agronomy

MDPI

Article

Combining Data Assimilation with Machine Learning to Predict the Regional Daily Leaf Area Index of Summer Maize (*Zea mays* L.)

Yongqiang Wang [1,2], Hui Zhou [3,*], Xiaoyi Ma [2,*] and Hu Liu [3]

[1] Agricultural College, Inner Mongolia Agricultural University, Hohhot 010019, China; wangyongqiang@nwafu.edu.cn
[2] Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas, Ministry of Education, Northwest A&F University, Xianyang 712100, China
[3] Yinshanbeilu Grassland Ecohydrology National Observation and Research Station, China Institute of Water Resources and Hydropower Research, Beijing 100038, China; liuhuycy@163.com
* Correspondence: zhouhui@iwhr.com (H.Z.); xma@nwafu.edu.cn (X.M.)

Abstract: The prediction of the daily crop leaf area index (LAI) plays a crucial role in forecasting crop growth trends and guiding field management decisions in the realm of scientific research. However, research on the daily prediction of LAI is scarce, and the challenges associated with acquiring sufficient training data pose limitations to the application of machine learning in this context. This study aimed to synergize the strengths of data assimilation and machine learning algorithms to forecast the daily LAI of maize. Initially, a data assimilation algorithm was employed to minimize the disparity between moderate-resolution imaging spectroradiometer-derived LAI and LAI generated through the CERES-Maize model. This effort resulted in a dataset comprising 289 LAI curves. Building upon this dataset, long short-term memory (LSTM) networks, support vector regression (SVR), and random forest (RF) algorithms were formulated, incorporating *N*-day LAI input history (N = 5, 10, 15, 20, and 25) to predict LAI for days N + 1 to N + 15. The outcomes revealed that, in contrast to the LAI simulated by the crop model before assimilation, the assimilated LAI closely approximated the observed LAI, with an R^2 value of 0.90 and an RMSE of 0.44 m^2/m^2. Furthermore, when compared to SVR and RF, the LSTM-based LAI prediction model exhibited superior accuracy at N = 15, achieving R^2 values of 0.99 and 0.99 for the training and testing datasets, respectively, along with RMSE values of 0.12 and 0.14 m^2/m^2. It was evident that data assimilation supplied an ample number of samples for the training of machine learning algorithms. The integration of data assimilation technology with machine learning algorithms proved to be an effective methodology for forecasting daily crop LAI.

Keywords: leaf area index; CERES-Maize model; data assimilation; machine learning

check for updates

Citation: Wang, Y.; Zhou, H.; Ma, X.; Liu, H. Combining Data Assimilation with Machine Learning to Predict the Regional Daily Leaf Area Index of Summer Maize (*Zea mays* L.). *Agronomy* **2023**, *13*, 2688. https://doi.org/10.3390/agronomy13112688

Academic Editor: Gniewko Niedbała

Received: 10 October 2023
Revised: 23 October 2023
Accepted: 23 October 2023
Published: 25 October 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The leaf area index (LAI) is defined as the total one-half of green leaves area per unit ground surface area [1]. The LAI is a crucial parameter for evaluating crop growth and estimating crop yield in agricultural production [2]. This index is closely influenced by photosynthesis, biomass accumulation, and other physiological processes of vegetation [3]. Therefore, predicting regional changes in the LAI can enable the assessment of future crop growth and guide decision-making related to field management at the regional scale [4].

Crop model simulation and observation (especially remote sensing observation) are the two main methods for obtaining LAI time series data for crops. Crop models based on mechanisms such as photosynthesis, respiration, and transpiration can accurately simulate the continuous evolution of the LAI of a crop at a certain location [5]. However, the simulation of crop models leads to highly uncertain results for regional LAI time series because many model inputs are poorly understood at the regional scale [6,7]. Moreover,

remote sensing sensors can monitor the growth of crops in areas with border crossings [8]. However, the inversion results of remote sensing data only represent the instantaneous LAI of crops, which is the main obstacle to the application of LAI inversion results in research with time characteristics, such as crop growth monitoring [9]. The aforementioned two methods of obtaining time series LAI data have unique advantages and limitations. Combining crop model simulation and remote sensing observations to achieve complementary advantages has become a research hotspot in the monitoring of regional-scale time series LAI data [10–12].

Data assimilation research involving the combination of crop growth models and instantaneous remote sensing observations has developed rapidly [13]. Jin et al. [4] integrated remotely sensed LAI data into the CERES-Maize model, resulting in a notable enhancement in the spatiotemporal consistency of summer maize LAI. However, most of the aforementioned studies simply "reproduced" the historical growth process of crops by using historical meteorological data measured throughout the crop growth period as the model input. These studies did not predict the future growth process of crops.

Fortunately, machine learning algorithms have the capability to extract patterns from historical time series datasets and, consequently, forecast future time series data [14]. Adaryani et al. [15] developed models for forecasting rainfall depth 5 min and 15 min in advance, leveraging support vector regression (SVR), long short-term memory (LSTM) networks, and convolutional neural networks (CNN). The results indicated that the predictive performance of SVR and LSTM networks surpassed that of CNN. Meanwhile, Alkabbani et al. [16] trained regional wind energy time series forecasting models utilizing artificial neural networks, deep neural networks, LSTM networks, bagged trees, and SVR. The findings demonstrated that the predictive performance of the SVR model excelled in this context. Hu et al. [17] employed the random forest (RF) technique for predicting time series salinity data, and the outcomes revealed that when compared to the traditional RF method, a hybrid RF model integrated with signal decomposition techniques demonstrated enhanced forecasting performance and yielded reasonably precise predictive results. However, it is worth noting that we have not encountered literature employing machine learning algorithms to forecast future time series LAI. Nevertheless, there are instances where scholars have applied machine learning techniques to analyze time series LAI data. Wang et al. [18] evaluated the performance of machine learning methods for LAI retrieval from time series moderate resolution imaging spectroradiometer (MODIS) reflectance data. Pipia et al. [19] employed a combination of optical and synthetic aperture radar time series data to address LAI data gaps using multioutput Gaussian processes. Huang et al. [20] proposed a methodology for reconstructing LAI time series data using a generative adversarial network and an improved Savitzky–Golay (S-G) filter. Their method demonstrated precision in interpolating low-quality LAI data while efficiently preserving high-quality information in the original time series. However, despite the improvements in the spatiotemporal continuity of LAI achieved by these methods, they still fall short of predicting future time series changes in LAI, instead focusing primarily on reproducing historical LAI patterns.

Therefore, we postulated that predicting future time series of LAI based on a historical LAI time series dataset established through data assimilation methods and machine learning algorithms held promise. This study combined the advantages of data assimilation technology, which can be used to establish numerous historical datasets, and LSTM, SVR, and RF models to develop a method for predicting the regional scale daily LAI of maize. The objectives of this study were to (i) investigate the feasibility of using data assimilation methods to establish samples for machine learning models and (ii) compare the ability to use LSTM, SVR, and RF models for predicting the regional scale daily LAI of summer maize.

2. Materials and Methods

Figure 1 displays the flowchart of the combination of data assimilation with the LSTM network, SVR, and RF for predicting the daily LAI of maize. We used the CERES-Maize model to simulate the LAI trends. MODIS LAI time series data were used as remote sensing observations. We used the Shuffled Complex Evolution University of Arizona (SCE-UA) algorithm to iteratively optimize the input parameters of the CERES-Maize model until the difference between the LAI simulated using the CERES-Maize model and the corresponding MODIS LAI was minimized. The CERES-Maize model with optimal input parameters provided 296 LAI curves. These curves were employed for training LSTM, SVR, and RF models to forecast daily LAI for the upcoming 15 days. Concurrently, an assessment and comparison of the predictive capabilities of these three models for daily LAI were conducted. The models and methods used in this study are described in detail in the following text.

