Article

Improving the Forecasting of Winter Wheat Yields in Northern China with Machine Learning–Dynamical Hybrid Subseasonal-to-Seasonal Ensemble Prediction

Junjun Cao 1,2, Huijing Wang 1, Jinxiao Li 3,*, Qun Tian 4 and Dev Niyogi 5,6,7

1 Key Laboratory of Geographical Process Analysis & Simulation of Hubei Province, College of Urban and Environmental Sciences, Central China Normal University, Wuhan 430079, China; caojun@ccnu.edu.cn (J.C.); wanghai@mails.ccnu.edu.cn (H.W.)
2 Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China
3 State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China
4 Guangdong Provincial Key Laboratory of Regional Numerical Weather Prediction, Guangzhou Institute of Tropical and Marine Meteorology, CMA, Guangzhou 510641, China; tianq@gd121.cn
5 Department of Geological Sciences, Jackson School of Geosciences, University of Texas at Austin, Austin, TX 78712, USA; happy1@utexas.edu
6 Department of Civil, Architecture, and Environmental Engineering, University of Texas at Austin, Austin, TX 78712, USA
7 Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA

*Correspondence: lijinxi@lasg.iap.ac.cn

Abstract: Subseasonal-to-seasonal (S2S) prediction of winter wheat yields is crucial for farmers and decision-makers to reduce yield losses and ensure food security. Recently, numerous researchers have utilized machine learning (ML) methods to predict crop yield, using observational climate variables and satellite data. Meanwhile, some studies also illustrated the potential of state-of-the-art dynamical atmospheric prediction in crop yield forecasting. However, the potential of coupling both methods has not been fully explored. Herein, we aimed to establish a skilled ML–dynamical hybrid model for crop yield forecasting (MHCF v1.0), which hybridizes ML and a global dynamical atmospheric prediction system, and applied it to northern China at the S2S time scale. In this study, we adopted three mainstream machine learning algorithms (XGBoost, RF, and SVR) and the multiple linear regression (MLR) model, and three major datasets, including satellite data from MOD13C1, observational climate data from CRU, and S2S atmospheric prediction data from IAP CAS, used to predict winter wheat yield from 2005 to 2014, at the grid level. We found that, among the four models examined in this work, XGBoost reached the highest skill with the S2S prediction as inputs, scoring $R^2$ of 0.85 and RMSE of 0.78 t/ha 3–4 months, leading the winter wheat harvest. Moreover, the results demonstrated that crop yield forecasting with S2S dynamical predictions generally outperforms that with observational climate data. Our findings highlighted that the coupling of ML and S2S dynamical atmospheric prediction provided a useful tool for yield forecasting, which could guide agricultural practices, policy-making and agricultural insurance.

Keywords: yield forecasting; climate variables; subseasonal-to-seasonal prediction; machine learning; winter wheat

1. Introduction

Growing populations and an increasing frequency of extreme weather events are threatening wheat production and bringing great challenges to global food security [1–5]. Besides, timely crop yield forecasting on subseasonal-to-seasonal (S2S) time scales (2 weeks to 3 months) is of great interest to agricultural production, decision-making, market futures and the insurance industry [6–8]. Wheat is the world’s most widely distributed cereal,
with the largest planting area [9]—especially for China, the world’s largest producer and consumer of wheat, where the planting area and yield of winter wheat account for about 94% and 95% of the total planting area and wheat yield, respectively [10]. Hence, this study uses, as an example, the winter-wheat-planting regions of northern China.

Numerous studies have been carried out on crop yield prediction, based on either crop growth modeling or statistical regression. Crop growth modeling aims to reproduce the key processes of plant growth and development in detail, from daily meteorological data, cultivar features, soil properties and agro-management information [11]. As such, crop growth models are crucial for providing farmers or specialists with real-time information about their crops, giving risk-assessment information, and monitoring decision-making, relevant to agricultural management, by quantifying the impact of weather, soil, and management interaction on crops [12]. However, the high computational costs and data requirements involved hinder scaling the approach to multiple crops and regions [13,14]. Traditional linear regression models are based on the empirical relationships between historical yields and other factors, such as climate variables, agrometeorological factors, and/or remote sensing data. However, as this method is unable to consider dynamical meteorological factors with the changing of growth stages, it has limited ability to disentangle the complex nonlinear relationships between independent variables and yields [15–19]. By contrast, machine learning (ML) algorithms—more advanced regression methods and more popular in agricultural production—use data or experience to improve the performance of specific algorithms [20], particularly by explaining a higher-order and nonlinear relationship. Increasingly, ML has been used for agricultural applications, such as crop type classification and crop yield prediction [21], and is becoming an indispensable and mainstream tool in precision agriculture.

Climate variables, such as temperature and precipitation, have been confirmed to have significant impacts on crop production, and explain approximately one-third of global crop yield variability [2,22,23]. Remote sensing data enable the rapid monitoring and forecasting of agricultural information, such as crop growth and grain yield [24–26], to be achieved on a large scale. Therefore, the majority of previous studies have used observational climate data and remote sensing data. Recently, some studies have illustrated that dynamical atmospheric prediction is more advantageous than observational climate data in predicting crop yields, which raises the idea that seasonal agricultural production forecasting could benefit directly from dynamical atmospheric prediction [7,27,28].

S2S dynamical atmospheric prediction, which aims to bridge the gap between medium-range weather forecasts and seasonal prediction, is an emerging and fast-developing field. The science community has rallied under the World Weather Research Programme–World Climate Research Programme S2S Prediction Project, dedicated to improving the forecasting skill and understanding of the sources of S2S predictability. Substantial progress has been made recently on predicting the onset, evolution, and decay of some large-scale extreme events, such as heatwaves and tropical cyclones [29]. Considering the potential significant advantages of S2S atmospheric prediction paired with ML, to the best of our knowledge, no study has combined these two methods in a hybrid approach to predict crop yields. In the present study, we seek to address this knowledge gap.

