remote sensing

MDPI

Article

Rice Yield Prediction and Model Interpretation Based on Satellite and Climatic Indicators Using a Transformer Method

Yuanyuan Liu [1,2], Shaoqiang Wang [1,2,3,*], Jinghua Chen [1,2], Bin Chen [1,2], Xiaobo Wang [1,2], Dongze Hao [2,4] and Leigang Sun [5,6]

[1] Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Lab of Reginal Ecological Processes and Environmental Change, School of Geography and Information Engineering, Chinese University of Geosciences, Wuhan 430074, China
[4] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[5] Institute of Geographical Sciences, Hebei Academy of Sciences, Shijiazhuang 050011, China
[6] Hebei Technology Innovation Center for Geographic Information Application, Shijiazhuang 050011, China
* Correspondence: sqwang@igsnrr.ac.cn

Abstract: As the second largest rice producer, India contributes about 20% of the world's rice production. Timely, accurate, and reliable rice yield prediction in India is crucial for global food security and health issues. Deep learning models have achieved excellent performances in predicting crop yield. However, the interpretation of deep learning models is still rare. In this study, we proposed a transformer-based model, Informer, to predict rice yield across the Indian Indo-Gangetic Plains by integrating time-series satellite data, environmental variables, and rice yield records from 2001 to 2016. The results showed that Informer had better performance (R^2 = 0.81, $RMSE$ = 0.41 t/ha) than four other machine learning and deep learning models for end-of-season prediction. For within-season prediction, the Informer model could achieve stable performances ($R^2 \approx 0.78$) after late September, which indicated that the optimal prediction could be achieved 2 months before rice maturity. In addition, we interpreted the prediction models by evaluating the input feature importance and analyzing hidden features. The evaluation of feature importance indicated that $\mathrm{NIR_V}$ was the most critical factor, while intervals 6 (mid-August) and 12 (mid-November) were the key periods for rice yield prediction. The hidden feature analysis demonstrated that the attention-based long short-term memory (AtLSTM) model accumulated the information of each growth period, while the Informer model focused on the information around intervals 5 to 6 (August) and 11 to 12 (November). Our findings provided a reliable and simple framework for crop yield prediction and a new perspective for explaining the internal mechanism of deep learning models.

Keywords: crop yield prediction; remote sensing; deep learning; feature importance; attention

check for updates

Citation: Liu, Y.; Wang, S.; Chen, J.; Chen, B.; Wang, X.; Hao, D.; Sun, L. Rice Yield Prediction and Model Interpretation Based on Satellite and Climatic Indicators Using a Transformer Method. *Remote Sens.* **2022**, *14*, 5045. https://doi.org/10.3390/rs14195045

Academic Editor: David M Johnson

Received: 17 August 2022
Accepted: 7 October 2022
Published: 10 October 2022

1. Introduction

Rice feeds more than 50% of the world's 7.5 billion people [1], especially, 90% of rice production and 87% of paddy rice harvested area come from Asia [2,3]. As the second largest rice producer in the world after China, India contributes about 20% of the world's rice production [4]. However, rice yield in India highly depends on the variability of monsoon affected by climate change in the tropics, which is the most vulnerable area in the world [5,6]. Under the pressure of climate change, increasing crop yield to meet the rising population demand and adapt to climate change is challenging [5,7]. Therefore, timely, accurate, and reliable rice yield prediction in India plays a vital role in food security, health issues, and marketing planning on the local, national, and global levels [8].

In recent decades, popular crop yield prediction methods include field surveys, crop growth models, and statistical machine learning models [9–11]. Traditionally, field surveys are used to estimate crop yield by collecting information from farmers. Field surveys can capture the actual growth of crops, but this method is time-consuming, laborious, and difficult to be applied on large scales [12,13]. Due to these limitations, crop growth models and statistical machine learning models are more reliable and convenient for large-scale predictions. Crop growth models can simulate crop growth, plant carbon assimilation, and phenology using the information of meteorology, soil, phenology, and management [14]. However, these models require sufficient field data to adjust the specific parameters [15], which restricts their performances at the regional-scale application [16].

Compared with crop growth models, machine learning methods can reduce dependence on crop-specific parameters as remote sensing technology advances and describe nonlinear relationships between input features and crop yield. Therefore, many machine learning approaches have been applied to predict crop yield at the regional scales based on climate and remote sensing data [10,17–19]. For instance, Shahhosseini et al. [17] applied four machine learning methods, including least absolute shrinkage and selection operator (LASSO), ridge, random forest (RF), extreme gradient boosting (XGBoost), and their ensembles as metamodels, to predict maize yield and found that XGBoost had the best performance with a relative mean square error (*RMSE*) of 13.5%. Felipe Maldane et al. [18] used RF, multiple linear regression (MLR), and artificial neural network (ANN) to estimate the sugarcane yield, and the results showed that the prediction of RF was the closest to the observed yield.

In addition, deep learning methods can capture complex relationships from high-dimensional data and recognize sequential patterns of time-series data [20,21]. Thus, deep learning methods have been widely used in solving real-world problems with the advancement of artificial intelligence [22–24]. Recently, popular deep learning methods, such as the convolutional neural network (CNN) and long short-term memory (LSTM), have been applied to predict crop yield and outperformed traditional machine learning methods [13,25–27]. For instance, Cao et al. [27] predicted the county-level winter wheat yield in China using deep learning methods, including DNN (deep neural networks), 1D-CNN (1D convolutional neural networks), and LSTM, while Tian et al. [13] found that the LSTM model outperformed the BPNN (back propagation neural network) and SVM (support vector machine) in predicting wheat yield.

Deep learning methods have achieved high accuracy in predicting crop yield. However, as ‘black-boxes’, deep learning methods have the disadvantage of low interpretability. Therefore, how to improve interpretability is a crucial research question [28]. Existing studies using deep learning methods mainly focus on improving yield prediction performance without considering model interpretation. Improving interpretability is essential to ensure the model’s reliability and to understand the environmental features relevant to crop yield. Recently, the attention mechanism has provided a channel for interpreting deep learning models by exploring the distribution of attention weights. Based on the attention mechanism, the transformer model was proposed in natural language processing [29], providing an approach to improve the interpretability of models. Precisely, the transformer models can explain the internal working of a model using self-attention weights. Compared with other traditional machine learning methods, transformer models have proved their state-of-the-art performance in natural language processing applications, such as generative language modeling [30,31]. Moreover, the latest agricultural studies have used these transformer models in crop classification [32–34]. To our best knowledge, there are still no studies that use transformer models to predict crop yield based on remote sensing and climate data.

In this study, we adopted a transformer model called Informer to improve the performance of district-level rice yield prediction and interpreted the internal mechanism of models from multiple perspectives. First, the Informer model was applied with multi-temporal inputs based on satellite and meteorological data from 2001 to 2016. Then, the

Informer model was compared with LASSO, RF, XGBoost, and AtLSTM for performance evaluation. Finally, the potential driving factors for yield prediction were investigated by analyzing the feature importance and attention weights. This study seeks to address two questions:

(1) How does the Informer model perform for rice yield prediction compared with several traditional machine learning and deep learning models?

(2) How can we interpret deep learning models for predicting crop yield based on attention mechanisms?

2. Materials

Three types of data, including time-series satellite variables, climate variables, and district-level rice yield, were collected from different sources. The detailed information of input data is shown in Table 1.

Table 1. Summary of the input features for model development.

Category	Variables	Related Crop Properties	Spatial Resolution	Source	Temporal Resolution
Satellite imagery	NDVI	Plant vigor	1000 m	MODIS	16-day
	EVI				
	$\mathrm{NIR_V}$				
	SIF		0.05 degree	CSIF	4-day
Climate	Tmax	Heat stress	0.5 degree	CRUNCEP	1-day
	Tmin				
	Srad				
	Pr	Water stress	0.05 degree	CHIRPS	1-day
Others	Historical average yield (t/ha)		District-level		N/A
	Crop area		500 m	MODIS	Yearly

Abbreviations: NDVI, normalized difference vegetation index; EVI, enhanced vegetation index; $\mathrm{NIR_V}$, near-infrared reflectance of vegetation; SIF, solar-induced chlorophyll fluorescence; Tmax, daily maximum temperature; Tmin, daily minimum temperature; Srad, solar radiation; Pr, daily total precipitation.

2.1. Study Area and Yield Data

The study area is the Indian Indo-Gangetic Plains in northwest India (Figure 1), a flat alluvial plain with an average evaluation of 177 m, consisting of three states: Punjab, Haryana, and Uttar Pradesh. The soil texture of this plain is mainly sandy loam to loam, with moderate water-holding capacity, low organic matter content, and high fertility. The water resources in this region are primarily derived from rainfall and groundwater, with 97% of cropland in Punjab, 83% in Haryana, and 68% in Uttar Pradesh being irrigated [35]. The dominant cropping system is Kharif rice in rotation with winter wheat in the Indian Indo-Gangetic Plains. The wet season when Kharif rice is grown spans approximately May–June through October–November. The dry season for wheat growing spans approximately November–December through March–April [36,37]. Districtwise Kharif rice yield data for 101 districts over 16 years (2001–2016) were retrieved from the Area and Production Statistics database (https://aps.dac.gov.in (accessed on 10 May 2021)). The location and topographic maps of the Indian Indo-Gangetic Plains are mapped in Figure 1.