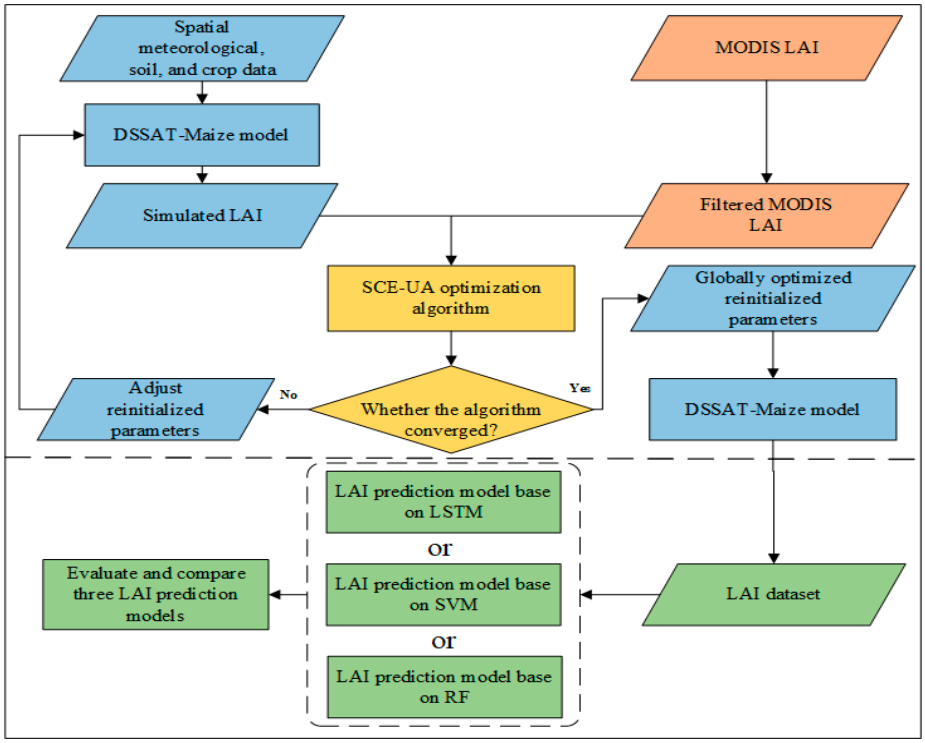


Figure 1. Flowchart of the construction of the daily LAI prediction model.

2.1. Study Area

The research zone covers a substantial portion of the Fenwei River Valley Plain (latitude range: 33.98–36.64° N, longitude range: 107.42–111.89° E; Figure 2). This region exhibits a semiarid climate characterized by an average annual precipitation of 550 mm, an average annual air temperature of 14.4 °C, and an elevation ranging from 350 to 800 m. The predominant land use in this study area is agricultural, with summer maize being one of the primary crops cultivated during the summer season. This crop is generally sowed in the middle of June and harvested in early October.

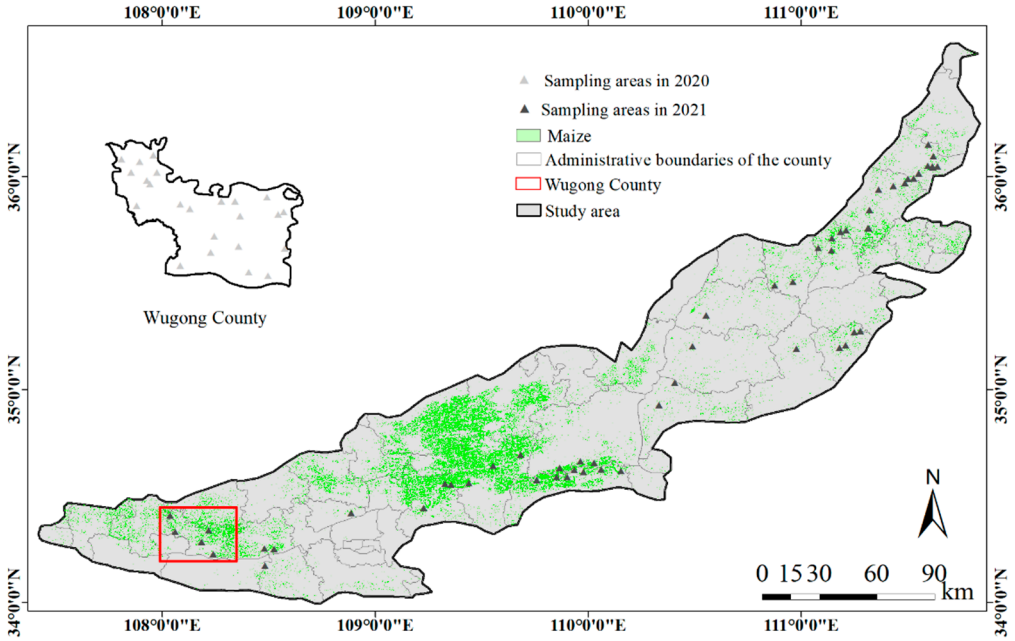


Figure 2. Study area and sampling plot distribution.

2.2. Field Data Collection

In 2020, we established 23 sampling areas measuring 10 m × 10 m in Wugong County in which we could monitor maize growth (Figure 2). These sampling areas were positioned at a distance from both roadways and tree cover. The precise central coordinates for each sampling area were determined with the aid of a global positioning system (GPS) receiver, ensuring their georeferencing compatibility with remote-sensing imagery. In each sampling area, there were three independent sampling points measuring 1 m × 1 m, which were equally spaced within the sampling area. The LAI was measured on 16–18 July, 30 July–1 August, 20–22 August, 6–8 September, and 14–16 September, 2020. Finally, summer maize yields in the 53 sample plots were measured after harvesting in early October 2020. LAI and yield values from the three sampling points were averaged to represent the unique LAI and yield values in each sample area, respectively. In addition, we collected data on the planting date, planting depth, plant spacing, and row spacing for each sampling area.

Similarly, in 2021, we arranged 56 sampling areas measuring 10 m × 10 m in the Fenwei River Valley Plain (Figure 2), at which we could monitor summer maize growth on a larger scale. We measured the LAI on 15–17 July, 26–28 July, 5–7 August, and 23–25 August 2021. Grain yield was measured after the maize had been harvested. We also collected data on the planting date, planting depth, plant spacing, and row spacing for each sampling area. Regarding LAI measurement, the length (from collar to tip) and width (widest point of the leaf) of all leaves of the selected plants were first measured, and the LAI was then calculated.

$$TA = 0.75 \times \sum_{i=1}^{u} l_i \times w_i \quad (1)$$

$$LAI = \frac{PD}{10000} \times \frac{1}{v} \times \sum_{p=1}^{v} TA_p \quad (2)$$

where TA represents the total leaf area of an individual maize plant, measured in square meters, with l and w denoting the length and width of the leaf in meters, respectively. u is the number of leaves in a single maize plant. The coefficient 0.75 represents the relationship between the measured leaf area and the area of a rectangle defined by the product of l and w, and this coefficient has been widely accepted by the majority of maize agronomists [21–23]. v is the number of selected maize plants, and PD is the planting density in plants per hectare.

2.3. Input Data for the CERES-Maize

CERES-Maize model simulation requires daily meteorological data, soil data, crop management information, and crop parameters. In this study, meteorological data, including daily maximum temperature, minimum temperature, solar radiation, and rainfall data, were obtained from the China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn/; accessed on 15 February 2022). In order to adapt the CERES-Maize model for regional analysis, it is imperative to transform meteorological data from a single-point scale to a spatial scale. As a solution, we employed ordinary Kriging interpolation to estimate the values of the aforementioned weather variables across a grid measuring 5 km × 5 km. Likewise, soil data in gridded format, with a resolution of 5 arc min, were acquired from the reference study [24]. To keep the resolution of the soil data consistent with the resolution of the meteorological data, we used inverse distance weighted interpolation on the scale of 5 km × 5 km to generate the grid values of the soil data. In addition, the crop management information, including planting date, planting depth and spacing, and planting density, were collected from the field (Section 2.2). Considering the suitable water and fertilizer conditions during the summer maize growing season in the study area, the CERES-Maize model was used to simulate the summer maize growth in the study area at the potential productivity level.

The DSSAT-CERES-Maize model incorporates three categories of crop parameters that influence the developmental process: cultivar-type, ecotype, and species-type parameters [25]. In accordance with the standard guidelines for utilizing the CERES-Maize model, experienced model users have the flexibility to make adjustments to the ecotype parameters, whereas the species-type parameters are recommended to remain unaltered for regular model users. The cultivar-type genetic parameters are calibrated based on observational data. Hence, six parameters (Table 1) were calibrated in this study.

Table 1. Maize cultivar coefficients.

Acronym	Unit	Min	Max
P1	(°C·d)	130	380
P2	-	0	4
P5	(°C·d)	600	1100
G2	number	400	1100
G3	(mg/day)	4	11.5
PHINT	(°C·d)	30	90

Note: P1 represents thermal time from emergence to end of the juvenile phase. P2 represents photoperiod sensitivity coefficient. P5 represents thermal time from silking to physiological maturity. G2 represents maximum possible number of kernels per plant. G3 represents kernel filling rate during the linear grain filling stage and under optimum conditions. PHINT represents interval in thermal time (degree days) between successive leaf tip appearances.