Specifically, we use S2S dynamical prediction system outputs and a variety of algorithms to build an ML–dynamical hybrid model for crop yield forecasting (MHCF v1.0). The motivations behind this study are to (1) compare the performance in crop yield forecasting based on observed meteorological data/remote sensing data and S2S atmospheric prediction system outputs; (2) investigate the potential of various algorithms for S2S crop yield forecasting; (3) evaluate how early MHCF v1.0 can forecast winter wheat yield with reasonable accuracy. Section 2 describes the data and methods used in this study, respectively. Section 3 presents results from the comparison of different data sources on the model accuracy and comparison of machine learning models, and reports findings on the spatial distribution of yield forecasting and the optimum lead time. Sections 4 and 5 provide some further discussion and conclusions, respectively.
2. Materials and Methods

2.1. Study Area

The study area is the main winter-wheat-producing area in northern China with a temperate continental monsoon climate, including most areas of Hebei, Henan, Shandong province, and southwestern Shanxi province, central Shaanxi province (Figure 1). The main prevailing rotation mode is winter wheat and summer corn in this region. In general, the growing season of winter wheat starts in October and ends in June the following year [30].

Figure 1. The spatial distribution of winter-wheat-planting areas in Northern China.

2.2. Data and Preprocessing

Four types of data were used in this study: crop yield data, satellite data, observational climate data, and S2S atmospheric prediction data. The detailed description and sources of the datasets are shown in Tables 1 and 2. Firstly, all variables were aggregated into 0.5° spatial resolution and monthly temporal resolution. Then all variables were further masked based on the winter-wheat-planting areas. The sample size used in the study comprised 219 grids × 10 years × 9 months (19,710) in total.

Table 1. Summary and abbreviations of input data.

| Category   | Abbreviation | Description                                                                 |
|------------|--------------|-----------------------------------------------------------------------------|
| Crop data  | Yield        | Winter wheat yield                                                          |
| Satellite data | EVI       | Enhanced vegetation index                                                   |
| Climate data | Tmx        | Maximum temperatures                                                        |
|            | Tmn         | Minimum temperatures                                                        |
|            | Tmp         | Mean temperatures                                                           |
|            | GDD         | Growing degree days                                                         |
|            | Pre         | Precipitation                                                               |
|            | SPEI        | Standardized precipitation evapotranspiration                               |
|            | VPD         | Vapor pressure deficit                                                       |
| S2S data   | temp925     | 925 hPa Air Temperature in K                                                |
|            | u925        | 925 hPa eastward wind in m/s                                                |
|            | v925        | 925 hPa northward wind m/s                                                  |
|            | humdy925    | 925 hPa specific humidity in kg/kg                                          |
|            | skt         | ground temperature in K (equal to skin temperature)                        |
|            | t2m         | surface (2 m) Air Temperature in K                                          |
|            | prec        | total precipitation rate in mm/h                                            |
|            | radia       | surface net shortwave radiation in W/m² (positive downward)                 |
Table 2. Detailed information of the datasets used for winter wheat yield prediction in China.

| Category   | Variables | Resolution   | Time Range | Data Source            |
|------------|-----------|--------------|------------|------------------------|
| Crop data  | Yield     | County       | Yearly     | 2004–2014              | Ministry of Agriculture of China |
| Satellite data | EVI      | 0.05°        | 16-day     | 2004–2014              | MOD13C1                        |
| Climate data | Tmx Tmn Tmp GDD Pre SPEI VPD | 0.5°         | Monthly    | 2004–2014              | CRU                            |
| S2S data   | temp925 u925 v925 humid925 skt t2m prec radia | 0.5°         | Monthly    | 2004–2014              | FGOALS-f2                      |

2.2.1. Cropland and Winter Wheat Yield Data

We collected county-level winter wheat yield data (t/ha) in the winter-wheat-producing regions of northern China from 2005 to 2014, which were gathered by the Agricultural Statistical Yearbook of the Ministry of Agriculture of China. The winter-wheat-planting area of 2014 was used to mask the winter wheat yields with a resolution of 1 km [31]. Overall, we selected 393 counties across the winter-wheat-planting areas in China. To match the spatial scale of other variables, the county-level winter wheat yields were assigned to a 0.5° grid through weighted averaging.

2.2.2. Satellite Data

Enhanced Vegetation Index (EVI) has been proven to be superior in crop yield prediction to Normalized Difference Vegetation Index (NDVI), which is more sensitive to higher canopy Leaf Area Index [15,32–35]. Thus, we chose EVI as the satellite indicator, which was derived from the MOD13C1 (Collection 6) product with a 16-day repeat and 0.05° spatial resolution [22]. Finally, the EVI was resampled by the MVC (maximum synthesis method) to have 0.5° spatial resolution and monthly time steps.

2.2.3. Observational Climate Data

Observational climate data were obtained from the Climatic Research Unit (CRU), including monthly maximum temperature, minimum temperature, mean temperature, precipitation, vapor pressure deficit (VPD), and growing degree day (GDD). One other variable was the Standardized Precipitation–Evapotranspiration Index (SPEI) [36]. VPD and GDD were calculated from the CRU variables using the following formulas [22].

\[

day \text{VPD} = e_{\text{sat}} - Vap \\
\text{e}_{\text{sat}} = 6.108 \times e^{\left(\frac{17.27 \times T}{237.3 + T}\right)}
\]

\[
GDD = \frac{T_{\text{mx}} + T_{\text{mn}}}{2} - T_{\text{base}}
\]

\[
T_{\text{mx}} = \begin{cases} 
T_{\text{mx}}, & 25\degree C > T_{\text{mx}} > T_{\text{base}} \\
25\degree C, & T_{\text{mx}} \geq 25\degree C \\
T_{\text{base}}, & T_{\text{mx}} \leq T_{\text{base}}
\end{cases}
\]

\[
(1)
\]

\[
(2)
\]
\[
T_{mn} = \begin{cases} 
T_{mn}, & 25^\circ C > T_{mn} > T_{base} \\
25^\circ C, & T_{mn} \geq 25^\circ C \\
T_{base}, & T_{mn} \leq T_{base}
\end{cases}
\]

Vap is the vapor pressure in CRU, \( e_{sat} \) is the saturated vapor pressure (in hPa) at mean air temperature (\( T_{mp} \)). \( T_{mx} \), \( T_{mn} \) are the monthly maximum temperature and minimum temperatures respectively, and \( T_{base} \) is the lower temperature limits (\(^\circ C\)). The lower temperature for winter wheat is 0 \(^\circ C\).