2.2. Satellite Imagery

We considered the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), near-infrared reflectance of vegetation ($\mathrm{NIR_V}$), and solar-induced chlorophyll fluorescence (SIF) for crop yield prediction in this study. Both NDVI and EVI were extracted from the 16-day Moderate Resolution Imaging Spectroradiometer (MODIS) MOD13A2 product with a spatial resolution of 1 km. $\mathrm{NIR_V}$ was calculated using the MODIS NIR reflectance ($\mathrm{NIR_T}$) and MODIS NDVI from the MOD13A2 product [38].

$$\mathrm{NIRV} = \mathrm{NDVI} \times \mathrm{NIR} \tag{1}$$

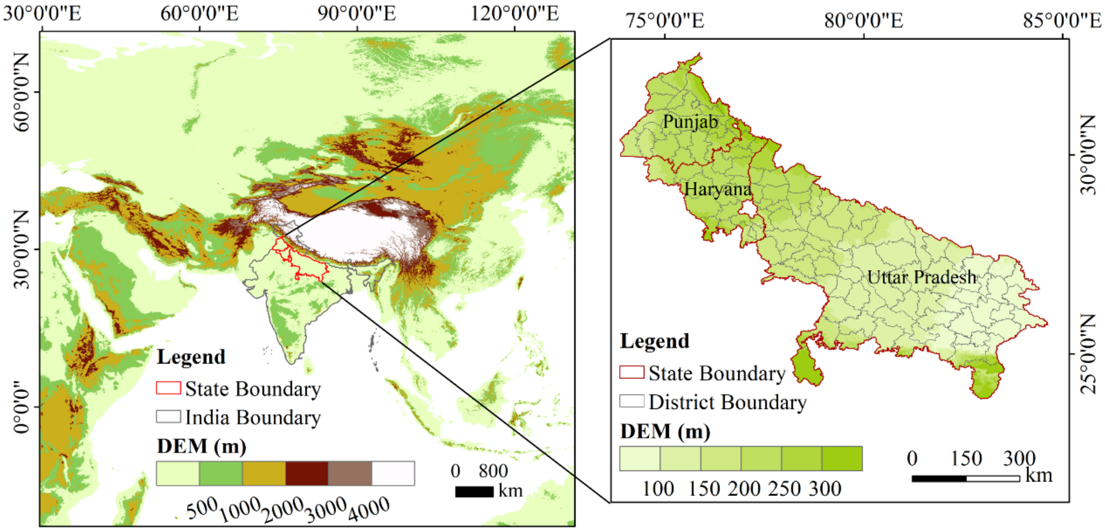

Figure 1. The location and topographic maps of the Indian Indo-Gangetic Plains. The topographic data from GEBCO gridded bathymetry data (https://www.gebco.net/ (accessed on 26 September 2022)).

Besides, the satellite-based SIF data were extracted from a global spatially contiguous SIF (CSIF) dataset product, which provides 4-day data with 0.05° spatial resolution [39]. By training a neural network, Zhang et al. [39] generated the CSIF dataset using surface reflectance from the MODIS and SIF from the Orbiting Carbon Observatory-2 (OCO-2). SIF has been proven effective in monitoring the photosynthetic activity of terrestrial ecosystems and is widely used in predicting crop yield [19].

2.3. Environmental Data

Environmental data consisted of four climate variables. We collected three climate variables from the 6 h atmospheric forcing dataset, CRUNCEP Version 7 [40]. The three climate variables, including daily maximum temperature (Tmax, °C), minimum temperature (Tmin, °C), and solar radiation (Srad, W m^{-2}), were extracted from the CRUNCEP dataset with a spatial resolution of 0.5°. In addition, the daily total precipitation (Pr, mm) was collected from the Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS) dataset, which has a 0.05° spatial resolution [41].

2.4. Data Preprocessing

The climatic variables were first calculated and aggregated to a 16-day interval from 25 May (or 24 May) to 2 December (or 1 December), resulting in 12 time-series input features (Table 2). Then, the climate and satellite data were unified into 500 m spatial resolutions and masked by the cropland layer derived from the MODIS Land Cover Type product (MCD12Q1 v6), which has a spatial resolution of 500 m and can be obtained from the Google Earth Engine (GEE) platform. We did not use a specific crop type classification map since the rice–wheat rotation accounts for about 85% of the cultivated land in the Indo-Gangetic Plains [42]. Therefore, rice is the dominant crop of the Kharif season in the study area, and the croplands of MODIS can effectively represent the distribution area of rice. In addition, we did not use the spatial production allocation model (SPAM) data, which had specific crop type maps, because it was only available in 2005 and 2010, while the cropping area changed from 2001 to 2016. The eight sequential variables (i.e., NDVI, EVI, NIR_V, SIF, Tmax, Tmin, Srad, and Pr) were aggregated spatially to the district level. Finally, a total of 96 features were applied to build models, consisting of eight sequential variables with 12 time intervals for each.

Table 2. The start dates and end dates of the 12 time intervals.

Number of the Interval	Start Date		End Date	
	Normal Year	Leap Year	Normal Year	Leap Year
1	25 May	24 May	9 June	8 June
2	10 June	9 June	25 June	24 June
3	26 June	25 June	11 July	10 July
4	12 July	11 July	27 July	26 July
5	28 July	27 July	12 August	11 August
6	13 August	12 August	28 August	27 August
7	29 August	28 August	13 September	12 September
8	14 September	13 September	29 September	28 September
9	30 September	29 September	15 October	14 October
10	16 October	15 October	31 October	30 October
11	01 November	31 October	16 November	15 November
12	17 November	16 November	2 December	1 December

3. Method

3.1. Informer Model

The transformer model based on the self-attention structure was proposed by Vaswani in 2017 [29]. The transformer model uses the self-attention mechanism rather than the sequential structure of RNNs, which has efficient parallel computing and can capture global information. Compared with RNN models, transformer-based models have superior performance in solving long-sequence time-series forecasting problems [43]. However, traditional transformer-based models need expensive costs of computation and dozens of CPUs owing to the self-attention mechanism and architecture. Therefore, it is unaffordable for the transformer models to solve realistic problems. To overcome the deficiencies of transformer models and improve prediction performance, the Informer model, which can achieve better efficiency and overcome resource consumption, was proposed by Zhou et al. [43].

We present the architecture of the Informer model for rice yield prediction in Figure 2. The Informer model consists of an input layer, two encoder layers, one decoder layer, and an output layer. The canonical self-attention [29] is replaced by the ProbSparse self-attention mechanism in the Informer model. The ProbSparse self-attention can efficiently reduce the time complexity and memory usage. The encoder consists of two identical multihead attention layers formed by combining multiple self-attention. The encoder module is designed to handle longer sequential inputs under memory usage limitations by using the self-attention distilling operation. The self-attention distilling operation can extract dominating attention and reduce the network size sharply, which is beneficial for receiving long-sequence inputs. The decoder used in the Informer model is a standard decoder structure containing a stack of two identical multihead attention layers [29]. Then, the output of the decoder is aggregated into the fully connected layer, and the fully connected layer acquires the final output. The detailed architecture of Informer can be referred to in a study by Zhou et al. [43].

To obtain the best performance of the Informer model, we set the dimension of the model to 512, the number of heads to 8, the dimension of the fully connected layer to 2048, the ProbSparse attention factor to 5, the experiment times to 3, the train epochs to 10, the batch size to 16, the early stopping patience to 3, and the optimizer learning rate to 0.0001. The number of encoder and decoder layers was set to 2 and 1, respectively. The Informer model was implemented using the open-source PyTorch library in Python 3.6.

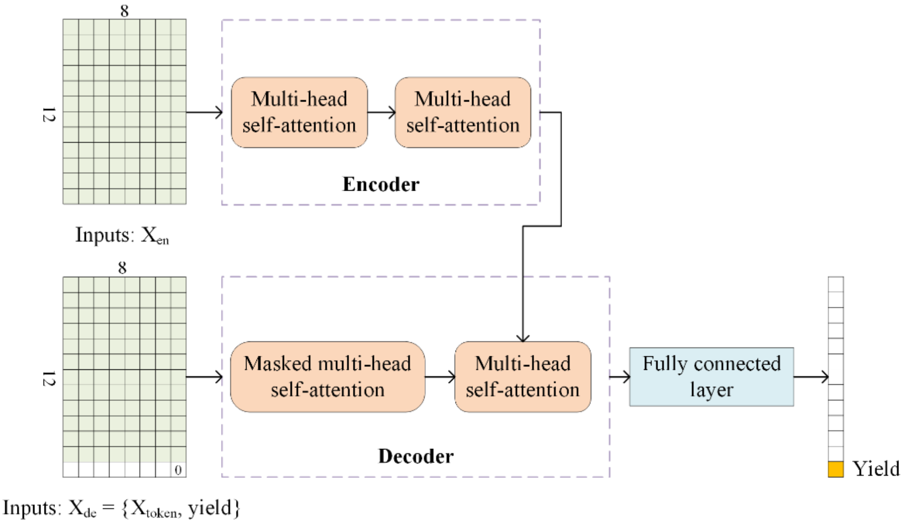
8
12
Multi-head self-attention
Multi-head self-attention
Encoder
Inputs: X_{en}
8
12
0
Masked multi-head self-attention
Multi-head self-attention
Decoder
Fully connected layer
Yield
Inputs: $X_{de} = \{X_{token}, \text{yield}\}$

Figure 2. The architecture of the Informer model.

3.2. Baseline Models

To validate the proposed Informer model, we compared the Informer model with four other models, including one linear regression method (least absolute shrinkage and selection operator regression (LASSO)), two machine learning methods (random forest (RF) and extreme gradient boosting (XGBoost)), and one deep learning method (attention-based long short-term memory (AtLSTM)). The following content explains the detailed information about the other four models.

LASSO is a regularized regression method and uses L1 regularization [44]. The main parameter of LASSO is alpha, which was tuned using the grid search with cross-validation. RF is an ensemble machine learning algorithm composed of numerous independent decision trees. RF has an accurate prediction ability, thanks to RF avoiding overfitting and injecting correct randomness [45]. XGBoost has parallel and distributed computing, which makes XGBoost have a quicker running speed than other popular machine learning methods [46]. The parameters of RF and XGBoost were tuned using a randomized search with cross-validation. LASSO and RF were realized using the scikit-learn library. XGBoost was trained using the XGBoost library.

The LSTM model was proposed by Hochreiter and Schmidhuber [21] and has solved the error flow and artificial long-time-lag tasks in existing recurrent neural networks (RNN) [21]. LSTM has been widely applied to real-world sequences processing problems such as natural language processing [47] and crop yield prediction [48,49]. The AtLSTM model is an attention-based LSTM model that is practical and reliable in many fields, such as the industrial field [50]. In this study, we used a six-layer AtLSTM model to predict rice yield, which consists of one input layer, two LSTM layers, one attention module layer, one fully connected layer, and one output layer (Figure 3). The input is a time series with climate and satellite data at 12 intervals of the rice growth period.