2.4. MODIS LAI Dataset

MODIS have the advantages of short revisit intervals and wide geographical coverage, which are conducive for capturing the growth characteristics and spatiotemporal variability of crops. In this study, we adopted the MODIS LAI dataset (https://lpdaac.usgs.gov/products/mcd15a3hv006/; accessed on 15 February 2022) for the study area from June to October 2020 and from June to October 2021. This dataset has a spatial resolution of 500 m. According to the morphological characteristics of the LAI curve of summer maize and the Google Earth image of the study area, we identified 289 maize pixels and extracted 289 LAI curves in the MODIS LAI dataset (79 LAI curves were derived from summer maize sampling areas in 2020 and 2021, while the remaining 210 LAI curves were obtained from non-sampling areas. Field investigations revealed that the central locations of all 210 maize pixels were situated within the summer maize cultivation areas). Because of cloud and aerosol interference, the MODIS LAI time series exhibits a zigzag pattern. To mitigate this, we applied an upper-envelope-based S-G filter to smooth the MODIS LAI curve. To obtain

the upper envelope of the time series of MODIS LAI for a particular pixel, the following steps were as follows [26,27]:

Step 1: Set the initial time series of MODIS LAI as LAI_a^0, where a = 1, 2, 3, ..., A;

Step 2: Apply the S-G algorithm to filter LAI_a^0, resulting in the time series of LAI, denoted as LAI_a^{sg};

Step 3: Construct a new time series of LAI, denoted as LAI_a^{new}, using Equation (3).

$$LAI_a^{new} = \begin{cases} LAI_a^{sg}, when\ LAI_a^{sg} \geq LAI_a^0 \\ LAI_a^0, when\ LAI_a^{sg} \leq LAI_a^0 \end{cases}, a = 1, 2, 3, \ldots, A \quad (3)$$

Step 4: Apply the S-G algorithm to filter LAI_a^{new}, resulting in the time series of LAI, denoted as LAI_a^{sg};

Step 5: Calculate the fitting effect index *dif* using Equation (4),

$$dif = \sum_{a=1}^{A} \left(\left| LAI_a^{sg} - LAI_a^0 \right| \times Q_a \right) \quad (4)$$

where Q_a represents the weight of each LAI in the time series. If *dif* was reached the predefined threshold, then that obtained in Step 4 was considered as the upper envelope of MODIS LAI. Otherwise, return to Step 3.

2.5. Construction of an LAI Dataset Based on the Data Assimilation Framework

Data assimilation is a crucial method that integrates two fundamental scientific research tools: observations and models [28]. It allows for the integration of multi-source, spatiotemporally discontinuous observational data into an evolving crop model, thereby enabling a more accurate simulation of crop growth processes. There are two primary categories of data assimilation methods: parameter optimization methods based on cost functions and filtering methods based on estimation theory. Among these, parameter optimization methods iteratively adjust parameters or initial conditions within the crop model that are closely related to growth and yield formation, which are often challenging to obtain through conventional means. The objective is to minimize the disparity between observed values and model-simulated values, ultimately achieving model calibration. In this paper, the data assimilation framework based on cost function-based parameter optimization is primarily comprised of three key components: the crop model, remote sensing observations (Section 2.4), and assimilation algorithms.

2.5.1. CERES-Maize Model

The Decision Support System for Agrotechnology Transfer (DSSAT), as referenced in [29], stands out as one of the most extensively employed crop model systems. This system effectively replicates the intricate physiological processes inherent to crop development, encompassing photosynthesis, respiration, and the allocation of dry matter. Specifically, it leverages the CERES-Maize model [25] to simulate the growth of maize. Within the framework of the CERES-Maize model, maize growth rate is contingent upon temperature and photoperiod. The daily minimum temperature plays a pivotal role in influencing biomass allocation toward grains. In the case of maize, the number of grains per plant depends on both the number of potential kernels per plant and the average crop growth rate during the silking-to-grain filling stage. Notably, when the number of kernels per plant significantly lags behind the potential number, the CERES-Maize model accounts for barren plants [30].

According to the general requirements for using the DSSAT model, the user must calibrate the cultivar coefficients (Table 1) by using field observation data. Therefore, we first utilized a genetic algorithm to calibrate the six cultivar coefficients using field-scale-measured data (i.e., phenological period, LAI, and yield) from 2016. Field-scale-measured data from 2017 was used to validate the six coefficients [31].

2.5.2. Integrating a Shuffled Complex Evolution University of Arizona (SCE-UA) Optimization Algorithm for Data Assimilation

In this study, a cost function was formulated to measure the disparity between MODIS-derived LAI and LAI predictions generated by the CERES-Maize model. Due to the confounding influence of mixed information, clouds, and rainfall at the 500 m spatial resolution, the MODIS LAI values frequently register as considerably lower than those simulated by the CERES-Maize model. Consequently, we opted to employ normalized LAI values rather than unadjusted LAI values in crafting the cost function.

$$LAI_{nor} = \frac{LAI_t - LAI_{\min}}{LAI_{\max} - LAI_{\min}} \quad (5)$$

where LAI_{nor} represents normalized LAI, LAI_t represents MODIS or CERES-Maize simulated LAI on Day t, and LAI_{max} and LAI_{min} represent the maximum and minimum LAI of MODIS or CERES-Maize simulated LAI during the crop growing season, respectively.

The selection of parameters that need to be reinitialized is pivotal in a cost function. According to the results of parameter sensitivity analysis [4], the parameters P2 and P5 have a greater influence on LAI. Therefore, only P2 and P5 were chosen as the reinitialization parameters, while P1, G2, G3, and PHINT will remain unchanged (Section 2.5.1). Therefore, the cost function was constructed as Equation (6):

$$J(X) = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{(LAI^k_{nor,CERES-Maize}(X) - LAI^k_{nor,MODIS})}{LAI^k_{nor,MODIS}}\right| \quad (6)$$

where J is the value of the cost function, X is a vector composed of parameters to be optimized, $LAI_{nor,CERES-Maize}$ is the LAI value simulated using the CERES-Maize model, $LAI_{nor,MODIS}$ is the LAI value derived from the remotely sensed data, and K represents the total number of LAI values derived from the remotely sensed data.

In pursuit of the most suitable parameter sets, we employed the SCE–UA optimization algorithm [32] with the aim of minimizing the cost function J by systematically adjusting the parameters within the CERES-Maize model. The SCE-UA algorithm is a global optimization algorithm that combines the advantages of deterministic search, random search, and competitive evolution. Under a multiparameter combination, its global search performance and calculation efficiency are excellent. In addition, the SCE-UA algorithm is insensitive to the initial values of the optimized parameters; thus, limited known prior knowledge is required when using the SCE-UA algorithm to combine remote sensing and crop growth models [33]. In this study, the SCE-UA algorithm was considered to be convergent, and the cost function value did not improve by more than 0.1% after 10 iterations.

2.6. Construction of Time Series Forecasting Model for Leaf Area Index

Short-term predictions of the LAI of summer maize cannot effectively forecast its growth status, whereas long-term predictions of the LAI of summer maize result in a decrease in the accuracy of the LAI prediction model due to the lack of input information. Therefore, this study selected an appropriate number of days to predict the LAI. LAI data for the previous N days (N = 5, 10, 15, 20, and 25) were used to predict LAI time series data for the subsequent N + 1 to N + 15 days. In addition, the phenology of summer maize is different under different weather conditions. Therefore, in addition to the LAI, we used the Julian day (Jd) corresponding to the LAI as an input for training the LSTM network, SVR, and RF, respectively. This is because LSTM excels in handling data with long-term dependencies, SVR demonstrates strong robustness in small sample regression tasks, and RF exhibits excellent resistance to noise and outliers in the data. To ensure the consistency of the input parameters, we scaled the LAI and its corresponding Jd to 0–1. Moreover, when constructing the above three models, 80% of the dataset was used as the training set, and 20% of the dataset was used as the test set.

2.6.1. Long Short-Term Memory Network (LSTM)

The core of LSTM networks [34] is a memory cell that stores the information state (Figure 3). Three gates, namely the input, output, and forget gates, are used to add information into or remove information from the memory cell. At time t, the LSTM unit's computation process can be dissected into the subsequent stages.

