CrossRef]
19. Cai, Y.; Guan, K.; Lobell, D.; Potgieter, A.; 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]
20. Denison, R.F.; Bryant, D.C.; Kearney, T.E. Crop yields over the first nine years of LTRAS, a long-term comparison of field crop systems in a Mediterranean climate. *Field Crop. Res.* **2004**, *86*, 267–277. [CrossRef]
21. Wolf, K.M.; Torbert, E.E.; Bryant, D.; Burger, M.; Denison, R.F.; Herrera, I.; Hopmans, J.; Horwath, W.; Kaffka, S.; Kong, A.Y.Y.; et al. The century experiment: The first twenty years of UC Davis' Mediterranean agroecological experiment. *Ecology* **2018**, *99*, 503. [CrossRef] [PubMed]
22. Fortes Gallego, R.; Prieto Losada MD, H.; García Martín, A.; Córdoba Pérez, A.; Martínez, L.; Campillo Torres, C. Using NDVI and guided sampling to develop yield prediction maps of processing tomato crop. *Span. J. Agric. Res.* **2015**, 13. [CrossRef]
23. Corani, G.; Benavoli, A. A Bayesian approach for comparing cross-validated algorithms on multiple data sets. *Mach. Learn.* **2015**, *100*, 285–304. [CrossRef]
24. Alberto, G.S.; Juan, F.S.; Waldo, O.B. Attribute Selection Impact on Linear and Nonlinear Regression Models for Crop Yield Prediction. *Sci. World J.* **2014**, *2014*, 1–10.
25. Middleton, M.; Närhi, P.; Arkimaa, H.; Hyvönen, E.; Kuosmanen, V.; Treitz, P.; Sutinen, R. Ordination and hyperspectral remote sensing approach to classify peatland biotopes along soil moisture and fertility gradients. *Remote. Sens. Environ.* **2012**, *124*, 596–609. [CrossRef]
26. Nguyen-Tuong, D.; Seeger, M.; Peters, J. Model Learning with Local Gaussian Process Regression. *Adv. Robot.* **2009**, *23*, 2015–2034. [CrossRef]
27. Appelhans, T.; Mwangomo, E.; Hardy, D.R.; Hemp, A.; Nauss, T. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. *Spat. Stat.* **2015**, *14*, 91–113. [CrossRef]
28. Liu, Y.; Yang, G.; Li, S. Application of BP-AdaBoost model in temperature compensation for fiber optic gyroscope bias. *Beijing Univ. Aeronaut. Astronaut.* **2014**, *40*, 235–239.
29. 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]
30. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
31. Bolton, D.K.; Friedl, M.A. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agric. For. Meteorol.* **2013**, *173*, 74–84. [CrossRef]
32. Han, J.; Zhang, Z.; Cao, J.; Luo, Y.; Zhang, L.; Li, Z.; Zhang, J. Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China. *Remote. Sens.* **2020**, *12*, 236. [CrossRef]
33. Saeed, U.; Dempewolf, J.; Becker-Reshef, I.; Khan, A.; Ahmad, A.; Wajid, S.A. Forecasting wheat yield from weather data and MODIS NDVI using Random Forests for Punjab province, Pakistan. *Int. J. Remote. Sens.* **2017**, *38*, 4831–4854. [CrossRef]
34. Guan, K.; Wu, J.; Kimball, J.S.; Anderson, M.C.; Frolking, S.; Li, B.; Hain, C.R.; Lobell, D. 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]
35. Kasampalis, D.A.; Alexandridis, T.K.; Deva, C.; Challinor, A.; Moshou, D.; Zalidis, G. Contribution of Remote Sensing on Crop Models: A Review. *J. Imaging* **2018**, *4*, 52. [CrossRef]