2.2.4. S2S Climate Prediction Data

The S2S atmospheric prediction outputs were obtained from the IAP-CAS FGOALS-f2 dynamical forecasting system, which has been applied at China National Climate Center for real-time S2S prediction [37–39]. The prediction model is a Climate System Model representing the interaction between the atmosphere, oceans, land, and sea ice, which uses the time-lag perturbation method to generate 35 ensemble samples, and rolls on the 20th of each month to predict the climate conditions in the next 6 months [40]. The prediction products include the monthly and daily average and 6-h average forecast data of the next 6 months of the 4 modules of the atmosphere, ocean, land, and sea ice [37,38]. Studies have shown that FGOALS-f2 is skilled in predicting extreme events, such as summer drought and tropical cyclones genesis, which inevitably affect crop growth and factual yield [41,42]. In this study, we used the monthly atmospheric prediction outputs of FGOALS-f2 with a 0.5 \(^\circ\) spatial resolution for the next whole month to forecast the winter wheat yield, including 925-hPa air temperature in K, 925-hPa eastward wind in m/s, 925-hPa northward wind in m/s, 925-hPa specific humidity in kg/kg, ground temperature in K, surface (2-m) air temperature in K, total precipitation rate in mm/h, and surface net shortwave radiation in W/m\(^2\).

2.3. Model Development

Four ML methods (MLR, SVR, RF, XGBoost) were adopted to establish prediction models between input variables and winter wheat yield. Before applying the models, the data preprocessing was carried out. Firstly, we randomly divided the whole dataset into 70% training data and 30% testing data [33,43,44], and then preprocessed the training data and testing data separately. This is because the testing data are often unavailable in real life, otherwise, it is equivalent to telling the model the answer to a part of the prediction in advance. Secondly, to have a mean of 0 and a standard deviation of 1, the training data and testing data were normalized respectively by the Z-score [22]. Next, the best hyper-parameters for each model were determined by the five-fold cross-validation using the GridSearchCV package [45,46]. Finally, the optimal models were determined based on the tuned hyperparameters, and the predicted \( R^2 \) was calculated on the testing dataset.

2.3.1. Multiple Linear Regression (MLR)

Multiple linear regression (MLR) is the most commonly used linear regression method that utilizes multiple explanatory variables to predict the response variable [44,47]. The goal of MLR is to establish a linear relationship between the independent variable and the dependent variable. Essentially, MLR is an extension of ordinary least squares, which involves more than one explanatory variable [48]. As mentioned above, the factors affecting crop yield are complex and interrelated. Therefore, it is reasonable for our study to choose MLR instead of simple linear regression.

2.3.2. Support Vector Machine Regression (SVR)

Support vector machine regression (SVR) is a model derived from Support Vector Machine (SVM). Similar to SVM, SVR finds a regression plane to minimize the distance of all inputs to this hyperplane. Generally speaking, SVR requires a kernel function to map all inputs from the original space to a high-dimensional space. Then, a linear function is constructed in this feature space to balance error minimization and overfitting [49,50].
The most commonly used kernel functions include linear kernels, polynomial kernels, and Gaussian radial basis kernels. In addition, the hyperparameters that need tuning are the penalty coefficient C and the kernel coefficient gamma.

2.3.3. Random Forest (RF)

The RF model is an ensemble learning method for both regression and classification, which was introduced by Breiman [51]. RF is composed of many decision trees. In the training phase, the input training dataset is divided into multiple different sub-training datasets using bootstrap sampling, and then each sub-training set is used to generate a single decision tree. After training each single decision tree, it gives a prediction. Finally, predictions of each decision tree are averaged to provide the final prediction [47,52]. Moreover, RF is more robust and insensitive to noise. The number of decision trees (n_estimators), the maximum depth (max_depth) are the need-to-be-tuned hyper-parameters in our study.

2.3.4. EXtreme Gradient Boost (XGBoost)

XGBoost is a scalable machine learning algorithm for tree boosting proposed by Chen and Guestrin [53]. The basic principle of the XGBoost algorithm is to build multiple weak learners on the full data and aggregate the modeling results of all weak learners to obtain better regression or classification performance. XGBoost incorporates a regularized model to prevent overfitting and the weak learner can be a regression tree or a linear model [53–55]. The XGBoost model in this research is established based on the decision tree. The predicted values in XGBoost are obtained by directly summing the leaf weights on all decision trees. The optimal parameters of the XGBoost algorithm are obtained using the GridSearchCV package [45,46].

2.4. Model Evaluation

We conducted leave-one-year-out cross-validation to assess the practicality of the models, i.e., using all years’ data from 2005 to 2014 except the target year to train the model and then make a prediction for the target year [28]. This approach is an extensively used cross-validation method because of its simplicity, universality, and superiority in avoiding the issue of over-fitting [28,33,56]. To evaluate the model performance, the coefficient of determination (R²), root-mean-square error (RMSE), and percent error (PE) were selected as the evaluation metrics in this paper. The RMSE and PE are calculated by Formulas (3) and (4).

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}
\]  

(3)

\[
PE = \frac{|y_i - \hat{y}_i|}{\hat{y}_i} \times 100\%
\]  

(4)

\( y_i \) is the predicted yield, \( \hat{y}_i \) is the observed yield, and \( n \) is the number of samples.

2.5. Experimental Design

Three experiments were designed to address the aims of our study as outlined in the introduction. Firstly, we trained crop yield models with observational climate data and S2S atmospheric prediction outputs separately to identify which data source is superior for crop yield forecasting. Secondly, we compared several yield prediction models with different algorithms to find the model with the highest accuracy. Thirdly, we performed in-season prediction to assess the lead time that can reasonably predict the winter wheat yield. The in-season prediction started at the beginning of the growing season and ended one month before harvest. The flow diagram of the yield modeling steps is presented in Figure 2.
Remote Sens. 2022, 14, x FOR PEER REVIEW 7 of 18

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{\text{pred}} - y_{\text{obs}})^2} \] (3)

\[ PE = \left| \frac{y_{\text{pred}} - y_{\text{obs}}}{y_{\text{obs}}} \right| \times 100\% \] (4)

\( y_{\text{pred}} \) is the predicted yield, \( y_{\text{obs}} \) is the observed yield, and \( n \) is the number of samples.