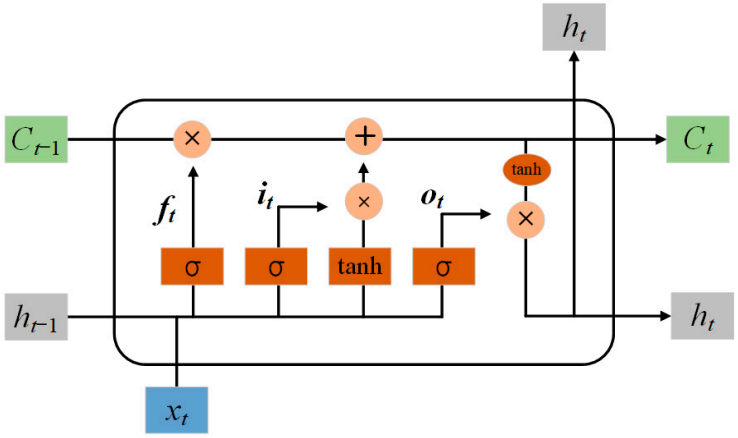


Figure 3. Architecture of the memory cell of an LSTM network.

(1) Begin by computing the candidate memory cell vector, denoted as $\widetilde{c}_t$.

$$\widetilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{7}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{8}$$

where tanh signifies the hyperbolic tangent function, W_c stands for the weight matrix, h_{t-1} represents the previous LSTM cell output vector, x_t denotes the current input vector of the LSTM cell, and b_c is the deviation vector.

(2) Next, determine the input gate i_t, which regulates the update of the current input vector into the memory cell's state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{9}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{10}$$

where σ corresponds to the sigmoid function, W_i is the weight matrix, and b_i serves as the deviation vector.

(3) Similarly, evaluate the forget gate f_t, which governs the updating of historical data within the memory cell's state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{11}$$

where W_f represents the weight matrix, and b_f is the deviation vector.

(4) Proceed to compute the state vector of the memory cell at the current time step c_t.

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c}_t \tag{12}$$

where * denotes the element-wise dot product of the matrix, and c_{t-1} refers to the state vector of the last LSTM unit.

(5) Calculate the output gate o_t, which is responsible for controlling the output of the memory cell's state.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{13}$$

where W_o signifies the weight matrix, and b_o is the deviation vector.

(6) Finally, derive the current output vector of the LSTM unit h_t.

$$h_t = o_t * \tanh(c_t) \tag{14}$$

The aim of this study was to predict LAI time series data, and the mean square error (MSE) was selected as the loss function. The loss L is defined as follows:

$$L(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(y^i - f_\theta(x^i)\right)^2 \tag{15}$$

where the loss of a single sample is the square of the difference between the LSTM output value f_θ (x^i) and the target output value y^i, m is the number of samples, and θ is the weight parameter for LSTM learning. As displayed in Figure 4, the developed prediction models had a three-layer LSTM structure. The number of hidden neurons in each layer was set as 160; the input parameters were a tuple (LAI_j, Jd_j; j = 1, 2, . . ., N); and the final output of the models was $y' = \{LAI_{N+1}, LAI_{N+2}, \cdots, LAI_{N+15}\}$. The parameter LAI_j represents the LAI information input into an LSTM network, and LAI_{N+1}, LAI_{N+2}, . . ., and LAI_{N+15} represent the LAI values predicted by the LSTM network.

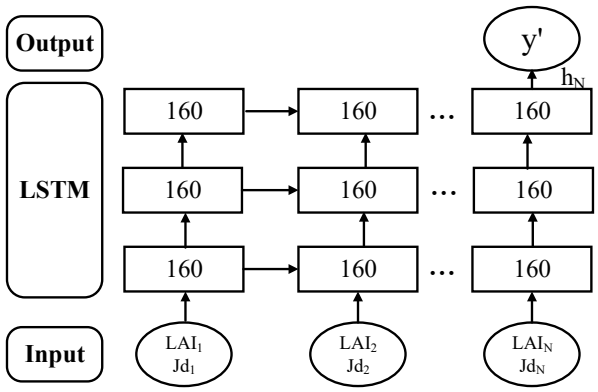


Figure 4. LSTM architecture of the developed LAI prediction model (Jd: Julian day).

2.6.2. Support Vector Regression (SVR)

SVR [35] models have gained widespread application for estimating crop LAI [36]. Notably, these SVR models demonstrate robust performance, particularly when confronted with limited training data. The SVR model utilized in this study was obtained from the LIBSVM machine learning library, accessible at https://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html (accessed on 15 May 2022).

The chosen kernel function for this model is the radial basis function. The crucial hyperparameters, namely the penalty factor C and the parameter γ, were determined using a grid search (GS) technique. This GS procedure involves specifying the search range for the model, partitioning feasible parameter values into predefined intervals, and subsequently generating a grid of parameter combinations. The steps undertaken to ascertain the optimal hyperparameters were as follows (Figure 5):

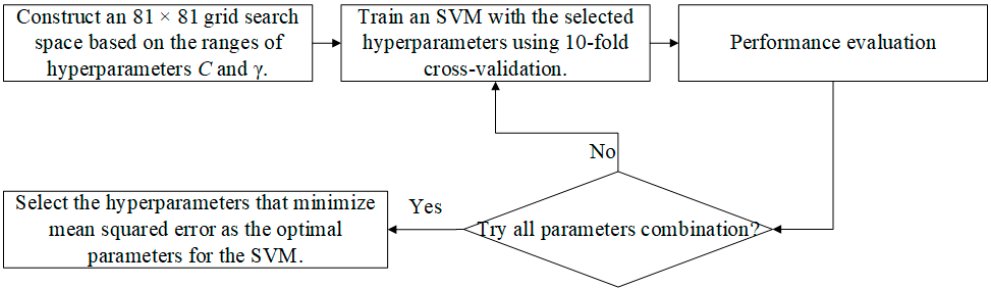


Figure 5. Flowchart for optimizing hyperparameters in support vector regression.

In Step 1, a range was established for the hyperparameters C and γ to determine their optimal values. The intervals for both C and γ were defined as $[2^{-10}, 2^{10}]$, with a step size of $2^{0.25}$ [37], leading to the creation of an 81 × 81 grid.

Subsequently, in Step 2, a 10-fold cross-validation approach was employed. From the pool of 289 sample datasets, 230 were selected as the training dataset (80% of the total dataset) [38]. This training dataset was evenly divided into 10 subsets, each taking its turn as the validation set. Model performance was evaluated using the remaining nine subsets, and the hyperparameter values that yielded the lowest mean squared error were chosen as the optimal settings through this cross-validation process.

In Step 3, a GS was conducted iteratively. Sequentially, new combinations of hyperparameters were selected from the grid, and Step 2 was repeated. This iterative process was repeated until the combination of hyperparameters that produced the smallest mean squared error was identified. These specific hyperparameter values were then considered as the optimal configuration for the SVR model.

2.6.3. Random Forest (RF)

RF is a regression technique that combines the performance of numerous decision tree algorithms to predict the value of a variable [39]. That is when RF receives an (x) input vector made up of the values of the different evidential features analyzed for a given training area, RF builds a number M of regression trees and averages the results. After M such trees $\{T(x)\}_1^M$ are grown, the RF regression predictor is

$$\hat{f}(x) = \frac{1}{M}\sum_{m}^{M} T(x) \tag{16}$$

In order to mitigate inter-tree correlations, the RF method enhances diversity among its constituent trees by having them grow from distinct training data subsets generated through a technique known as bagging. Bagging involves the creation of training data through random resampling of the original dataset with replacement. This approach results in some data points being employed multiple times during training, while others may not be used at all. This strategy enhances the model's stability, rendering it more robust when confronted with minor fluctuations in input data while simultaneously elevating prediction accuracy. Conversely, during the growth of each tree within the RF, the algorithm selects the best feature/split point from a randomly chosen subset of relevant features extracted from the overall set of input features. Consequently, this practice may diminish the strength of individual trees but contributes to a reduction in inter-tree correlation, ultimately diminishing the generalization error.

Furthermore, the samples that are excluded from the training dataset for the m-th tree within the bagging process are designated as part of a separate subset known as the out-of-bag (oob) set. These oob samples can be employed by the m-th tree to assess its performance. RF also offers an evaluation of the relative significance of various evidential attributes. This particular facet proves valuable in multi-source investigations, especially when dealing with high data dimensionality. Understanding the impact of each feature on

the predictive model is crucial for selecting the most relevant evidential attributes [40]. In order to gauge the importance of each variable, RF conducts a process where one input evidential feature is altered while all others are held constant. Subsequently, it measures the reduction in accuracy through the use of oob error estimation [39].