The LSTM cell consists of three gates: a forget gate g_t^f, an input gate g_t^i, and an output gate g_t^o, which can control the cell state (S_t) and output (O_t). The calculation formula used for generating hidden features O_t is shown in Equations (2)–(6):

$$g_t^f = \sigma(W_f \cdot [O_{t-1}, I_t] + b_f) \quad (2)$$

$$g_t^i = \sigma(W_i \cdot [O_{t-1}, I_t] + b_i) \quad (3)$$

$$S_t = g_t^f \cdot S_{t-1} + g_t^i \cdot tanh(W_c \cdot [O_{t-1}, I_t] + b_c) \quad (4)$$

$$g_t^o = \sigma(W_o \cdot [O_{t-1}, I_t] + b_o) \quad (5)$$

$$O_t = g_t^o \cdot tanh(S_t) \quad (6)$$

where g_t^f, g_t^i, and g_t^o are the forget gate, input gate, and output gate, respectively. W_f, W_i, W_c, and W_o are the weight matrices of the forget gate, input gate, input candidate element, and output gate, respectively. b_f, b_i, b_c, and b_o are biases of the forget gate, input gate, input candidate element, and output gate, respectively. S_t, $S_{t\text{-}1}$, O_t, and $O_{t\text{-}1}$ are the current cell state, previous cell state, current cell output, and previous cell output. I_t is the current input. σ is the sigmoid function.

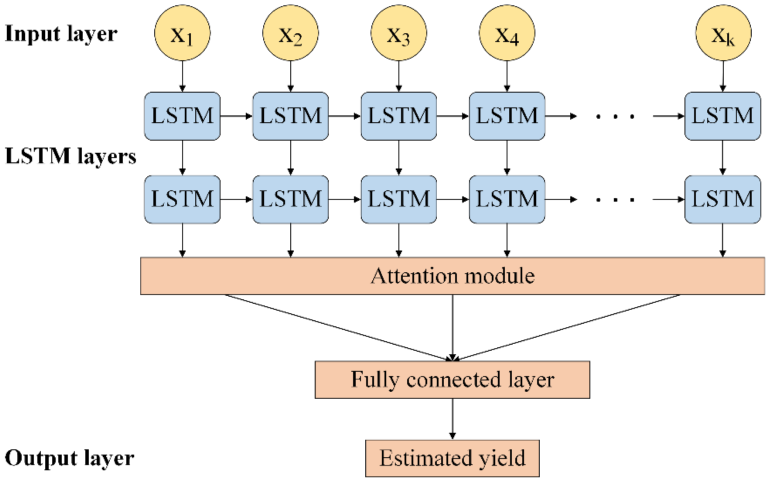

Figure 3. The architecture of the AtLSTM (attention-based long short-term memory network) model.

To better utilize multi-time-step information for crop yield prediction, the attention module adjusts the contribution of hidden features by normalizing the weight parameters. The attention weight α_t of each hidden feature O_t^* obtained by the LSTM layers is generated using a *softmax* function (Equations (7) and (8)):

$$\alpha_t = softmax(W_a \cdot O_t + b_a) \tag{7}$$

$$O_t^* = \alpha_t O_t \tag{8}$$

where W_a and b_a are learnable weight matrices in the attention module and shared by all time steps. α_t is the attention weight. O_t^* is the adjusted hidden feature. Then, the output of the attention module is aggregated into the fully connected layer.

To derive the best performance of the AtLSTM model, we set the hidden features to 32, the batch size to 16, and the number of LSTM layers to 2, and applied the dropout mechanism to the LSTM layers to prevent overfitting by setting the dropout rates to 20%. We applied an Adam optimizer with a learning rate of 0.001 to optimize the parameters of the network. The mean squared error (MSE) is a loss function used to evaluate the difference between predicted and actual values. The AtLSTM model was implemented using the PyTorch library in Python 3.6.

3.3. Model Interpretation Approaches

3.3.1. Input Feature Importance Evaluation

The feature importance of two machine learning methods (RF and XGBoost) was used to rank features. Greater values of feature importance indicate higher relative importance. We evaluated the feature importance of two machine learning methods (RF and XGBoost) from 2009 to 2016.

3.3.2. Hidden Feature Analysis

We used two interpretation methods to interpret hidden features of two deep learning models, including attention weight distribution for AtLSTM and self-attention weight matrix for Informer. First, we applied attention values to evaluate the relative importance of hidden features of AtLSTM when predicting rice yield. Compared with the traditional LSTM model, the AtLSTM model can give higher attention weights to essential features and assign lower weights to redundant features [51]. Therefore, the attention weights of AtLSTM can represent the relative importance of hidden features and make the model more interpretable. We analyzed the attention values of AtLSTM over time.

Second, the self-attention weight matrix was adopted to interpret hidden features of the Informer model. The attention mechanism is one of the main structures of the transformer model. Each input feature has a different attention weight using the attention mechanism [52]. The analysis of attention weights can interpret the internal mechanism of the transformer model. The higher the weight of low-level features at a specific time step, the more critical that time step is for high-level features. In this study, we calculated the self-attention weight matrix of the first and second encoder layers, indicating the importance of features in each time step. The attention matrix of this study is the average self-attention weight matrix of multiple heads and all samples.

3.4. Model Evaluation

To evaluate the practicality of the models, we applied the leave-one-year-out prediction to all the models for each test year from 2009 to 2016. We first eliminated the test year and then trained the model to predict the test year with all the remaining years. The leave-one-year-out strategy has been applied in previous studies [8,10,27]. Although we cannot capture the real-world future data, the leave-one-year-out strategy ensures that each test year has the same number of training years, making the model performance independent of the amount of training data [53]. We applied 2009–2016 as the test years to ensure that each test year has the same training data without using too much future data. We further used the better-performing models to evaluate the within-season yield prediction. The within-season yield prediction using data from different periods can detect when the models can achieve the optimal prediction.

Our study was conducted at the district level. Therefore, we selected R^2 (coefficients of determination), *RMSE* (root mean square error), *MAPE* (mean absolute percentage error), and *d* (Willmott's index of agreement) [54] between the observed and predicted district-level yield for each prediction year to assess performance.

$$R^2 = \left(\frac{\sum_{i=1}^{\mathrm{n}} (O_i - \overset{-}{O})(P_i - \overset{-}{P})}{\sqrt{\sum_{i=1}^{\mathrm{n}} (O_i - \overset{-}{O})^2} \sqrt{\sum_{i=1}^{\mathrm{n}} (P_i - \overset{-}{P})^2}} \right)^2 \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\mathrm{n}}} \tag{10}$$

$$MAPE = \frac{100\%}{\mathrm{n}} \sum_{i=1}^{n} \left| \frac{P_i - O_i}{O_i} \right| \tag{11}$$

$$\mathrm{d} = 1 - \frac{\sum\limits_{i=1}^{n} |Pi - Oi|}{\sum\limits_{i=1}^{n} \left(|Pi - \overline{O}| + |Oi - \overline{O}| \right)} \tag{12}$$

where n is the number of samples. P_i and O_i represent the predicted and observed values. $\overline{P}$ and $\overline{O}$ represent the mean values of the predicted and observed yields.

4. Results

4.1. Model Performance and Comparison

The Informer model was compared with the four other methods for the end-of-season prediction using the leave-one-year-out strategy. R^2, *RMSE* (t/ha), *MAPE* (%), and *d* of rice yield prediction in each of the past 8 years by the five models are shown in Figure 4. Overall, the Informer model showed a better predictive capability than all other models, with an average R^2 of 0.81, *RMSE* of 0.41 t/ha, *MAPE* of 15.47%, and *d* of 0.74 from 2009 to

2016. Moreover, the Informer model performed best in multiple testing years. Additionally, the machine learning (i.e., RF and XGBoost) and deep learning (i.e., AtLSTM) models provided better results than the traditional linear regression model (i.e., LASSO) in rice yield prediction in India, demonstrating that the machine learning methods could handle the nonlinear relationship between yield and input variables. The scatter plot showed that the observed and predicted rice yield had a linear and positive pattern for the five models in the testing years from 2009 to 2016 (Figures 5 and S1). Overall, we found that the Informer model showed the best performance in predicting rice yield in multiple testing years. In addition, the AtLSTM model had the most satisfactory performance in 2013. Meanwhile, we found the lowest performance in the LASSO model compared with the other models (Figures 5 and S1).

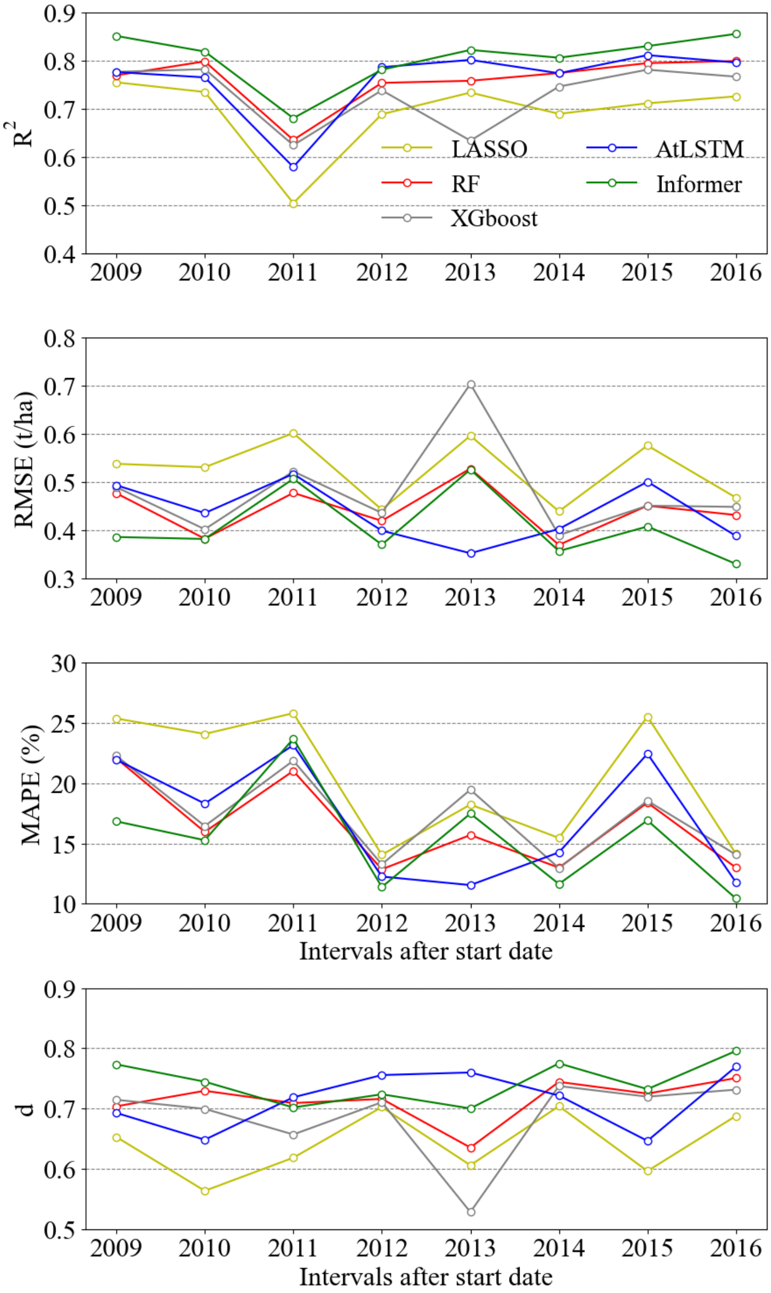

Figure 4. R^2, *RMSE* (t/ha), *MAPE* (%), and *d* of five models in rice yield prediction.