2.5. Experimental Design

Three experiments were designed to address the aims of our study as outlined in the introduction. Firstly, we trained crop yield models with observational climate data and S2S atmospheric prediction outputs separately to identify which data source is superior for crop yield forecasting. Secondly, we compared several yield prediction models with different algorithms to find the model with the highest accuracy. Thirdly, we performed in-season prediction to assess the lead time that can reasonably predict the winter wheat yield. The in-season prediction started at the beginning of the growing season and ended one month before harvest. The flow diagram of the yield modeling steps is presented in Figure 2.

Figure 2. A flow diagram of the yield modeling steps in this study.

3. Results

3.1. Comparison of Different Data Sources on Model Accuracy

To compare the performance of observational climate data with S2S dynamical atmospheric prediction outputs in yield forecasting, three groups of data were integrated into the selected models—namely, S2S dynamical atmospheric prediction outputs (S2S-based), observational climate data (Observation-based), and the combination of the two (S2S+Observation-based). As mentioned earlier, the hybrid of S2S atmospheric prediction outputs and ML have been called MHCF v1.0, while the combination of observational climate data and ML are called the basic model, i.e., benchmark. We found that the performance of the S2S-based model was better than the observation-based model in all selected ML models, but not for MLR (Figure 3). S2S-based alone outperformed the observation-based, as well as the combination of both, among the ML-based models, with an \( R^2 \) of 0.85, 0.84, and 0.76 in XGBoost, RF, and SVR (RMSEs of 0.78 t/ha, 0.82 t/ha, and 0.99 t/ha), respectively. Meanwhile, the prediction \( R^2 \) values of the observation-based models were 0.81, 0.8 and 0.65 in XGBoost, RF and SVR (RMSEs were 0.89 t/ha, 0.92 t/ha and 1.21 t/ha), respectively.

Additionally, we noticed that the S2S+observation-based model performed almost equivalently to the S2S-based model alone for XGBoost and RF, and worse for SVR, indicating that S2S atmospheric prediction outputs plus observational climate variables as model inputs do not add extra contributions. This phenomenon may largely be because the S2S atmospheric prediction outputs are generated based on a coupled prediction system of the atmosphere, ocean, land, and sea ice, which is different from the observational climate data that are collected based on the conditions of the historical atmosphere. Therefore, the overlapping of variables from different systems will result in them restricting each other and will ultimately worsen the model performance for the ML methods.
observation-based, as well as the combination of both, among the ML-based models, with an R² of 0.85, 0.84, and 0.76 in XGBoost, RF, and SVR (RMSEs of 0.78 t/ha, 0.82 t/ha, and 0.99 t/ha), respectively. Meanwhile, the prediction R² values of the observation-based models were 0.81, 0.80, and 0.65 in XGBoost, RF, and SVR (RMSEs were 0.89 t/ha, 0.92 t/ha, and 1.21 t/ha), respectively.

Figure 3. Model performance of the three groups of variables. (a) the prediction R² of the four models (XGBoost, RF, SVR, and MLR); (b) the RMSE (unit: t/ha) of the predicted yield.

Additionally, we noticed that the S2S+observation-based model performed almost equivalently to the S2S-based model alone for XGBoost and RF, and worse for SVR, indicating that S2S atmospheric prediction outputs plus observational climate variables as model inputs do not add extra contributions. This phenomenon may largely be because the S2S atmospheric prediction outputs are generated based on a coupled prediction system of the atmosphere, ocean, land, and sea ice, which is different from the observational climate data that are collected based on the conditions of the historical atmosphere. Therefore, the overlapping of variables from different systems will result in them restricting each other and will ultimately worsen the model performance for the ML methods.

We also evaluated the contribution of EVI to the candidate models (Figure 4). As shown in Figures 4 and 5, whether for R² or RMSE, the inclusion or not of EVI did not significantly change the difference between the two models of MHCF v1.0 and Benchmark among the four selected algorithms. The result showed that EVI contributed little to the predictive capability. Therefore, in the following analysis, both the MHCF v1.0 and Benchmark models do not include EVI.

Figure 4. Model performance. (a,c) R² of MHCF v1.0 and Benchmark with EVI or without EVI, respectively; (b,d) RMSE (t/ha) of MHCF v1.0 and Benchmark with EVI or without EVI, respectively.

Figure 5. The correlation coefficient between Enhanced Vegetation Index (EVI) and yield during the whole growing season. Consecutive Months represent the correlation coefficient for each month which includes the previous months' information; specific months represent the correlation coefficient for the specific month. The red rectangle indicates the peak season.

Figure 6. The relationship between yield and Different Variables.
3.2. Comparison of Machine Learning Models

As stated in Section 3.1, the S2S-based model alone achieved the highest accuracy among the ML-based models (XGB > RF > SVR) (Figures 3 and 4). For MLR, the performance of the S2S-based model was the worst among the three groups of input variables, followed by observation-based, and then S2S+observation-based, which had the most input data, and achieved the highest R^2, which was completely different from the results obtained by the ML-based models. One possible explanation for this is that the MLR method essentially captures the correlation between input variables and yield [47], and the correlation between observational climate data and yield is much greater than that between S2S atmospheric prediction and yield (Figure 6). Thus, the ML models (i.e., XGBoost, RF, and SVR) outperformed the linear method (i.e., MLR), since most relationships between yield and different variables are nonlinear, and ensemble learning methods are better able to capture these relationships than the linear method [22].

3.3. Spatial Patterns of Yield Predictions at the Grid Level

The spatial distributions of predicted yields are also provided (Figure 7). We only evaluated the predictions made by XGBoost and RF for 2005, 2010, and 2014, using the S2S atmospheric prediction and observational climate data as training inputs, respectively, as they showed the best prediction skill among the four candidate models. Figure 7 shows the spatial patterns of predicted yield by XGBoost, and Figure 8 shows the results predicted by RF. In general, the spatial patterns of the predicted yield were consistent with the recorded yield. The high-yield grids were mainly concentrated in the east of the planting area, while the grids with low yield were mainly distributed in the west. As shown in Figure 7, some high-yield grids in the east were slightly underestimated. The prediction yields obtained by RF in Figure 8 are roughly similar to those of XGBoost.