2.7. Statistics for Model Performance Evaluation

In this study, we used the coefficient of determination (R^2, [41]) and root mean square error (RMSE, [42]) as indicators to assess the performance of our models. R^2 ranges between 0 and 1, where a value closer to 1 signifies a superior fit of the model to the data, while a value closer to 0 indicates a poorer fit. This intuitive scale makes R^2 easy to comprehend and interpret. The RMSE indicates the accuracy of the model in terms of the deviation between the measured and predicted values. The smaller the RMSE, the smaller the difference between the measured and simulated values. The RMSE has been established as a primary metric in prior research [43]. The mathematical expressions of R^2 and RMSE are as follows:

$$R^2 = \frac{\sum_{i=1}^{n} (O_i - \overline{O})^2 (S_i - \overline{S})^2}{\sum_{i=1}^{n} (O_i - \overline{O})^2 \sum_{i=1}^{n} (S_i - \overline{S})^2} \quad (17)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (O_i - S_i)^2}{n}} \quad (18)$$

where O_i is the observed LAI value, S_i is the simulated LAI value, $\overline{O}$ is the average observed value, $\overline{S}$ is the average simulated value, and n represents the number of samples.

3. Results and Discussion

3.1. Generation of the S-G Filtered MODIS LAI

We utilized a 4-day MODIS LAI dataset spanning from June to September in both 2020 and 2021. Figure 6 illustrates the typical growth stages of summer maize in the Fenwei Valley Plain, along with a comparative analysis involving MODIS LAI, S-G filtered LAI, and CERES-Maize simulated LAI. The MODIS LAI data is susceptible to distortions caused by cloud cover and aerosol contamination, resulting in a serrated appearance. To mitigate this, we applied an upper-envelope-based S-G filter method to eliminate outliers in the MODIS LAI dataset. This refinement process yielded a smoother time series of S-G filtered LAI, which was subsequently employed in the assimilation process with the CERES-Maize model. It is important to note that the S-G filtered LAI tends to exhibit lower values in comparison to the CERES-Maize simulated LAI. This discrepancy can be attributed to the relatively coarse spatial resolution of MODIS data (500 m), which combines information from various land types, contributing to the observed differences.

3.2. Comparison of the LAI before and after Data Assimilation

One randomly selected sampling point was used as an example to assess the accuracy of LAI simulation from the CERES-Maize model before and after data assimilation (Figure 7a). The LAI curve simulated by the model before assimilation lay below the LAI curve after assimilation, indicating that the assimilated LAI values were closer to the observed ones. Furthermore, the accuracy of LAI simulations by the crop model before and after assimilation at all sampling points was compared (Figure 7b). As depicted in the figures, before assimilation, the LAI simulated by the crop model was noticeably lower than the observed LAI, with a corresponding R^2 of 0.55 and RMSE of 1.17 m^2/m^2. Assimilating MODIS LAI significantly improved the crop model's LAI simulation capability, resulting in an R^2 of 0.90 and RMSE of 0.44 m^2/m^2, with data points clustering closer to the 1:1 line. The above results indicate that data assimilation can reduce the error in simulating LAI

in crop models. This conclusion is consistent with previous studies [4,26]. The primary reason is that remotely sensed LAI reflects actual crop growth information, particularly aspects such as irrigation and fertilization, which might not have been comprehensively accounted for in crop models.

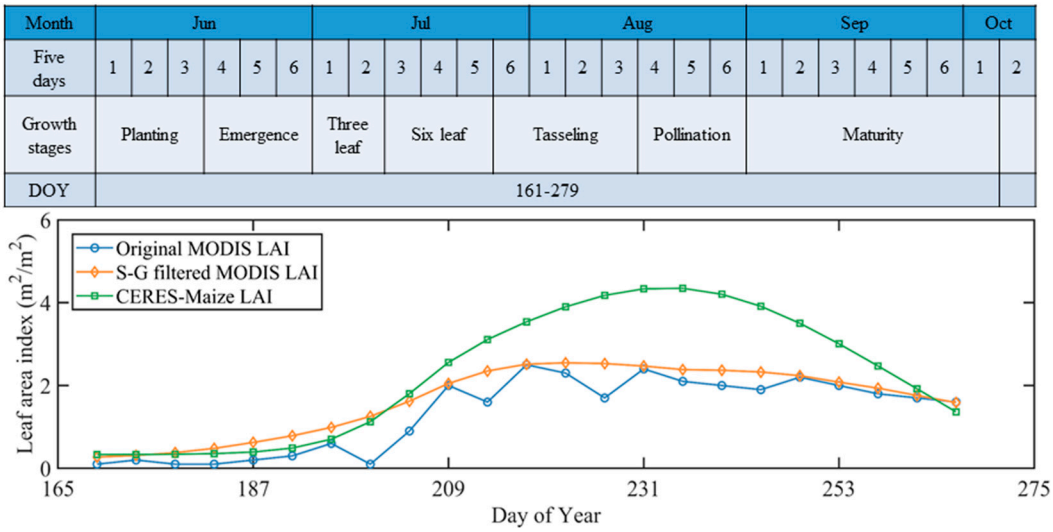


Figure 6. General growth stages of summer maize in Fenwei Valley Plain, and the temporal comparison of MODIS LAI, S-G filtered LAI, and CERES-Maize simulated LAI.

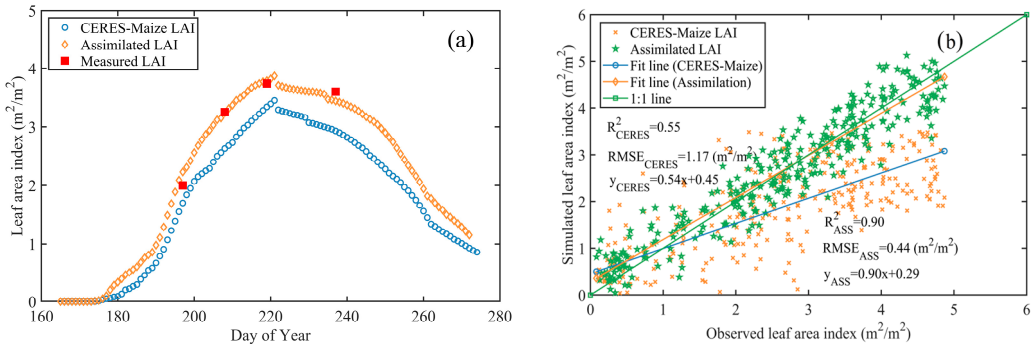


Figure 7. Comparison of the CERES-Maize simulated LAI with and without data assimilation at the point scale. Note: (**a**) is a randomly selected point, and (**b**) is the result of LAI for all sampling points.

3.3. Analysis and Comparison of Accuracy in Different Leaf Area Index Prediction Models

The assimilated time series of LAI were separately fed into LSTM, SVR, and RF algorithms. We conducted a comparative analysis of the accuracy of these three machine learning algorithms in predicting LAI for $N + 1$ to $N + 15$ days, where N took on values of 5, 10, 15, 20, and 25 (Table 2). As N increased, the accuracy of LAI predictions in both the training and testing sets initially improved and then declined. In the study conducted by Wen et al. [44], similar findings were also observed. This is because when training with longer sequences, the algorithms have a greater capacity to capture the underlying temporal patterns in the data. However, this increased capacity can also make the models more susceptible to overfitting. Overfitting occurs when the models start to memorize noise and random fluctuations in the data rather than generalizing meaningful patterns [45]. As a result, the predictive accuracy may decrease as the model becomes overly complex and loses its ability to generalize. At $N = 15$, all algorithms achieved their highest predictive accuracy on both the training and testing datasets.

Table 2. Comparison of the performance of the LAI prediction models.