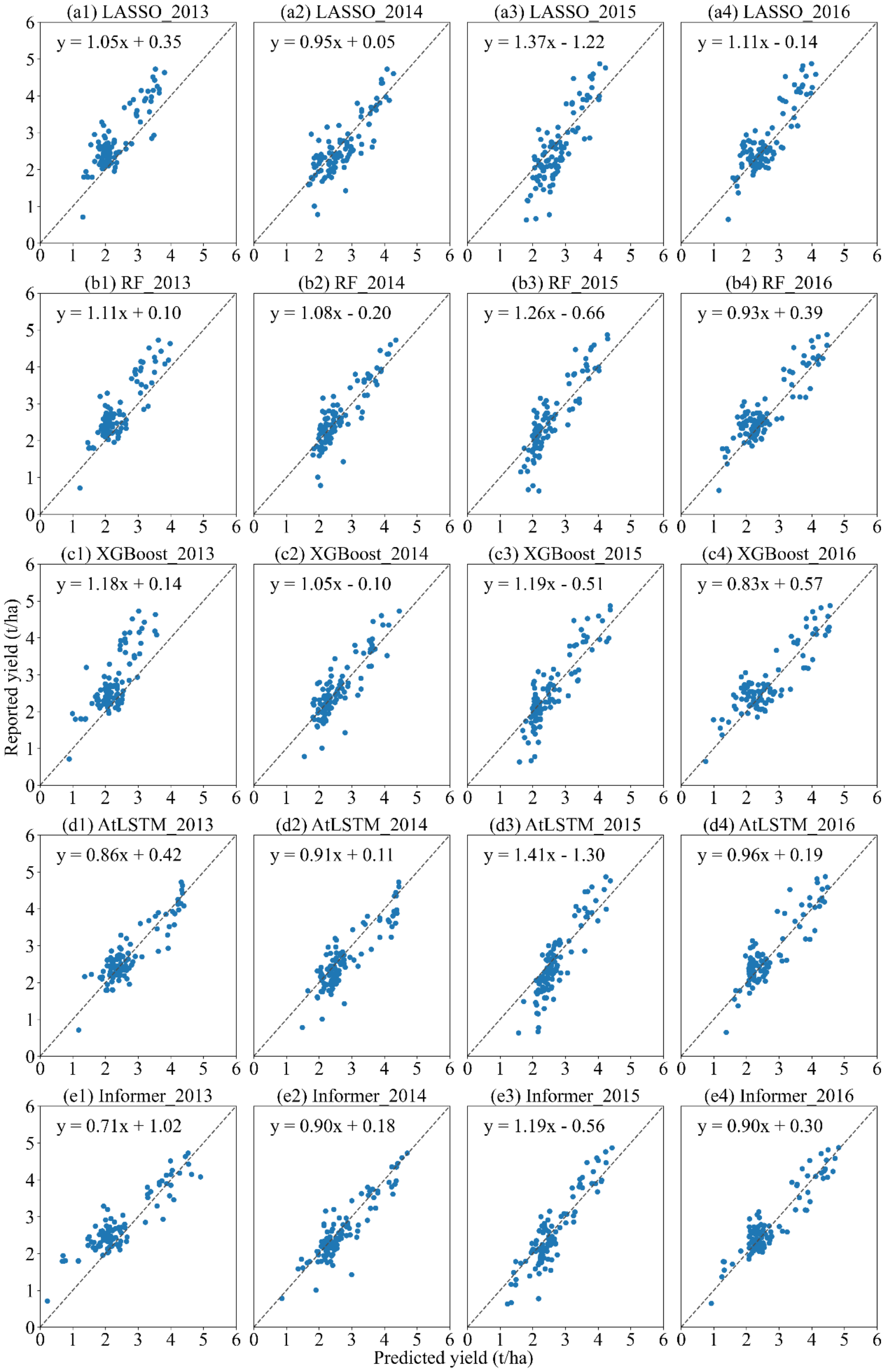

Figure 5. Scatter plots of reported yields vs. predicted yields of (**a1**–**a4**) LASSO, (**b1**–**b4**) RF, (**c1**–**c4**) XGBoost, (**d1**–**d4**) AtLSTM, and (**e1**–**e4**) Informer in 4 testing years: (1) 2013; (2) 2014; (3) 2015; (4) 2016.

The spatial patterns of the predicted yield for all testing years are presented in Figures 6 and S2. Overall, the spatial patterns of the predicted yield were in good agreement with the observed yield, especially for the four machine learning models. The Punjab State, the northwest of the study area, had high-yield districts. In contrast, the Uttar Pradesh State had low-yield districts in the southeast of the study area. The proposed Informer model best agreed with the observed yield for multiple testing years, while the

LASSO performed worst, especially in the high-yield area. The relative error maps of the predicted rice yield in all testing years are shown in Figures 7 and S3. Comparing all five models, the Informer model outperformed the other models in multiple test years, such as 2012, 2014, 2015, and 2016, indicating that the Informer model can effectively improve the performance of yield prediction. In contrast, the LASSO model performed the worst, especially in 2010 and 2015, compared with the other models. The three baseline methods, RF, XGBoost, and AtLSTM, were more advantageous than the linear LASSO but not as well as the Informer model.

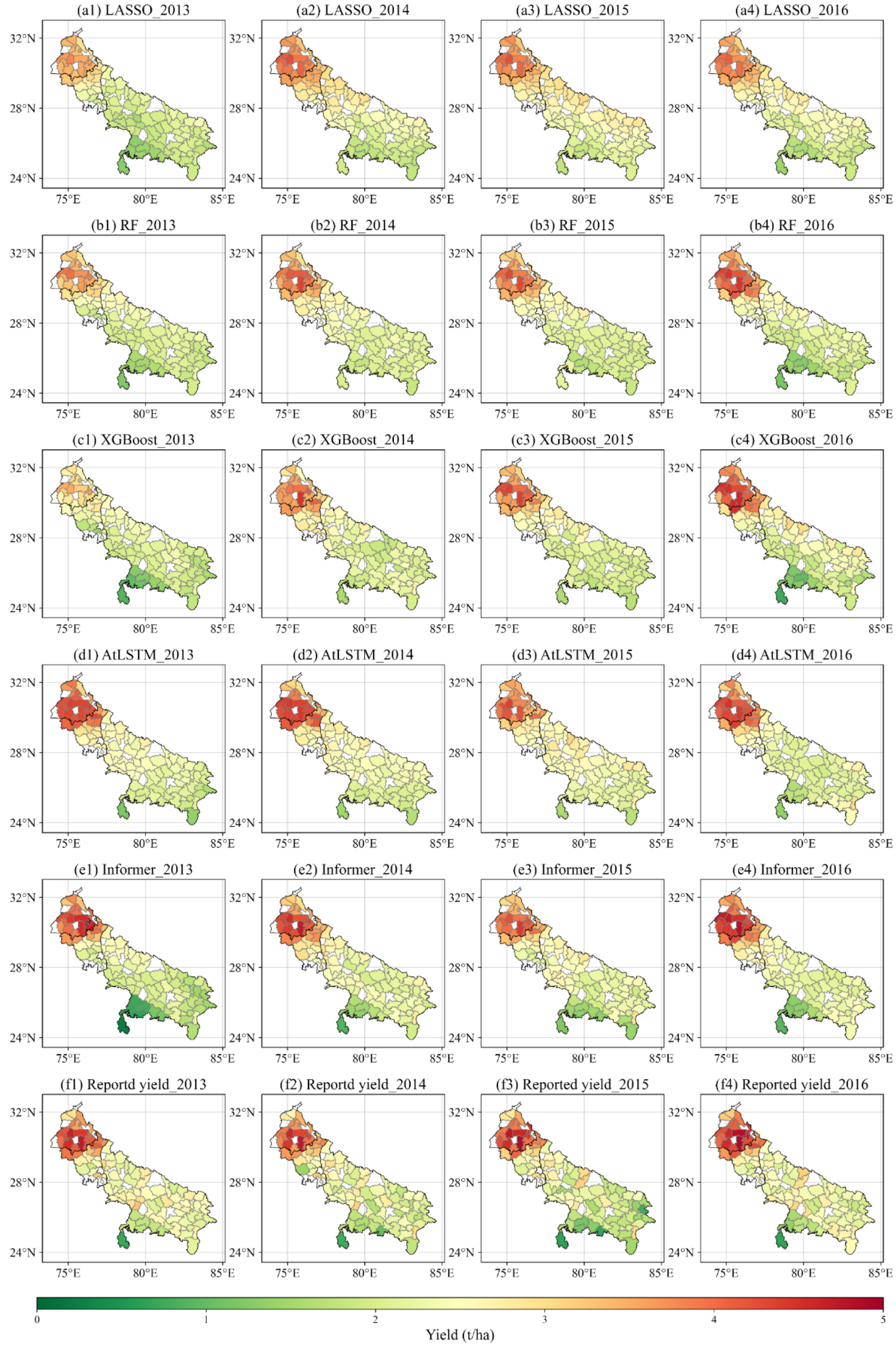

Figure 6. Spatial patterns of the predicted yield of (**a1**–**a4**) LASSO, (**b1**–**b4**) RF, (**c1**–**c4**) XGBoost, (**d1**–**d4**) AtLSTM, (**e1**–**e4**) Informer, and (**f1**–**f4**) reported yield from (1) 2013 to (4) 2016.