![Figure 5. The correlation coefficient between Enhanced Vegetation Index (EVI) and yield during the whole growing season. Consecutive Months represent the correlation coefficient for each month which includes the previous months’ information; specific months represent the correlation coefficient for the specific month. The red rectangle indicates the peak season.](image-url)
Figure 6. The correlation coefficient between observational climate variables (a) or FGOALS-f2 ensemble S2S atmospheric prediction (b) and yield. **"** indicates a correlation coefficient (r) with statistical significance levels of \( p \)-value < 0.01.

To further compare the performances of different sources of input data, we present the mean prediction PE of winter wheat yield predicted by XGBoost. It was found that grids with low yield tended to have larger PE (\( \geq 25\% \)) (Figure 9a,b). The PE differences between Benchmark and MHCF v1.0 were obtained by the PE of Benchmark to minus that of MHCF v1.0 (Figure 9c). The model performance from integrating different sources of meteorological data also corresponded with crop yield. In other words, the high-yield grids using the MHCF v1.0 model achieved a satisfactory \( R^2 \) compared to the low-yield grids, whereas, in contrast, the combination of low-yield grids and Benchmark model achieved better predictive precision.
Figure 7. Spatial patterns of the yield predicted with XGBoost (2005, 2010, and 2014). (a–c) the recorded yield; (d–f) the yield predicted by MHCF v1.0; (g–i) the yield predicted by the Benchmark model.

Figure 8. Spatial patterns of the yield predicted with RF (2005, 2010 and 2014). (a–c) the recorded yield; (d–f) the yield predicted by MHCF v1.0; (g–i) the yield predicted by the Benchmark model.
To further compare the performances of different sources of input data, we present spatial patterns of the mean PE for XGBoost. (a) PE calculated by Benchmark model using XGBoost; (b) PE calculated by MHCF v1.0 using XGBoost; (c) PE difference between (a,b) (former minus the latter). Positive values indicate that the Benchmark model has a greater error, while negative values indicate that the MHCF v1.0 model has a greater error.

3.4. Optimum Lead Time of Yield Prediction

To investigate the limit of making an optimum early yield prediction, we conducted an in-season prediction experiment on winter wheat yield using XGBoost and RF. The in-season predictions were made in monthly intervals, which means we added the climate information for a more recent month at each time step. In general, better performance from the model was achieved with more input data, i.e., as the prediction period approached the end of the growing season, the $R^2$ increased and the RMSE decreased [57]. As shown in Figure 10, XGBoost and RF performed comparably in terms of predictive capability for yield forecasting, sharing the same trend. Although RF achieved a stable $R^2$ and RMSE earlier than XGBoost, XGBoost had a higher $R^2$ and lower RMSE than RF. Overall, the performance of both XGBoost and RF increased with the accumulation of data. In terms of XGBoost, the $R^2$ ranged from 0.77 to 0.85 and 0.75 to 0.81 (RMSE from 0.97 to 0.78 t/ha and 1.01 to 0.89 t/ha) for MHCF v1.0 and Benchmark model, respectively (Figure 10), while for RF, the $R^2$ ranged from 0.78 to 0.84 and 0.75 to 0.8 (RMSE from 0.94 to 0.82 t/ha and 1.03 to 0.9 t/ha) for the two models, respectively (Figure 10). The results confirmed that the S2S atmospheric predictions were superior to the observational climate variables as model inputs, and XGBoost was slightly superior to RF in predicting winter wheat yield.

Regarding the optimum forecasting time, the MHCF v1.0 models (XGBoost and RF) reached a stable $R^2$ of 0.85 and 0.84 in February and January, respectively, with a lead time of about four months before the harvesting of winter wheat (Figure 10a). By contrast, the Benchmark models (XGBoost and RF) resulted in the highest $R^2$ of 0.81 and 0.8 in March and January for XGBoost and RF, respectively, with a lead time of three months (Figure 10b). Our work achieves a more satisfactory lead time before harvest compared with previous studies [22,44,57]. The findings demonstrate that S2S atmospheric predictions have the potential to achieve an earlier optimum yield forecast than observational climate data as model inputs.
Figure 10. Temporal progression of model performance ($R^2$ and RMSE) based on the four models: (a,c) models trained with S2S atmospheric predictions; (b,d) models trained with observational climate variables.

4. Discussion

4.1. Model Performance

In this work, we have further developed the four most commonly used models (XGBoost, RF, SVR, MLR) for winter wheat yield forecasting. Linear regression generally explains the linear relationship between input and output variables, but from Figure 6, we could speculate that yield is not the result of a linear relationship with the input variables. In this study, the MLR method performed the worst for simulating yield among the developed models, which simply confirmed this point. ML algorithms have shown strong predictive capability, especially for XGBoost and RF. Thus, ML methods can achieve superior performance in capturing the complex relationships between S2S atmospheric predictions and yield, and they provide a window of opportunity to predict crop yield at regional, or even global, scales. Here, we built yield prediction models for winter wheat within the framework of ML because of its simplicity and efficiency, but also its limitations. By contrast, deep learning, with its higher accuracy at the cost of computational intensity and model complexity, may have great potential for improving global grain yield forecasting, and it is essential to find a balance between complexity and efficiency for future research [58]. Moreover, the hybridization of process-based crop growth models and/or ML with deep learning, as well as multi-source data, also has the potential to offer an improved and optimized technique for yield forecasting [17,59].

4.2. Comparison of Different Data Sources

Our findings demonstrated that the S2S atmospheric predictions outperformed those based on observational climate data. This may be largely due to S2S dynamical atmospheric predictions being able to provide more information over the simple use of observational climate data to simulate possible future scenarios, confronted with climate change, together with climate teleconnections between a region of interest and other parts of the globe [27,60]. Compared to previous works, which usually used observed climate data and satellite
data, the inclusion of S2S atmospheric prediction data in our study greatly improved the accuracy of winter wheat yield prediction. More S2S dynamic prediction models from various institutions (e.g., IAP-CAS, ECMWF, NECP-CFSv2, etc.) could be used for comparison to improve yield forecasting for future study [41,42]. In further research, deep learning methods should be considered to improve the accuracy of S2S atmospheric prediction [34,61]. Besides, the horizontal resolution of S2S atmospheric prediction systems should be increased to a convection-permitting resolution (<10 km), to provide crop yield predictions at finer spatial scales.