Model	Data Set	Input Length of LAI Sequence (N)									
		5		10		15		20		25	
		R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
LSTM	Training	0.75	0.69	0.85	0.5	0.99	0.12	0.87	0.45	0.80	0.65
	Testing	0.74	0.71	0.84	0.52	0.99	0.14	0.86	0.46	0.79	0.67
SVM	Training	0.7	0.75	0.81	0.64	0.97	0.20	0.82	0.60	0.72	0.77
	Testing	0.69	0.77	0.80	0.66	0.97	0.21	0.81	0.63	0.71	0.79
RF	Training	0.66	0.83	0.78	0.68	0.92	0.32	0.8	0.66	0.73	0.72
	Testing	0.65	0.84	0.76	0.7	0.92	0.33	0.78	0.70	0.72	0.73

When using short (N = 5) or long (N = 25) sequences of LAI to predict future LAI values, all algorithms exhibited a decrease in predictive accuracy. Furthermore, in comparison to SVR and RF, the LSTM-based LAI prediction model achieved the highest accuracy at N = 15, with R^2 values of 0.99 and 0.99 for the training and testing datasets, respectively, and RMSE values of 0.12 and 0.14 m^2/m^2.

Thus, Figure 8 further demonstrated the accuracy of LSTM networks, SVR, and RF algorithms in predicting LAI for N + 1 to N + 15 days when N was set to 15. Despite the strong correlation between the assimilated LAI and the predictions made by the three machine learning algorithms (with data points evenly distributed on both sides of the 1:1 line), the LSTM-based LAI prediction model exhibited the smallest RMSE. Specifically, in the training dataset, LSTM's RMSE was 0.08 m^2/m^2 and 0.20 m^2/m^2 lower than that of SVR and RF, respectively. In the testing dataset, LSTM's RMSE was 0.07 m^2/m^2 and 0.19 m^2/m^2 lower than that of SVM and RF, respectively.

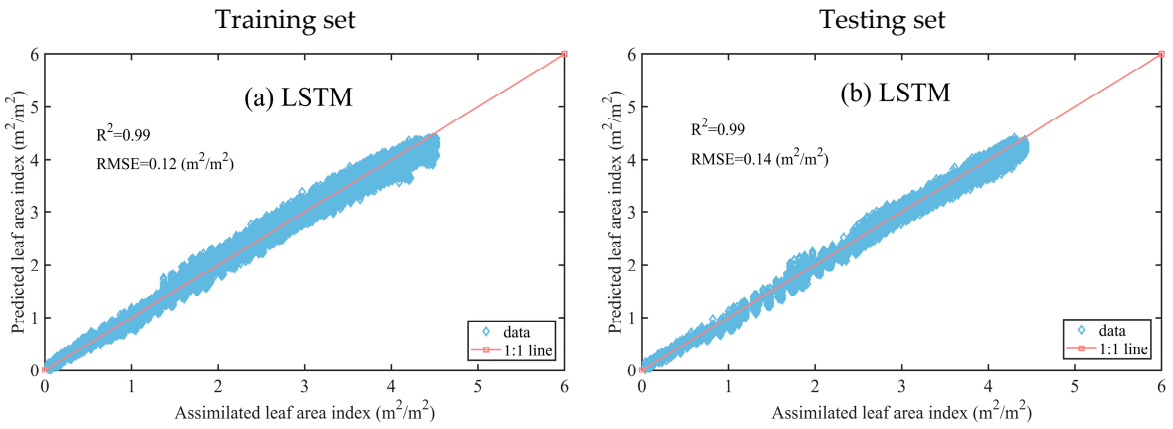


Figure 8. *Cont.*

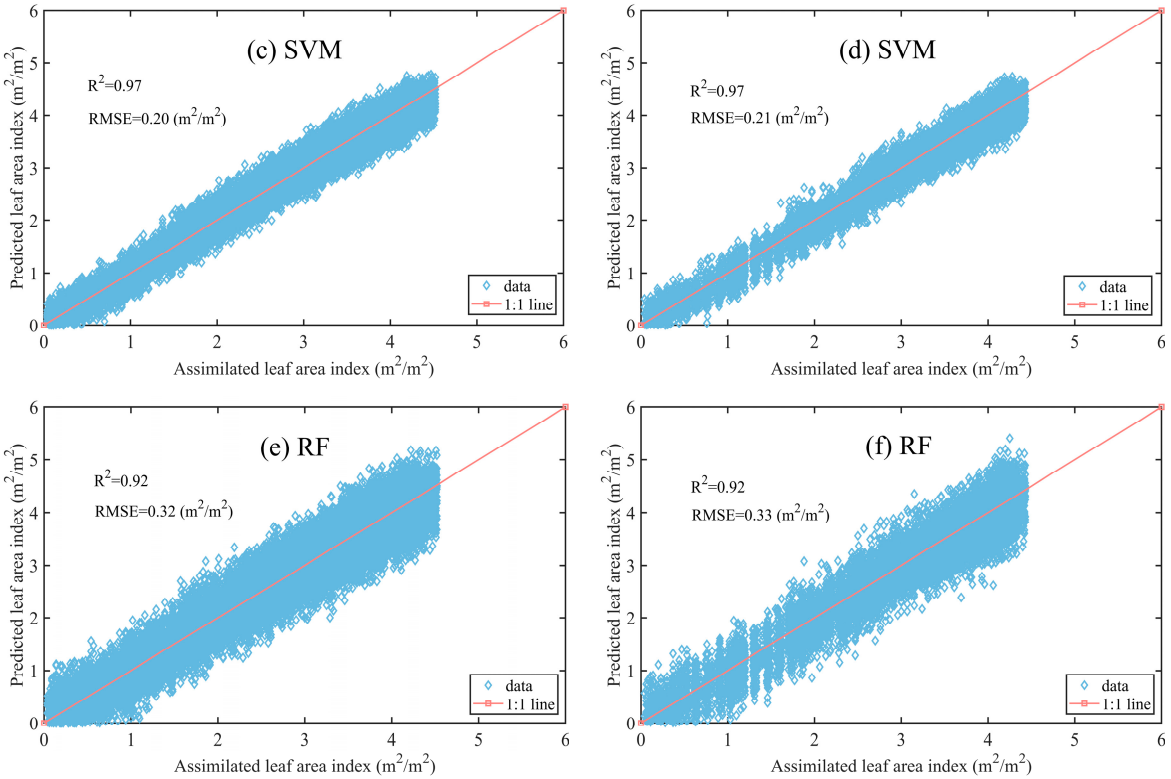


Figure 8. Results of the LAI prediction model with N = 15.

3.4. Predictive Capability of Different Machine Learning Models for LAI at Various Growth Stages

Furthermore, the predictive capabilities of the LSTM model, SVR model, and RF model for three different periods of summer corn LAI were compared (Figure 9). Figure 8b illustrates the method used to select three LAI curves. Initially, the maximum LAI for each LAI time series curve during the respective period was determined. These maximum LAI values were then sorted in descending order, and the LAI time series curves corresponding to the 25th, 50th, and 75th percentiles of these maximum LAI values were selected. Similar methods were employed to select LAI curves for the other two periods. During the rapid increase in LAI (Figure 9a), at the point of maximum LAI attainment (Figure 9b), and during the LAI decline (Figure 9c), LSTM predictions exhibited good consistency with the assimilated LAI. However, the SVR (Figure 9d–f) and RF (Figure 9g–i) models displayed larger errors in predicting the LAI time series for these three periods.

The training dataset used to predict the LAI of summer maize was simulated using the CERES-Maize model under the potential productivity level. However, the growth and development of maize is a complex process influenced by multiple environmental factors, including drought, pest, and nutrient stress. Future studies should expand the training dataset and consider the factors affecting the growth and development of summer maize to improve the universality of the developed LAI prediction models.