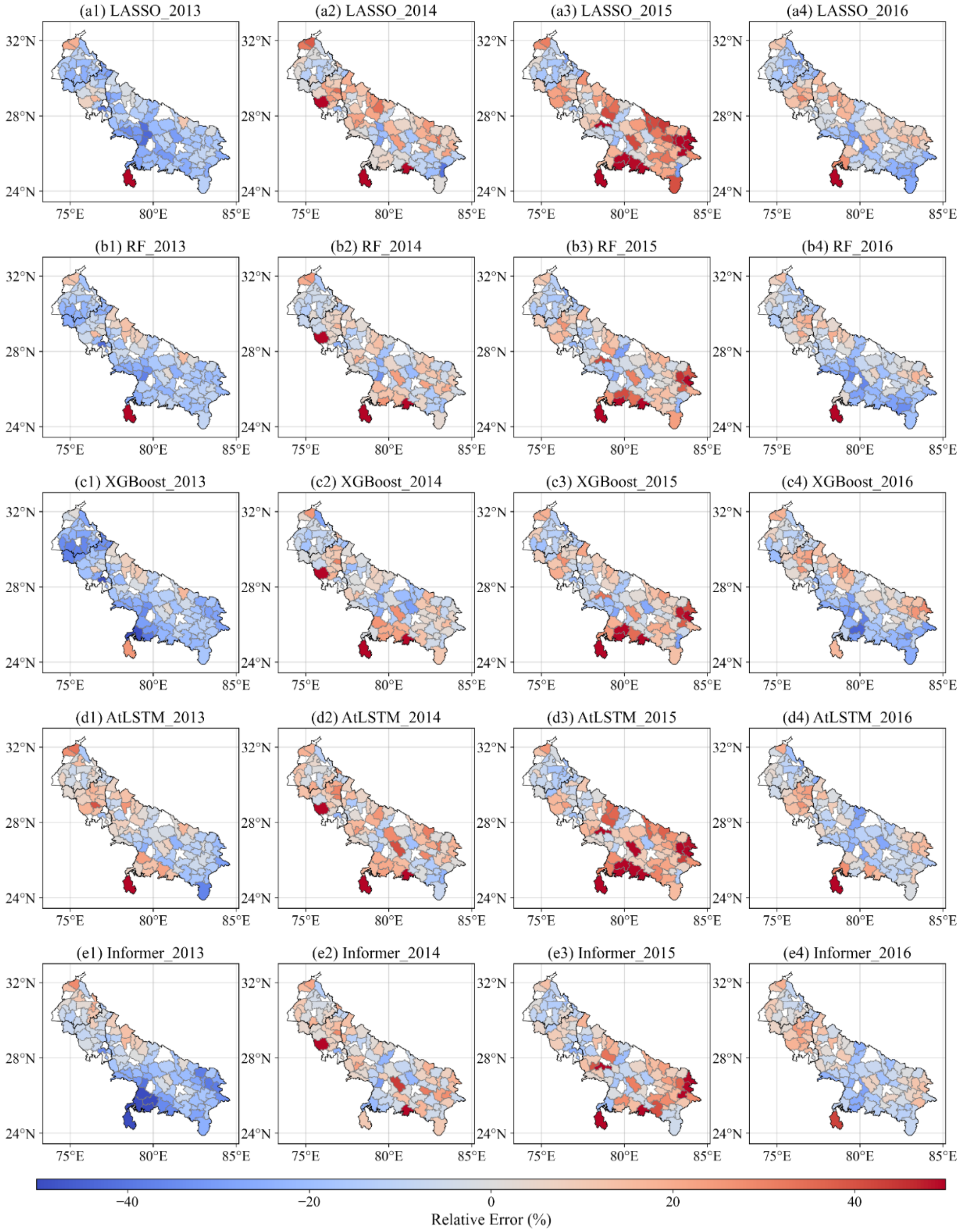

Figure 7. Spatial patterns of the relative errors of (**a1**–**a4**) LASSO, (**b1**–**b4**) RF, (**c1**–**c4**) XGBoost, (**d1**–**d4**) AtLSTM, and (**e1**–**e4**) Informer from (1) 2013 to (4) 2016.

4.2. Within-Season Yield Prediction

To further explore the performances of different models over time, the Informer model was compared with other machine learning models, including RF, XGBoost, and AtLSTM. We excluded the LASSO model in the comparison since the LASSO model had the poorest performance among all models (Figures 5 and S1). For overall comparison among models, we used each model's average R^2, *RMSE* (t/ha), *MAPE* (%), and *d* from 2009 to 2016. The within-season prediction was achieved by each model using data from different periods, with results shown in Figure 8. With the progress of the growing season, the yield prediction performance could be improved by more training data. The yield prediction performances of all the models were stable since interval 8 (mid-September), which indicated that the models could achieve the optimal prediction within 2 months before rice maturity. For rice yield prediction in all testing years, the proposed Informer model performed better than the other models since interval 9 (late-September) with the highest R^2 of 0.78, lowest *RMSE* of 0.44 t/ha, lowest *MAPE* of 16.56%, and highest *d* of 0.72, while the AtLSTM model had better performance than the other models during interval 5 (late-July) to 8 (mid- September) with the lowest *RMSE* and *MAPE*. The results demonstrate that the Informer model had the best performance in predicting rice yield, especially with more information at the end of the growing season. AtLSTM had the second-best within-season prediction performance among all models, illustrating its excellent ability to predict yield using long-term series data and the attention mechanism.

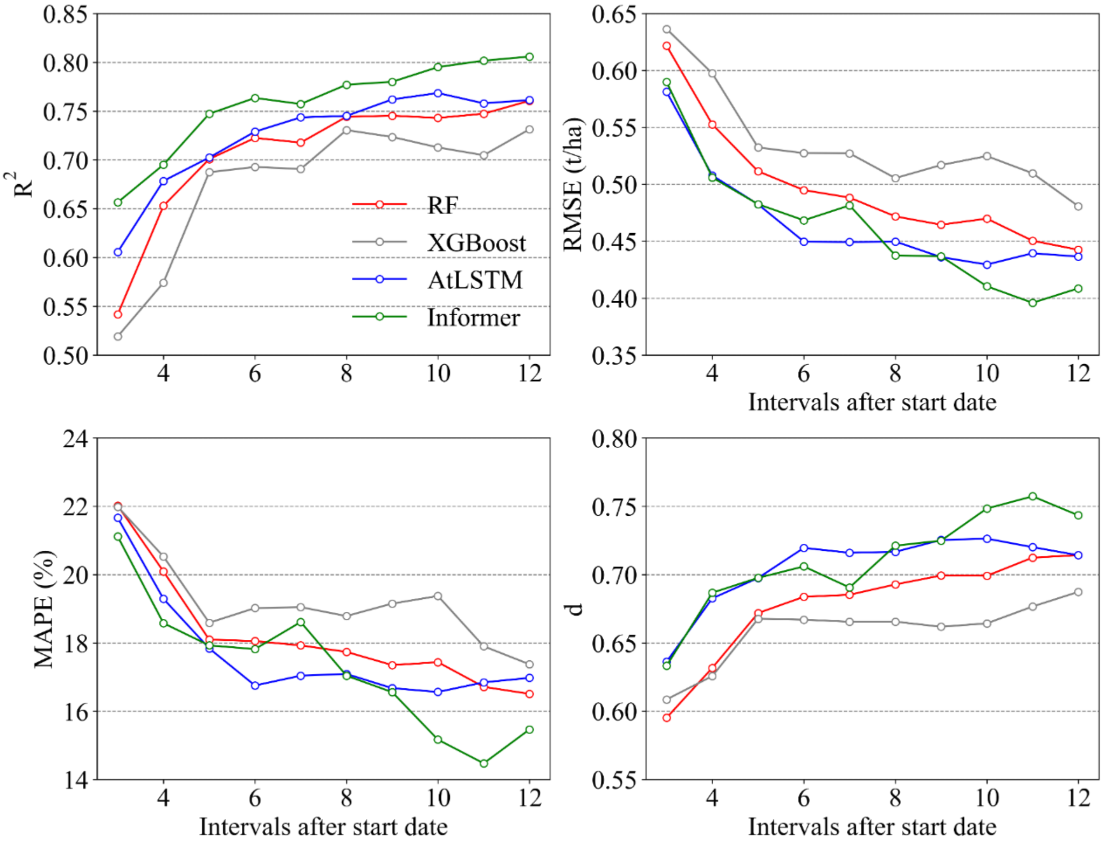

Figure 8. Average R^2, *RMSE* (t/ha), *MAPE* (%), and *d* of different models with time in all testing years (2009–2016).

The spatial patterns of the average absolute relative error in all testing years using the Informer model and information from different periods are shown in Figure 9. The average absolute relative errors of the Informer model were mainly prominent in the early growth periods, indicating that adequate information was critical for the yield prediction. As the growing season progressed, the performance of the Informer model was significantly improved in the Punjab and Haryana States. In contrast, fewer improvements were obtained in the southern Uttar Pradesh State. The spatial patterns of relative errors had been stable since mid-September, consistent with the results shown in Figure 8.