For the satellite data, only EVI was included in our study, and the results showed that EVI contributed little to the yield prediction model. The existing study showed that EVI provides most information in the peak season because that peak-season EVI contains biotic or abiotic stress information and may not be captured by the accumulated climate information [22]. However, climate variables over the whole growing period play critical roles in determining the wheat growth and final wheat yield. Thus, the performance of EVI for yield prediction was insignificant for predictions based on the whole growing season (Figure 5). Moreover, the newly emerging satellite Sun-Induced Chlorophyll Fluorescence (SIF) contains information about the physiological, biochemical, and the fraction of absorbed photosynthetically active radiation (fPAR) [62,63], and some studies indicated that SIF provided feasible satellite data for crop yield estimation at a larger scale, and it was not inferior to EVI [33,43]. For future research, SIF will undoubtedly play an important role in crop yield prediction.

4.3. Optimum Lead Time of Winter Wheat Yield Prediction

Our results also indicated the capability of MHCF v1.0 in winter wheat yield prediction, which achieved the highest R$^2$, ranging from 0.76–0.85, with a 3–4-month lead time (Figure 10). The model performance and lead time are consistent with, or better than, existing previous works; for example, a study about wheat yield prediction in Australia showed that their model achieved R$^2$ ranging from 0.73–0.75, with a 2-month lead time before harvest [22], and another study also predicted winter wheat yield in China and achieved higher accuracy, with R$^2$ ranging from 0.79–0.81, 1–2 months in advance of the harvest dates [56]. These findings show that the inclusion of S2S atmospheric predictions, as in our work, improves yield prediction performance, and that has increased as the S2S atmospheric prediction outputs can normally be available. In our study, we used the S2S atmospheric prediction data with a forecast length of 1 month in advance. Further research can compare the accuracy of S2S prediction outputs with different forecast lengths (1–6 months in advance) for the yield prediction model.

5. Conclusions

This study aimed to establish a forecasting model (MHCF v1.0) for the major production areas of winter wheat yield in northern China, using ML driven by an S2S atmospheric prediction system. To this end, we incorporated various datasets, including satellite data, observational climate data, and S2S atmospheric prediction data, for the four models (MLR, SVR, RF, XGBoost). We designed several experiments to test the model performance and compare the predictive performance of S2S atmospheric predictions, observational climate data, and the combination of meteorological and satellite data. Firstly, our results indicated that S2S atmospheric predictions are superior in their forecasting performance, in terms of crop yield, as compared with observational climate data, and the inclusion of S2S atmospheric predictions largely improves yield prediction performance. Secondly, we demonstrated that ML methods, especially ensemble learning models, perform significantly better than linear-regression-based methods (XGB > RF > SVR > MLR). Thirdly, a skilled prediction, 3–4 months before the harvest, was achieved by XGBoost, with R$^2$ ranging from 0.81–0.85 and RMSE ranging from 0.78–0.89 t/ha. This research proved that MHCF v1.0 is a novel and promising technique for regional yield prediction, and we hope that it can be extended to crop yield forecasts in other regions, and even on a global scale.
Author Contributions: Conceptualization, J.C.; methodology, J.C., J.L. and Q.T.; software, H.W.; validation, H.W.; data curation, J.C., J.L.; writing—original draft preparation, H.W.; writing—review and editing, J.C., J.L., Q.T. and D.N.; visualization, H.W.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by grants from the National Natural Science Foundation of China ((Project Nos. 41701111, 42005117), and the Guangdong Major Project of Basic and Applied Basic Research (Project No. 2020B0301030004).

Data Availability Statement: The datasets that support our work are as follows. The IAP-CAS FGOALS-f2 S2S dynamical atmospheric prediction outputs were provided by the Institute of Atmospheric Physics, Chinese Academy of Sciences (https://apps.ecmwf.int/datasets/data/s2s-realtime-daily-averaged-anso/levtype=sfc/type=cf/, accessed on 17 August 2021). The winter wheat yield data were obtained from the Ministry of Agriculture of China (https://doi.org/10.6084/m9.figshare.18093674, accessed on 9 January 2022). The winter-wheat-planting areas were used to mask the winter wheat yields (https://data.mendeley.com/datasets/jbs44b2hrk/2, accessed on 15 March 2022). EVI data were derived from MOD13C1 (Collection 6) records (https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MOD13C1, accessed on 9 September 2021). Historical observational climate variables were obtained from the CRU (https://crudata.uea.ac.uk/cru/data/hrg/, accessed on 24 Jun 2021).

Acknowledgments: The authors acknowledge the ArcGIS software provided by Esri and the Python software provided by the Python Software Foundation.

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

References

1. Tilman, D.; Balzer, C.; Hill, J.; Befort, B.L. Global food demand and the sustainable intensification of agriculture. Proc. Natl. Acad. Sci. USA 2011, 108, 20260–20264. [CrossRef] [PubMed]
2. Ray, D.K.; Gerber, J.S.; Macdonald, G.K.; West, P.C. Climate variation explains a third of global crop yield variability. Nat. Commun. 2015, 6, 5989. [CrossRef] [PubMed]
3. Prosekov, A.Y.; Ivanova, S.A. Food security: The challenge of the present. Geoforum 2018, 91, 73–77. [CrossRef]
4. Cole, M.B.; Augustin, M.A.; Robertson, M.J.; Manners, J.M. The science of food security. NPJ Sci. Food 2018, 2, 1–8. [CrossRef] [PubMed]
5. Cogato, A.; Meggio, F.; Migliorati, M.D.A.; Marinello, F. Extreme weather events in agriculture: A systematic review. Sustainability 2019, 11, 2547. [CrossRef]
6. Chipanshi, A.; Zhang, Y.; Koudiao, L.; Newlands, N.; Davidson, A.; Hill, H.; Warren, R.; Qian, B.; Daneshfar, B.; Bedard, F.; et al. Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agric. For. Meteorol. 2015, 206, 137–150. [CrossRef]
7. Iizumi, T.; Shin, Y.; Kim, W.; Kim, M.; Choi, J. Global crop yield forecasting using seasonal climate information from a multi-model ensemble. Clim. Serv. 2018, 21, 11–23. [CrossRef]
8. 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]
9. FAO; IFAD; UNICEF; WFP; WHO. The State of Food Security and Nutrition in the World 2021: Transforming Food Systems for Food Security, Improved Nutrition and Affordable Healthy Diets for All; Food and Agriculture Organization: Rome, Italy, 2021; ISBN 9789251015702.