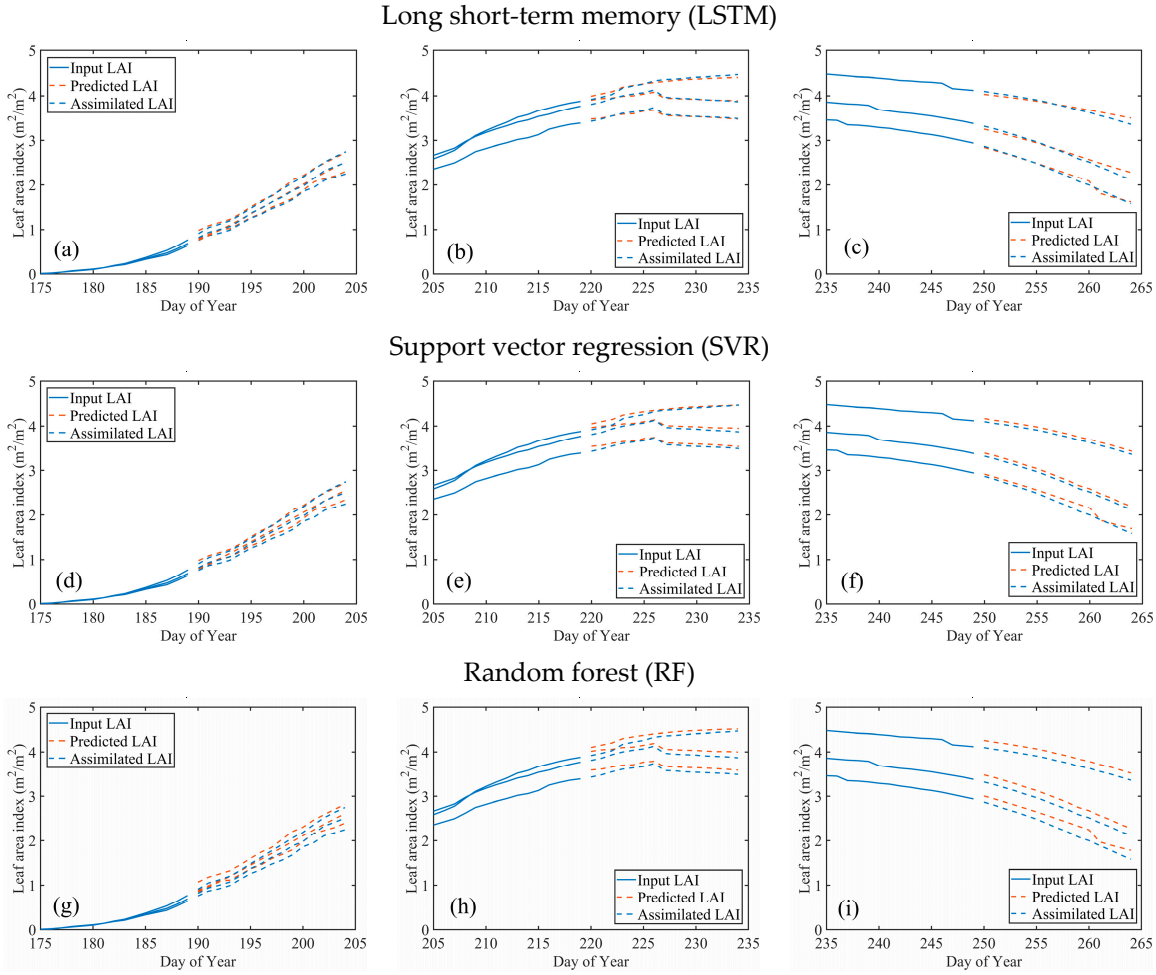


Figure 9. Results of the optimal leaf area index (LAI) prediction model for each maize growth stage. In the context of three distinct time periods, (**a**–**c**) represent the capabilities of LSTM in predicting LAI, while (**d**–**f**) correspond to the capabilities of SVR in predicting LAI during these three time periods. Finally, (**g**–**i**) denote the capabilities of RF in predicting LAI across the same three time periods.

4. Conclusions

This study aimed to synergize the strengths of data assimilation and machine learning algorithms to forecast the daily LAI of maize. Initially, a data assimilation algorithm was employed to minimize the disparity between MODIS-derived LAI and LAI generated through the CERES-Maize model. This effort resulted in a dataset comprising 289 LAI curves. Building upon this dataset, long short-term memory (LSTM) networks, support vector regression (SVR), and random forest (RF) algorithms were formulated, incorporating N-day LAI input history (N = 5, 10, 15, 20, and 25) to predict LAI for days N + 1 to N + 15. The following conclusions are obtained from this study:

(1) In contrast to the LAI simulated by the crop model before assimilation, the assimilated LAI closely approximated the observed LAI, with an R^2 value of 0.90 and an RMSE of 0.44 m^2/m^2. (2) When compared to SVR and RF, the LSTM-based LAI prediction model exhibited superior accuracy at N = 15, achieving R^2 values of 0.99 and 0.99 for the training and testing datasets, respectively, along with RMSE values of 0.12 and 0.14 m^2/m^2. (3) Second, the length of the input LAI time series influenced the prediction performance of the developed LSTM, SVR, and RF prediction models. When the input LAI times series included data from dates that were too far from or too close to the prediction date, the performance of the LAI prediction models worsened.

The issue of scale effects consistently arises in the context of data assimilation research for crop modeling. Agricultural landscapes in various regions often exhibit a scattered and

fragmented nature, resulting in substantial uncertainties when utilizing sensors with low spatial resolution (>250 m) for retrieving crop-related data due to the amalgamation of land surface information [7]. Furthermore, crop models, originally devised for field-scale applications, inherently differ in scale from remotely sensed observations. In this investigation, we employed MODIS LAI data with a spatial resolution of 500 m, which presents scale-related challenges at the field level. Hence, for forthcoming research endeavors, high-resolution observations (e.g., 5–30 m) are imperative to acquire precise crop parameters and effectively mitigate the spatial disparities between crop models and observational data.

Overall, it was evident that data assimilation supplied an ample number of samples for the training of machine learning algorithms. The integration of data assimilation technology with machine learning algorithms proved to be an effective methodology for forecasting daily crop LAI.

Author Contributions: Conceptualization, Y.W., H.Z. and X.M.; data curation, Y.W. and H.L.; methodology, Y.W. and H.Z.; writing—original draft, Y.W.; writing—review and editing, H.Z. and X.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by a special funding project of IWHR (MK2022J12), Projects for the Central Government to Guide Local Scientific and Technological Development (2021ZY0031), Science and Technology Plan Program of Inner Mongolia Autonomous Region (2021GG0060), and the National Key R&D Program of China (No. 2017YFD1900600).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Qiao, L.; Gao, D.; Zhao, R.; Tang, W.; An, L.; Li, M.; Sun, H. Improving estimation of LAI dynamic by fusion of morphological and vegetation indices based on UAV imagery. *Comput. Electron. Agric.* **2022**, *192*, 106603. [CrossRef]
2. Peng, Y.; Li, Y.; Dai, C.; Fang, S.; Gong, Y.; Wu, X.; Zhu, R.; Liu, K. Remote prediction of yield based on LAI estimation in oilseed rape under different planting methods and nitrogen fertilizer applications. *Agric. For. Meteorol.* **2019**, *271*, 116–125. [CrossRef]
3. Parker, G.G. Tamm review: Leaf Area Index (LAI) is both a determinant and a consequence of important processes in vegetation canopies. *For. Ecol. Manag.* **2020**, *477*, 118496. [CrossRef]
4. Jin, H.; Li, A.; Wang, J.; Bo, Y. Improvement of spatially and temporally continuous crop leaf area index by integration of CERES-Maize model and MODIS data. *Eur. J. Agron.* **2016**, *78*, 1–12. [CrossRef]
5. Adnan, A.; Diels, J.; Jibrin, J.; Kamara, A.; Shaibu, A.; Craufurd, P.; Menkir, A. CERES-Maize model for simulating genotype-by-environment interaction of maize and its stability in the dry and wet savannas of Nigeria. *Field Crops Res.* **2020**, *253*, 107826. [CrossRef]
6. Abebe, G.; Tadesse, T.; Gessesse, B. Assimilation of leaf Area Index from multisource earth observation data into the WOFOST model for sugarcane yield estimation. *Int. J. Remote Sens.* **2022**, *43*, 698–720. [CrossRef]
7. Jin, X.; Kumar, L.; Li, Z.; Feng, H.; Xu, X.; Yang, G.; Wang, J. A review of data assimilation of remote sensing and crop models. *Eur. J. Agron.* **2018**, *92*, 141–152. [CrossRef]
8. Lin, W.; Yuan, H.; Dong, W.; Zhang, S.; Liu, S.; Wei, N.; Lu, X.; Wei, Z.; Hu, Y.; Dai, Y. Reprocessed MODIS Version 6.1 Leaf Area Index Dataset and Its Evaluation for Land Surface and Climate Modeling. *Remote Sens.* **2023**, *15*, 1780. [CrossRef]
9. Ovakoglou, G.; Alexandridis, T.K.; Clevers, J.G.; Gitas, I.Z. Downscaling of MODIS leaf area index using landsat vegetation index. *Geocarto Int.* **2022**, *37*, 2466–2489. [CrossRef]
10. Han, C.; Zhang, B.; Chen, H.; Liu, Y.; Wei, Z. Novel approach of upscaling the FAO AquaCrop model into regional scale by using distributed crop parameters derived from remote sensing data. *Agric. Water Manag.* **2020**, *240*, 106288. [CrossRef]
11. Yeom, J.-M.; Jeong, S.; Deo, R.C.; Ko, J. Mapping rice area and yield in northeastern asia by incorporating a crop model with dense vegetation index profiles from a geostationary satellite. *GIScience Remote Sens.* **2021**, *58*, 1–27. [CrossRef]
12. Zhang, J.; Chen, Y.; Zhang, Z. A remote sensing-based scheme to improve regional crop model calibration at sub-model component level. *Agric. Syst.* **2020**, *181*, 102814. [CrossRef]
13. Chen, Y.; Tao, F. Improving the practicability of remote sensing data-assimilation-based crop yield estimations over a large area using a spatial assimilation algorithm and ensemble assimilation strategies. *Agric. For. Meteorol.* **2020**, *291*, 108082. [CrossRef]
14. Herzen, J.; Lässig, F.; Piazzetta, S.G.; Neuer, T.; Tafti, L.; Raille, G.; Van Pottelbergh, T.; Pasieka, M.; Skrodzki, A.; Huguenin, N. Darts: User-friendly modern machine learning for time series. *J. Mach. Learn. Res.* **2022**, *23*, 5442–5447.
15. Adaryani, F.R.; Mousavi, S.J.; Jafari, F. Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN. *J. Hydrol.* **2022**, *614*, 128463. [CrossRef]