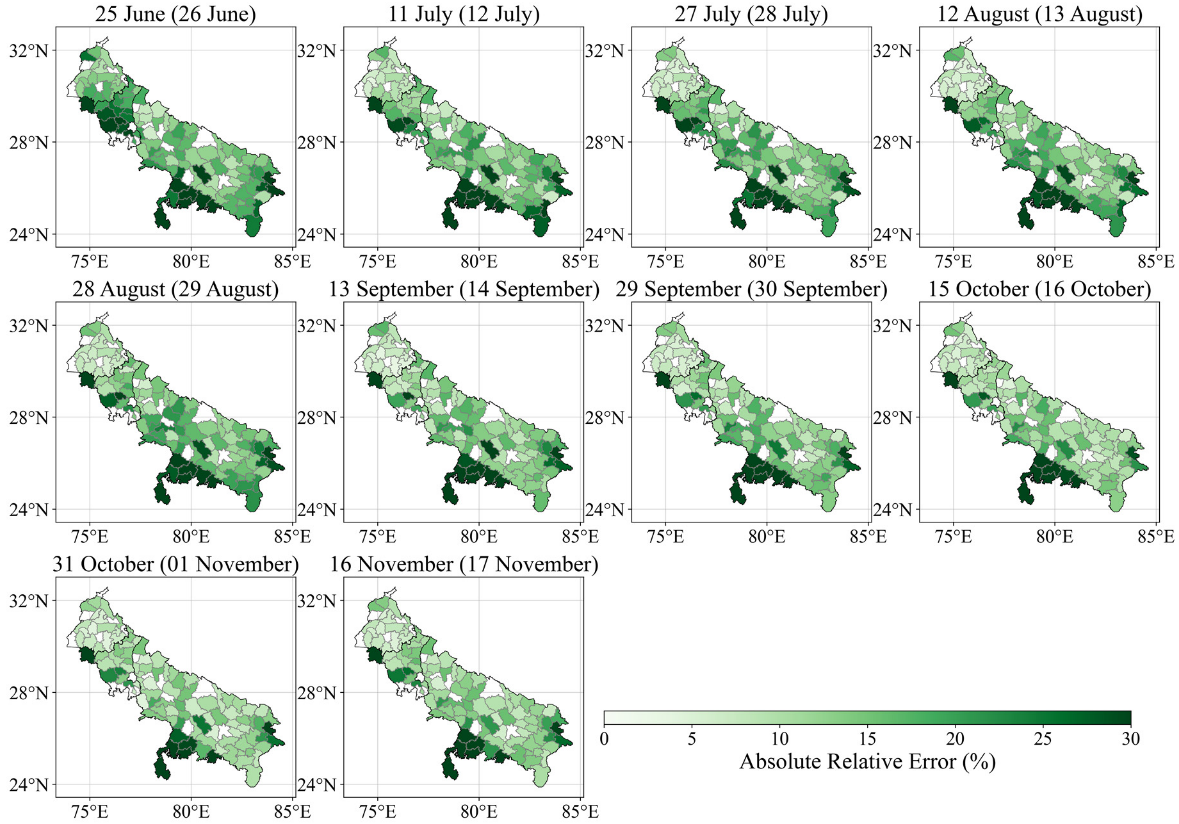

Figure 9. Spatial patterns of the average absolute relative error (%) for the Informer model with the progression of the growing season in all testing years (2009–2016).

4.3. Input Feature Importance Evaluation

The average feature importance of the top 15 features from 2009 to 2016 is given in Figure 10, and the feature importance values of each year are shown in Figures S4 and S5. Overall, the results indicate that NIR_V was the most important feature to predict rice yield among eight factors, demonstrating that NIR_V was superior to other vegetation indices in predicting crop yield. Moreover, Srad and Tmax had higher feature importance values than other climatic variables. In addition, intervals 6 and 12 had high values of feature importance among the 12 intervals in each year.

4.4. Hidden Feature Analysis

Therefore, the attention weights of AtLSTM can represent the relative importance of hidden features and make the model more interpretable. We analyzed the attention values of AtLSTM over time. The average attention weight values of the AtLSTM model from 2009 to 2016 are given in Figure 11, and the attention weight values for each year from 2009 to 2016 are given in Figure S6. Overall, we found that AtLSTM had higher attention values in the reproductive and ripening phases than in the vegetation phase. AtLSTM generated the attention values, reflecting the influence of extracted hidden features on the prediction results. Notably, the attention weight value remained below 0.09 from intervals 1 to 5 and was higher than 0.09 after interval 6 from 2009 to 2016. The attention weight value increased rapidly during intervals 1 to 6, and was stable after interval 6, indicating that the cumulative information became more and more useful in the vegetation phase and stabilized in the reproductive phase.

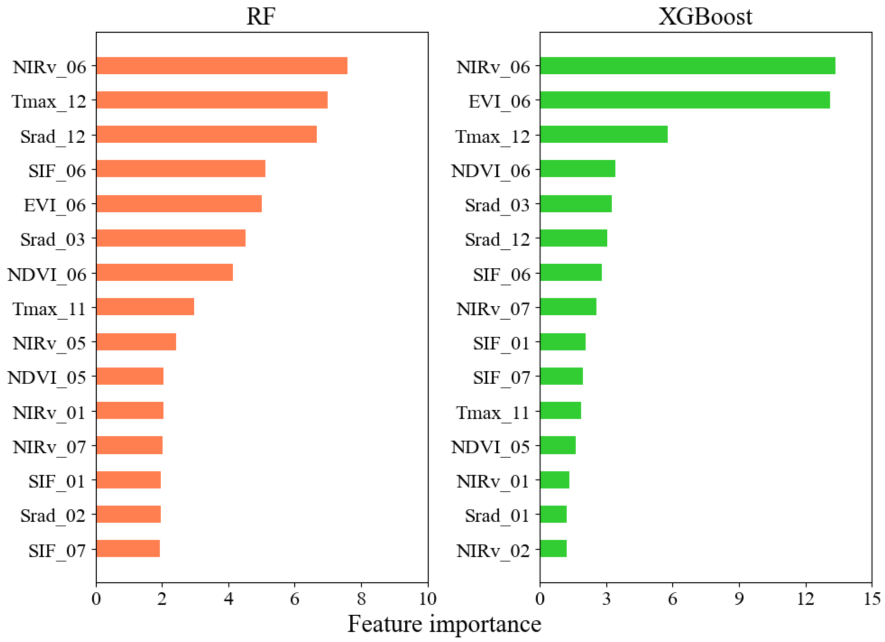

Figure 10. Average feature importance (rescaled to 0–100) of the top 15 features from the random forest and XGBoost models from 2009 to 2016.

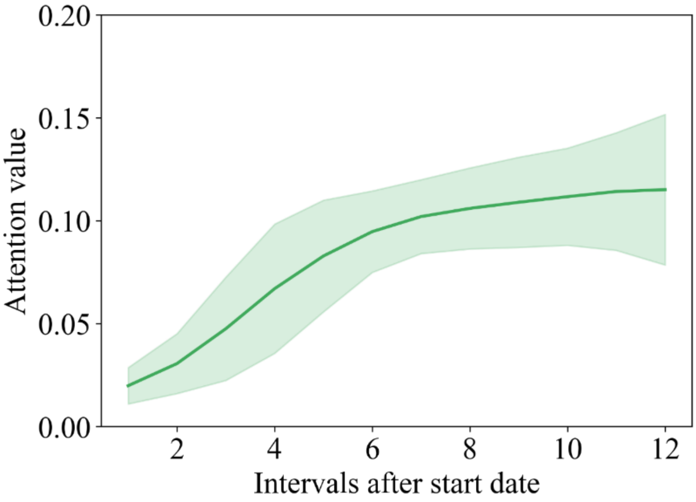

Figure 11. The average attention weight distribution of the AtLSTM model from 2009–2016. The green line stands for the average attention weight value of all test samples from 2009 to 2016, and the shadow represents the standard deviation of all test samples from 2009 to 2016.

To illustrate the importance of different intervals, we presented the average self-attention weight matrices of the Informer model from 2009 to 2016 in Figure 12 and the self-attention weight matrices of each year from 2009 to 2016 in Figures S7 and S8. In layer 1, the results indicated that intervals 3, 4, 8, and 11 had more significant contributions to high-level features than the others. In layer 2, the self-attention weight matrices showed that intervals 5 to 6 and 11 to 12 had high self-attention weight values. These results demonstrate that the leaf development and tillering stages around intervals 5 to 6 (August) and the flowering and grain-filling stages around intervals 11 to 12 (November) were important for predicting rice yield.

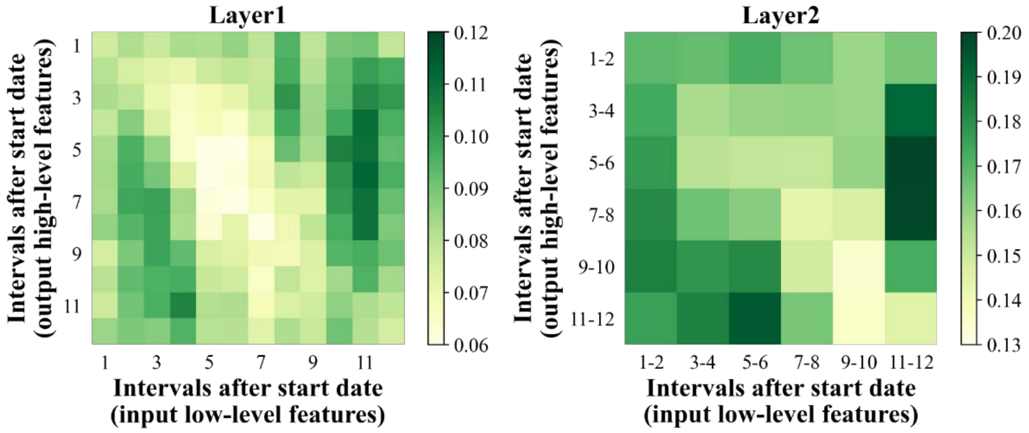

Figure 12. Average self–attention weight of the Informer model from 2009 to 2016. Each visual–ized weight matrix represents the averaged weight matrix of all heads of the same encoder layer in the Informer model. The results are the average of all test samples from 2009 to 2016.

5. Discussion

5.1. Advantage Analysis of the Informer Model

To the best of our knowledge, the transformer model based on the self-attention mechanism has not been explored in modeling crop yield in previous studies. In this study, we used a transformer model called Informer to predict rice yield based on remote sensing and climate data. Our results show that the Informer model exhibited the best performance in predicting rice yield compared with the other four machine learning models. This was mainly because the Informer model had the self-attention mechanism. Enabled by the self-attention mechanisms, the Informer model can compute correlations between each pair of remote sensing and climate variables at all time steps, and effectively extract the temporal dependence of long time series [29,34].

Compared with traditional deep learning models, such as CNN and RNN, the Informer model abandons recurrence and convolutions completely, which effectively improves model scalability and training efficiency and makes the model more parallelized [29]. In addition, another advantage of the Informer model is that it can address some deficiencies of regular transformer models, such as the consumption of considerable resources [43]. The Informer model's structure makes it better than traditional deep learning models in handling long sequence time-series forecasting problems. Therefore, the Informer model proposed in our study can be extended to other cropping areas for predicting crop yield and guiding agricultural practice.