Funding: This work was funded by grants from the National Natural Science Foundation of China ((Project Nos. 41701111, 42005117), and the Guangdong Major Project of Basic and Applied Basic Research (Project No. 2020B0301030004).

Data Availability Statement: The datasets that support our work are as follows. The IAP-CAS FGOALS-f2 S2S dynamical atmospheric prediction outputs were provided by the Institute of Atmospheric Physics, Chinese Academy of Sciences (https://apps.ecmwf.int/datasets/data/s2s-realtime-daily-averaged-anso/levtype=sfc/type=cf/, accessed on 17 August 2021). The winter wheat yield data were obtained from the Ministry of Agriculture of China (https://doi.org/10.6084/m9.figshare.18093674, accessed on 9 January 2022). The winter-wheat-planting areas were used to mask the winter wheat yields (https://data.mendeley.com/datasets/jbs44b2hrk/2, accessed on 15 March 2022). EVI data were derived from MOD13C1 (Collection 6) records (https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MOD13C1, accessed on 9 September 2021). Historical observational climate variables were obtained from the CRU (https://crudata.uea.ac.uk/cru/data/hrg/, accessed on 24 Jun 2021).

Acknowledgments: The authors acknowledge the ArcGIS software provided by Esri and the Python software provided by the Python Software Foundation.

Conflicts of Interest: The authors declare no conflict of interest.
16. Pan, Y.; Li, L.; Zhang, J.; Liang, S.; Zhu, X.; Sulla-Menashe, D. Winter wheat area estimation from MODIS-EVI time series data using the Crop Proportion Phenology Index. Remote Sens. Environ. 2012, 119, 232–242. [CrossRef]

17. Feng, P.; Wang, B.; Liu, D.L.; Waters, C.; Xiao, D.; Shi, L.; Yu, Q. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agric. For. Meteorol. 2020, 285–286, 107922. [CrossRef]

18. Wang, M.; Tao, F.L.; Shi, W.J. Corn yield forecasting in northeast china using remotely sensed spectral indices and crop phenology metrics. J. Integr. Agric. 2014, 13, 1538–1545. [CrossRef]

19. Zhang, J.; Feng, L.; Yao, F. Improved maize cultivated area estimation over a large scale combining MODIS-EVI time series data and crop phenological information. ISPRS J. Photogramm. Remote Sens. 2014, 94, 102–113. [CrossRef]

20. Goldberg, D.E.; Holland, J.H. Genetic algorithms and machine learning. Mach. Learn. 1988, 3, 95–99. [CrossRef]

21. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 2020, 177, 105709. [CrossRef]

22. 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]

23. Liu, Y.; Li, N.; Zhang, Z.; Huang, C.; Chen, X.; Wang, F. The central trend in crop yields under climate change in China: A systematic review. Sci. Total Environ. 2020, 704, 135355. [CrossRef] [PubMed]

24. Liaqat, M.U.; Cheema, M.J.M.; Huang, W.; Mahmood, T.; Zaman, M.; Khan, M.M. Evaluation of MODIS and Landsat multiband vegetation indices used for wheat yield estimation in irrigated Indus Basin. Comput. Electron. Agric. 2017, 138, 39–47. [CrossRef]

25. Rembold, F.; Meroni, M.; Urbano, F.; Royer, A.; Atzberger, C.; Lemoine, G.; Eerens, H.; Haesen, D. Remote sensing time series analysis for crop monitoring with the SPIRITS software: New functionalities and use examples. Front. Environ. Sci. 2015, 3, 1–11. [CrossRef]

26. Wu, B.; Meng, J.; Li, Q.; Yan, N.; Du, X.; Zhang, M. Remote sensing-based global crop monitoring: Experiences with China’s CropWatch system. Int. J. Digit. Earth 2014, 7, 113–137. [CrossRef]

27. 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]

28. Peng, B.; Guan, K.; Pan, M.; Li, Y. Benefits of Seasonal Climate Prediction and Satellite Data for Forecasting U.S. Maize Yield. Geophys. Res. Lett. 2018, 45, 9662–9671. [CrossRef]

29. Vitart, F.; Robertson, A.W. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. npj Clim. Atmos. Sci. 2018, 1, 3. [CrossRef]

30. Wang, X.; Huang, J.; Feng, Q.; Yin, D. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. Remote Sens. 2020, 12, 1744. [CrossRef]

31. Luo, Y.; Zhang, Z.; Li, Z.; Chen, Y.; Chen, Y.; Zhang, L.; Cao, J.; Tao, F.; Tao, F. Identifying the spatiotemporal changes of annual harvesting areas for three staple crops in China by integrating multi-data sources. Environ. Res. Lett. 2020, 15, 074003. [CrossRef]

32. Franch, B.; Vermote, E.F.; Becker-Reshef, I.; Claverie, M.; Huang, J.; Zhang, J.; Justice, C.; Sobrino, J.A. Improving the timeliness of winter wheat production forecast in the United States of America, Ukraine and China using MODIS data and NCAR Growing Degree Day information. Remote Sens. Environ. 2015, 161, 131–148. [CrossRef]

33. Cao, J.; Zhang, Z.; Tao, F.; Zhang, L.; Luo, Y.; Zhang, J.; Han, J.; Xie, J. Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches. Agric. For. Meteorol. 2021, 297, 108275. [CrossRef]

34. Ma, Y.; Zhang, Z.; Kang, Y.; Özdoğan, M. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. Remote Sens. Environ. 2021, 259, 112408. [CrossRef]

35. Zhou, Y.; Yang, B.; Chen, H.; Zhang, Y.; Huang, A.; La, M. Effects of the Madden–Julian Oscillation on 2-m air temperature prediction over China during boreal winter in the S2S database. Clim. Dyn. 2019, 52, 6671–6689. [CrossRef]