16. Alkabbani, H.; Hourfar, F.; Ahmadian, A.; Zhu, Q.; Almansoori, A.; Elkamel, A. Machine Learning-based Time Series Modelling for Large-Scale Regional Wind Power Forecasting: A Case Study in Ontario, Canada. *Clean. Energy Syst.* **2023**, *5*, 100068. [CrossRef]
17. Hu, J.; Liu, B.; Peng, S. Forecasting salinity time series using RF and ELM approaches coupled with decomposition techniques. *Stoch. Environ. Res. Risk Assess.* **2019**, *33*, 1117–1135. [CrossRef]
18. Wang, T.; Xiao, Z.; Liu, Z. Performance evaluation of machine learning methods for leaf area index retrieval from time-series MODIS Reflectance Data. *Sensors* **2017**, *17*, 81. [CrossRef]
19. Pipia, L.; Muñoz-Marí, J.; Amin, E.; Belda, S.; Camps-Valls, G.; Verrelst, J. Fusing optical and SAR time series for LAI gap filling with multioutput Gaussian processes. *Remote Sens. Environ.* **2019**, *235*, 111452. [CrossRef]
20. Huang, A.; Shen, R.; Di, W.; Han, H. A methodology to reconstruct LAI time series data based on generative adversarial network and improved Savitzky-Golay filter. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102633. [CrossRef]
21. Zhen, X.; Shao, H.; Zhang, W.; Huo, W.; Batchelor, W.D.; Hou, P.; Wang, E.; Mi, G.; Miao, Y.; Li, H. Testing a bell-shaped function for estimation of fully expanded leaf area in modern maize under potential production conditions. *Crop J.* **2018**, *6*, 527–537. [CrossRef]
22. Birch, C.; Hammer, G.; Rickert, K. Improved methods for predicting individual leaf area and leaf senescence in maize (*Zea mays*). *Aust. J. Agric. Res.* **1998**, *49*, 249–262. [CrossRef]
23. Eldoma, I.M.; Li, M.; Zhang, F.; Li, F.-M. Alternate or equal ridge–furrow pattern: Which is better for maize production in the rain-fed semi-arid Loess Plateau of China? *Field Crops Res.* **2016**, *191*, 131–138. [CrossRef]
24. Han, E.; Ines, A.; Koo, J. Global high-resolution soil profile database for crop modeling applications. *Harv. Dataverse* **2015**, *1*, 1–37.
25. Jones, C.A.; Kiniry, J.R.; Dyke, P. *CERES-Maize: A Simulation Model of Maize Growth and Development*; Texas A & M University Press: College Station, TX, USA, 1986.
26. Zhuo, W.; Huang, J.; Gao, X.; Ma, H.; Huang, H.; Su, W.; Meng, J.; Li, Y.; Chen, H.; Yin, D. Prediction of winter wheat maturity dates through assimilating remotely sensed leaf area index into crop growth model. *Remote Sens.* **2020**, *12*, 2896. [CrossRef]
27. Chen, J.; Jönsson, P.; Tamura, M.; Gu, Z.; Matsushita, B.; Eklundh, L. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. *Remote Sens. Environ.* **2004**, *91*, 332–344. [CrossRef]
28. Law, K.; Stuart, A.; Zygalakis, K. *Data Assimilation*; Springer: Cham, Switzerland, 2015; Volume 214, p. 52.
29. 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]
30. DeJonge, K.C.; Ascough II, J.C.; Ahmadi, M.; Andales, A.A.; Arabi, M. Global sensitivity and uncertainty analysis of a dynamic agroecosystem model under different irrigation treatments. *Ecol. Model.* **2012**, *231*, 113–125. [CrossRef]
31. Wang, Y.; Guo, F.; Shen, H.; Xing, X.; Ma, X. Global sensitivity analysis and evaluation of the DSSAT model for summer maize (*Zea mays* L.) under irrigation and fertilizer stress. *Int. J. Plant Prod.* **2021**, *15*, 523–539. [CrossRef]
32. Duan, Q.; Gupta, V.K.; Sorooshian, S. Shuffled complex evolution approach for effective and efficient global minimization. *J. Optim. Theory Appl.* **1993**, *76*, 501–521. [CrossRef]
33. Huang, J.; Ma, H.; Sedano, F.; Lewis, P.; Liang, S.; Wu, Q.; Su, W.; Zhang, X.; Zhu, D. Evaluation of regional estimates of winter wheat yield by assimilating three remotely sensed reflectance datasets into the coupled WOFOST–PROSAIL model. *Eur. J. Agron.* **2019**, *102*, 1–13. [CrossRef]
34. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
35. Joachims, T. *Making Large-Scale SVM Learning Practical*; Technical Report; Technische Universität Dortmund: Dortmund, Germany, 1998.
36. Hosseini, M.; McNairn, H.; Mitchell, S.; Robertson, L.D.; Davidson, A.; Ahmadian, N.; Bhattacharya, A.; Borg, E.; Conrad, C.; Dabrowska-Zielinska, K. A comparison between support vector machine and water cloud model for estimating crop leaf area index. *Remote Sens.* **2021**, *13*, 1348. [CrossRef]
37. Zhang, P.; Shu, S.; Zhou, M. An online fault detection model and strategies based on SVM-grid in clouds. *IEEE CAA J. Autom. Sin.* **2018**, *5*, 445–456. [CrossRef]
38. Shamshiri, R.; Eide, E.; Høyland, K.V. Spatio-temporal distribution of sea-ice thickness using a machine learning approach with Google Earth Engine and Sentinel-1 GRD data. *Remote Sens. Environ.* **2022**, *270*, 112851. [CrossRef]
39. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
40. Gislason, P.O.; Benediktsson, J.A.; Sveinsson, J.R. Random forests for land cover classification. *Pattern Recognit. Lett.* **2006**, *27*, 294–300. [CrossRef]
41. Battude, M.; Al Bitar, A.; Morin, D.; Cros, J.; Huc, M.; Sicre, C.M.; Le Dantec, V.; Demarez, V. Estimating maize biomass and yield over large areas using high spatial and temporal resolution Sentinel-2 like remote sensing data. *Remote Sens. Environ.* **2016**, *184*, 668–681. [CrossRef]
42. Balkovič, J.; van der Velde, M.; Schmid, E.; Skalský, R.; Khabarov, N.; Obersteiner, M.; Stürmer, B.; Xiong, W. Pan-European crop modelling with EPIC: Implementation, up-scaling and regional crop yield validation. *Agric. Syst.* **2013**, *120*, 61–75. [CrossRef]
43. Cammarano, D.; Basso, B.; Holland, J.; Gianinetti, A.; Baronchelli, M.; Ronga, D. Modeling spatial and temporal optimal N fertilizer rates to reduce nitrate leaching while improving grain yield and quality in malting barley. *Comput. Electron. Agric.* **2021**, *182*, 105997. [CrossRef]

44. Wen, Y.; Zhang, W.; Luo, R.; Wang, J. Learning text representation using recurrent convolutional neural network with highway layers. *arXiv* **2016**, arXiv:1606.06905.
45. Montesinos López, O.A.; Montesinos López, A.; Crossa, J. Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*; Springer: Cham, Switzerland, 2022; pp. 109–139.

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