5.2. Interpretation of Rice Yield Prediction Models

We employed machine learning and deep learning methods not only to obtain satisfactory yield prediction performance but also to explain the internal mechanism of models. Recently, attention values have been widely used to interpret deep learning models in many fields, such as financial time series prediction [51] and crop yield estimation [55]. First, we applied feature importance to interpret machine learning models. The feature importance analysis showed that NIR_V was the most critical feature (Figure 10) since NIRV can better estimate GPP and has great potential in predicting crop yield [19]. Moreover, Srad and Tmax were the crucial factors in the rice yield prediction among four climatic variables (Figure 10). The reason was that they could directly influence rice's photosynthesis and carbon cumulation.

Second, we analyzed the hidden features to interpret deep learning models using attention weights. The analysis of hidden features provides a new sight to understand the internal working of deep learning models for rice yield prediction. In this study, we applied attention weights and a self-attention weight matrix to interpret AtLSTM and Informer, respectively. The attention weight distributions of AtLSTM illustrated that the reproductive and ripening phases were more important than the vegetation phase (Figure 11). The reason may be that AtLSTM had a recurrent network structure that could store the cumulative information of crop growth with time [34]. In addition, we found that the Informer model had higher attention weights around intervals 5 to 6 (August) and intervals 11 to 12 (November) (Figure 12), which indicated that those periods provided critical information to the later layer of the Informer model. This is because the leaf development and tillering stages (August) before the heading stage of Kharif rice are closely related to rice yield [56], while the flowering and grain-filling stages (November) are sensitive to high temperature and have a great impact on rice yield [57–59]. The results agreed with the feature importance of machine learning methods, which proved the reliability of our results. The Informer model obtains the later-layer features by calculating the attention weight and weighted summation of earlier-layer features. The results show that the self-attention weights of the Informer model disagreed with the attention weights of the AtLSTM model, which could attribute to the unique structure of these two deep learning methods. The AtLSTM model accumulated the temporal information in the hidden layer by the LSTM cell, resulting in the reproductive phase storing more information than the vegetation phase. Then, the attention weight value of the reproductive phase was higher than the vegetation phase. In contrast, the Informer model acquired temporal dependencies between any two intervals instead of accumulating temporal information over time.

5.3. Uncertainties and Future Work

Overall, the proposed Informer model achieved competitive performance for rice yield prediction across the Indian Indo-Gangetic Plains, which indicated that the Informer model had great potential to predict crop yield. Furthermore, interpreting deep learning models can demonstrate model reliability by evaluating input feature importance and analyzing hidden features. However, our study still has some limitations that need to be improved. First, the current study only considers the climate and satellite factors to predict crop yield due to the scarcity of field management data on the regional scale. Still, other features, such as soil properties, tillage, and fertilization rates, are also closely related to rice yield. Our study could reduce the prediction accuracy due to the lack of considering the above features. We should further consider more related features to improve the predictive performance as more high-quality data related to crop yield become available in the future. In addition, we should consider more high-resolution data, such as data based on drones, and develop the high-resolution-based model in the future, which can make the Informer model have better performance. Second, although we have interpreted the machine learning and deep learning models based on feature importance and attention values, we ignore model uncertainty analysis in this study. The estimation of uncertainty is critical for evaluating the model's confidence in predicting crop yield, which has been suggested by some previous studies [11,60]. Therefore, we should further analyze the uncertainty of models to quantify the reliability of deep learning models in future studies. Third, we determined the optimal parameters of the Informer model by trying multiple trials and optimizing training loss, which may limit the performance of the Informer model. For future studies, automatedly searching optimal parameters will improve the accuracy of the Informer model in predicting yield. Fourth, this study used districtwise Kharif rice yield data over 16 years because smaller-scale yield statistics were not available, which caused the training sample size to be small. The small training set can make it difficult for deep learning methods to train well. Thus, more detailed statistics on crop yield can improve the performance of deep learning methods. Finally, in multiple testing years, the Informer model has outperformed the other four widely used machine learning models. However, it is difficult for the Informer model to predict crop yield under extreme scenarios, such as heatwaves, floods, and droughts. With the rise in global temperatures, the crop yield will be greatly affected by heat stress in summer, especially in the Indian Indo-Gangetic Plains [37,61]. To deal with the above issues, combining the Informer model with crop growth models, such as the decision support system for agrotechnology transfer

(DSSAT) [62], can be considered to improve the predictive ability of the Informer model under environmental stress.

6. Conclusions

Accurate rice yield estimation is vital for food security, health issues, and agriculture trading. This study proposed the Informer model to predict district-level rice yield based on multitemporal satellite data, sequential climatic products, and historical rice yield. To evaluate the proposed Informer model, we compared the Informer model with four widely used machine learning methods, including LASSO, RF, XGBoost, and AtLSTM. Overall, the results showed that the Informer model performed best among the five models, with an average R^2 of 0.81, *RMSE* of 0.41 t/ha, *MAPE* of 15.47%, and *d* of 0.74 from 2009 to 2016. In addition, we evaluated the performance of different models within the growing season. The proposed Informer model achieved optimal prediction ($R^2 \approx 0.78$) 2 months before rice maturity. The feature importance analysis showed that NIR_V was the most critical factor, while intervals 6 (mid-August) and 12 (mid-November) were the essential periods for rice yield prediction. The hidden feature analysis demonstrated that the AtLSTM model accumulated the information of each growth period, while the Informer model focused on the information around intervals 5 to 6 (August) and 11 to 12 (November). This study proposed a modeling framework to predict crop yield and explain the internal mechanism of deep learning models at the regional scale.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs14195045/s1. Figure S1. Scatter plots of reported yields vs. predicted yields of (a) LASSO, (b) RF, (c) XGBoost, (d) AtLSTM, and (e) Informer in four testing years: (1) 2009, (2) 2010, (3) 2011, and (4) 2012; Figure S2. Spatial patterns of the predicted yield of (a) LASSO, (b) RF, (c) XGBoost, (d) AtLSTM, (e) Informer, and (f) reported yield from (1) 2009 to (4) 2012; Figure S3. Spatial patterns of the relative errors of (a) LASSO, (b) RF, (c) XGBoost, (d) AtLSTM, and (e) Informer from (1) 2009 to (4) 2012; Figure S4. Feature importance (rescaled to 0–100) of the top 15 features from the random forest algorithm from 2009 to 2016; Figure S5. Feature importance (rescaled to 0–100) of the top 15 features from the XGBoost algorithm from 2009 to 2016; Figure S6. The attention weight distribution of the AtLSTM model from 2009 to 2016; Figure S7. Self-attention weight analysis of the Informer model in layer 1 from 2009 to 2016; Figure S8. Self-attention weight analysis of the Informer model in layer 2 from 2009 to 2016.

Author Contributions: Conceptualization, Y.L.; methodology, Y.L., S.W. and D.H.; formal analysis, Y.L.; investigation, Y.L.; writing—original draft, Y.L.; supervision, S.W.; writing—review and editing, S.W., J.C., B.C. and X.W.; validation, D.H. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China, No. 31861143015; the Strategic Priority Research Program of the Chinese Academy of Sciences, No. XDA23100202; and the Science and Technology Planning Project of Hebei Academy of Sciences of China, No. 22A03.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no financial and personal relationships with other people or organizations that can inappropriately influence their work.

References

1. Singha, M.; Dong, J.; Zhang, G.; Xiao, X. High resolution paddy rice maps in cloud-prone Bangladesh and Northeast India using Sentinel-1 data. *Sci. Data* **2019**, *6*, 26. [CrossRef]
2. Zhang, G.; Xiao, X.; Dong, J.; Xin, F.; Zhang, Y.; Qin, Y.; Doughty, R.B.; Moore, B., 3rd. Fingerprint of rice paddies in spatial-temporal dynamics of atmospheric methane concentration in monsoon Asia. *Nat. Commun.* **2020**, *11*, 554. [CrossRef]
3. Zhang, G.; Xiao, X.; Biradar, C.M.; Dong, J.; Qin, Y.; Menarguez, M.A.; Zhou, Y.; Zhang, Y.; Jin, C.; Wang, J.; et al. Spatiotemporal patterns of paddy rice croplands in China and India from 2000 to 2015. *Sci. Total Environ.* **2017**, *579*, 82–92. [CrossRef] [PubMed]
4. FAOSTAT. FAOSTAT Online Database. Available online: http://faostat3.fao.org/browse/Q/QC/E (accessed on 23 June 2015).