36. Begueira, S.; Vicente-Serrano, S.M.; Reig, F.; Latorre, B. Standardized precipitation evapotranspiration index (SPEI) revisited: Parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. Int. J. Climatol. 2014, 34, 3001–3023. [CrossRef]

37. Li, J.; Bao, Q.; Liu, Y.; Wang, L.; Yang, J.; Wu, G.; Wu, X.; He, B.; Wang, X.; Zhang, X.; et al. Effect of horizontal resolution on the simulation of tropical cyclones in the Chinese Academy of Sciences FGOALS-f3 climate system model. Geosci. Model Dev. 2021, 14, 6113–6133. [CrossRef]

38. Li, J.; Bao, Q.; Liu, Y.; Wu, G.; Wang, L.; He, B.; Wang, X.; Yang, J.; Wu, X.; Shen, Z. Dynamical seasonal prediction of tropical cyclone activity using the fgoals-f2 ensemble prediction system. Weather Forecast. 2021, 36, 1759. [CrossRef]

39. Vitart, F.; Ardilouze, C.; Bonet, A.; Brookshaw, A.; Chen, M.; Codorean, C.; Déqué, M.; Ferranti, L.; Fucile, E.; Fuentes, M.; et al. The subseasonal to seasonal (S2S) prediction project database. Bull. Am. Meteorol. Soc. 2017, 98, 163–173. [CrossRef]

40. Li, J.; Bao, Q.; Liu, Y.; Wu, G.; Wang, L.; He, B.; Wang, X.; Li, J. Evaluation of FAMIL2 in Simulating the Climatology and Seasonal-to-Interannual Variability of Tropical Cyclone Characteristics. J. Adv. Model. Earth Syst. 2019, 11, 1117–1136. [CrossRef]

41. Feng, X.; Klingaman, N.; Zhang, S.; Guo, L. Building sustainable science partnerships between early-career researchers to better understand and predict east asia water cycle extremes. Bull. Am. Meteorol. Soc. 2020, 101, E785–E789. [CrossRef]

42. Ren, H.L.; Wu, Y.; Bao, Q.; Ma, J.; Liu, C.; Wan, J.; Li, Q.; Wu, X.; Liu, Y.; Tian, B.; et al. The China Multi-Model Ensemble Prediction System and Its Application to Flood-Season Prediction in 2018. J. Meteorol. Res. 2019, 33, 540–552. [CrossRef]
43. 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]

44. Guo, Y.; Fu, Y.; Hao, F.; Zhang, X.; Wu, W.; Jin, X.; Robin Bryant, C.; Senthilnath, J. Integrated phenology and climate in rice yields prediction using machine learning methods. *Ecol. Indic.* **2021**, *120*, 106935. [CrossRef]

45. Cawley, G.C.; Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.

46. Molinaro, A.M.; Simon, R.; Pfeiffer, R.M. Prediction error estimation: A comparison of resampling methods. *Bioinformatics* **2005**, *21*, 3301–3307. [CrossRef]

47. Bouras, E.H.; Jarlan, L.; Er-Raki, S.; Balaghi, R.; Amazirh, A.; Richard, B.; Khabba, S. Cereal yield forecasting with satellite drought-based indices, weather data and regional climate indices using machine learning in morocco. *Remote Sens.* **2021**, *13*, 3101. [CrossRef]

48. Aiken, L.S.; West, S.G.; Pitts, S.C.; Baraldi, A.N.; Wurpts, I.C. Multiple Linear Regression, Second Edition 2. *Handb. Psychol.* **2012**, *18*, 511–542. [CrossRef]

49. Gun, R.S. Support Vector Machines for classification and regression. *Analyst* **1998**, *135*, 230–267. [CrossRef]

50. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]

51. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

52. Wang, Y.; Zhang, Z.; Feng, L.; Du, Q.; Runge, T. Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States. *Remote Sens.* **2020**, *12*, 1232. [CrossRef]

53. 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; Volume 42, pp. 785–794.

54. Chen, T.; He, T. XGBoost: eXtreme Gradient Boosting, R Package Version 1.5.0.1; 8 November 2021, pp. 1–4. Available online: https://doi.org/10.6084/m9.figshare.19478261 (accessed on 31 March 2022).

55. Song, Y.; Liu, X.; Zhang, L.; Jiao, X.; Qiang, Y.; Qiao, Y.; Liu, Z. Prediction of double-high biochemical indicators based on lightGBM and XGBoost. In Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, Wuhan, China, 12–13 July 2019; pp. 189–193. [CrossRef]

56. 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]

57. Li, L.; Wang, B.; Feng, P.; Wang, H.; He, Q.; Wang, Y.; Liu, D.L.; Li, Y.; He, J.; Feng, H.; et al. Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. *Agric. For. Meteorol.* **2021**, *308–309*, 108558. [CrossRef]

58. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N. Prabhat Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef] [PubMed]

59. Shahhosseini, M.; Hu, G.; Huber, I.; Archontoulis, S.V. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Sci. Rep.* **2021**, *11*, 1606. [CrossRef] [PubMed]

60. Ogotu, G.E.O.; Franssen, W.H.P.; Supit, I; Omondi, P.; Hutjes, R.W.A. Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate forecasts. *Agric. For. Meteorol.* **2018**, *250–251*, 243–261. [CrossRef]

61. Ham, Y.G.; Kim, J.H.; Luo, J.J. Deep learning for multi-year ENSO forecasts. *Nature* **2019**, *573*, 568–572. [CrossRef]

62. Cao, J.; An, Q.; Zhang, X.; Xu, S.; Si, T.; Niyogi, D. Is satellite Sun-Induced Chlorophyll Fluorescence more indicative than vegetation indices under drought condition? *Sci. Total Environ.* **2021**, *792*, 148396. [CrossRef] [PubMed]

63. Song, L.; Guanter, L.; Guan, K.; You, L.; Huete, A.; Ju, W.; Zhang, Y. Satellite sun-induced chlorophyll fluorescence detects early response of winter wheat to heat stress in the Indian Indo-Gangetic Plains. *Glob. Chang. Biol.* **2018**, *24*, 4023–4037. [CrossRef] [PubMed]