5. Gupta, R.; Mishra, A. Climate change induced impact and uncertainty of rice yield of agro-ecological zones of India. *Agric. Syst.* **2019**, *173*, 1–11. [CrossRef]
6. Soora, N.K.; Aggarwal, P.K.; Saxena, R.; Rani, S.; Jain, S.; Chauhan, N. An assessment of regional vulnerability of rice to climate change in India. *Clim. Chang.* **2013**, *118*, 683–699. [CrossRef]
7. Zabel, F.; Muller, C.; Elliott, J.; Minoli, S.; Jagermeyr, J.; Schneider, J.M.; Franke, J.A.; Moyer, E.; Dury, M.; Francois, L.; et al. Large potential for crop production adaptation depends on available future varieties. *Glob. Chang. Biol.* **2021**, *27*, 3870–3882. [CrossRef]
8. Feng, L.; Wang, Y.; Zhang, Z.; Du, Q. Geographically and temporally weighted neural network for winter wheat yield prediction. *Remote Sens. Environ.* **2021**, *262*, 112514. [CrossRef]
9. Paudel, D.; Boogaard, H.; de Wit, A.; Janssen, S.; Osinga, S.; Pylianidis, C.; Athanasiadis, I.N. Machine learning for large-scale crop yield forecasting. *Agric. Syst.* **2021**, *187*, 103016. [CrossRef]
10. 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]
11. 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]
12. 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]
13. Tian, H.; Wang, P.; Tansey, K.; Zhang, J.; Zhang, S.; Li, H. An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China. *Agric. For. Meteorol.* **2021**, *310*, 108629. [CrossRef]
14. Guan, K.; Sultan, B.; Biasutti, M.; Baron, C.; Lobell, D.B. Assessing climate adaptation options and uncertainties for cereal systems in West Africa. *Agric. For. Meteorol.* **2017**, *232*, 291–305. [CrossRef]
15. Kang, Y.; Özdoğan, M. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. *Remote Sens. Environ.* **2019**, *228*, 144–163. [CrossRef]
16. Chen, Y.; Zhang, Z.; Tao, F. Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data. *Eur. J. Agron.* **2018**, *101*, 163–173. [CrossRef]
17. Shahhosseini, M.; Martinez-Feria, R.A.; Hu, G.; Archontoulis, S.V. Maize yield and nitrate loss prediction with machine learning algorithms. *Environ. Res. Lett.* **2019**, *14*, 124026. [CrossRef]
18. Felipe Maldaner, L.; de Paula Corrêdo, L.; Fernanda Canata, T.; Paulo Molin, J. Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches. *Comput. Electron. Agric.* **2021**, *181*, 105945. [CrossRef]
19. Peng, B.; Guan, K.; Zhou, W.; Jiang, C.; Frankenberg, C.; Sun, Y.; He, L.; Köhler, P. Assessing the benefit of satellite-based Solar-Induced Chlorophyll Fluorescence in crop yield prediction. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *90*, 102126. [CrossRef]
20. LeCun, Y.; Bengio, Y.; Hinton, G.J.n. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
21. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
22. Pryzant, R.; Ermon, S.; Lobell, D. Monitoring Ethiopian wheat fungus with satellite imagery and deep feature learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 39–47.
23. Ghali, R.; Akhloufi, M.A.; Jmal, M.; Souidene Mseddi, W.; Attia, R. Wildfire Segmentation Using Deep Vision Transformers. *Remote Sens.* **2021**, *13*, 3527. [CrossRef]
24. Alhussein, M.; Aurangzeb, K.; Haider, S.I. Hybrid CNN-LSTM model for short-term individual household load forecasting. *IEEE Access* **2020**, *8*, 180544–180557. [CrossRef]
25. You, J.; Li, X.; Low, M.; Lobell, D.; Ermon, S. Deep gaussian process for crop yield prediction based on remote sensing data. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, CA, USA, 4–9 February 2017; pp. 4559–4565.
26. Wolanin, A.; Mateo-García, G.; Camps-Valls, G.; Gómez-Chova, L.; Meroni, M.; Duveiller, G.; Liangzhi, Y.; Guanter, L. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environ. Res. Lett.* **2020**, *15*, 024019. [CrossRef]
27. Cao, J.; Zhang, Z.; Luo, Y.; Zhang, L.; Zhang, J.; Li, Z.; Tao, F. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. *Eur. J. Agron.* **2021**, *123*, 126204. [CrossRef]
28. Montavon, G.; Samek, W.; Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **2018**, *73*, 1–15. [CrossRef]
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
30. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv* **2019**, arXiv:02860.
31. Wolf, T.; Chaumond, J.; Debut, L.; Sanh, V.; Delangue, C.; Moi, A.; Cistac, P.; Funtowicz, M.; Davison, J.; Shleifer, S. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.

32. Xu, J.; Zhu, Y.; Zhong, R.; Lin, Z.; Xu, J.; Jiang, H.; Huang, J.; Li, H.; Lin, T. DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping. *Remote Sens. Environ.* **2020**, *247*, 111946. [CrossRef]
33. Reedha, R.; Dericquebourg, E.; Canals, R.; Hafiane, A. Vision Transformers For Weeds and Crops Classification Of High Resolution UAV Images. *arXiv* **2021**, arXiv:02716.
34. Xu, J.; Yang, J.; Xiong, X.; Li, H.; Huang, J.; Ting, K.C.; Ying, Y.; Lin, T. Towards interpreting multi-temporal deep learning models in crop mapping. *Remote Sens. Environ.* **2021**, *264*, 112599. [CrossRef]
35. Chauhan, B.S.; Mahajan, G.; Sardana, V.; Timsina, J.; Jat, M.L. Productivity and Sustainability of the Rice–Wheat Cropping System in the Indo-Gangetic Plains of the Indian subcontinent. *Adv. Agron.* **2012**, *117*, 315–369.
36. Pathak, H.; Ladha, J.; Aggarwal, P.; Peng, S.; Das, S.; Singh, Y.; Singh, B.; Kamra, S.; Mishra, B.; Sastri, A. Trends of climatic potential and on-farm yields of rice and wheat in the Indo-Gangetic Plains. *Field Crops Res.* **2003**, *80*, 223–234. [CrossRef]
37. 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]
38. Badgley, G.; Field, C.B.; Berry, J.A. Canopy near-infrared reflectance and terrestrial photosynthesis. *Sci. Adv.* **2017**, *3*, e1602244. [CrossRef] [PubMed]
39. Zhang, Y.; Joiner, J.; Alemohammad, S.H.; Zhou, S.; Gentine, P. A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks. *Biogeosciences* **2018**, *15*, 5779–5800. [CrossRef]
40. Viovy, N. CRUNCEP version 7—Atmospheric forcing data for the community land model, Research Data Archive at the National Center for Atmospheric Research. *Comput. Inf. Syst. Lab. Boulder CO USA* **2018**. [CrossRef]
41. Funk, C.; Peterson, P.; Landsfeld, M.; Pedreros, D.; Verdin, J.; Shukla, S.; Husak, G.; Rowland, J.; Harrison, L.; Hoell, A.; et al. The climate hazards infrared precipitation with stations–a new environmental record for monitoring extremes. *Sci. Data* **2015**, *2*, 150066. [CrossRef]
42. Singh, B.; Gajri, P.; Timsina, J.; Singh, Y.; Dhillon, S. Some issues on water and nitrogen dynamics in rice-wheat sequences on flats and beds in the Indo-Gangetic Plains. *Model. Irrig. Crop. Syst. Spec. Atten. Rice-Wheat Seq.* **2002**, 1–15.
43. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021.
44. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Society. Ser. B* **1996**, *58*, 267–288. [CrossRef]
45. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
46. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
47. Palangi, H.; Li, D.; Shen, Y.; Gao, J.; He, X.; Chen, J.; Song, X.; Ward, R. Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. *Audio Speech Lang. Process. IEEE/ACM Trans.* **2016**, *24*, 694–707. [CrossRef]
48. Jiang, H.; Hu, H.; Zhong, R.; Xu, J.; Xu, J.; Huang, J.; Wang, S.; Ying, Y.; Lin, T. A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US Corn Belt at the county level. *Glob. Chang. Biol.* **2020**, *26*, 1754–1766. [CrossRef] [PubMed]
49. Kang, Y.; Ozdogan, M.; Zhu, X.; Ye, Z.; Hain, C.; Anderson, M. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environ. Res. Lett.* **2020**, *15*, 064005. [CrossRef]
50. Han, Y.; Fan, C.; Xu, M.; Geng, Z.; Zhong, Y. Production capacity analysis and energy saving of complex chemical processes using LSTM based on attention mechanism. *Appl. Therm. Eng.* **2019**, *160*, 114072. [CrossRef]
51. Zhang, X.; Liang, X.; Zhiyuli, A.; Zhang, S.; Xu, R.; Wu, B. AT-LSTM: An Attention-based LSTM Model for Financial Time Series Prediction. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *569*, 052037. [CrossRef]
52. Kobayashi, G.; Kuribayashi, T.; Yokoi, S.; Inui, K. Attention is not only a weight: Analyzing transformers with vector norms. *arXiv* **2020**, arXiv:10102.
53. Li, Y.; Guan, K.; Yu, A.; Peng, B.; Zhao, L.; Li, B.; Peng, J. Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the U.S. *Field Crops Res.* **2019**, *234*, 55–65. [CrossRef]
54. Willmott, C.J.; Ackleson, S.G.; Davis, R.E.; Feddema, J.J.; Klink, K.M.; Legates, D.R.; O'Donnell, J.; Rowe, C.M. Statistics for the evaluation and comparison of models. *J. Geophys. Res.* **1985**, *90*, 8995. [CrossRef]
55. Lin, T.; Zhong, R.; Wang, Y.; Xu, J.; Jiang, H.; Xu, J.; Ying, Y.; Rodriguez, L.; Ting, K.C.; Li, H. DeepCropNet: A deep spatial-temporal learning framework for county-level corn yield estimation. *Environ. Res. Lett.* **2020**, *15*, 034016. [CrossRef]
56. Nguyen, H.T.; Lee, B.-W. Assessment of rice leaf growth and nitrogen status by hyperspectral canopy reflectance and partial least square regression. *Eur. J. Agron.* **2006**, *24*, 349–356. [CrossRef]
57. Mandal, D.; Kumar, V.; Ratha, D.; Lopez-Sanchez, J.M.; Bhattacharya, A.; McNairn, H.; Rao, Y.S.; Ramana, K.V. Assessment of rice growth conditions in a semi-arid region of India using the Generalized Radar Vegetation Index derived from RADARSAT-2 polarimetric SAR data. *Remote Sens. Environ.* **2020**, *237*, 111561. [CrossRef]
58. Sánchez, B.; Rasmussen, A.; Porter, J.R. Temperatures and the growth and development of maize and rice: A review. *Glob. Chang. Biol.* **2014**, *20*, 408–417. [CrossRef]

59. Fahad, S.; Adnan, M.; Hassan, S.; Saud, S.; Hussain, S.; Wu, C.; Wang, D.; Hakeem, K.R.; Alharby, H.F.; Turan, V.; et al. Rice Responses and Tolerance to High Temperature. In *Advances in Rice Research for Abiotic Stress Tolerance*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 201–224.
60. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6405–6416.
61. Wang, X.; Folberth, C.; Skalsky, R.; Wang, S.; Chen, B.; Liu, Y.; Chen, J.; Balkovic, J. Crop calendar optimization for climate change adaptation in rice-based multiple cropping systems of India and Bangladesh. *Agric. For. Meteorol.* **2022**, *315*, 108830. [CrossRef]
62. Jones, J.W.; Hoogenboom, G.; Porter, C.H.; Boote, K.J.; Batchelor, W.D.; Hunt, L.; Wilkens, P.W.; Singh, U.; Gijsman, A.J.; Ritchie, J.T. The DSSAT cropping system model. *Eur. J. Agron.* **2003**, *18*, 235–265. [CrossRef]