remote sensing

MDPI

Article

Monitoring of Low Chl-a Concentration in Hulun Lake Based on Fusion of Remote Sensing Satellite and Ground Observation Data

Siyuan Zhang [1], Yinglan A [2,3,*], Libo Wang [1], Yuntao Wang [2,3], Xiaojing Zhang [1], Yi Zhu [1] and Guangwen Ma [4]

[1] College of Water Sciences, Beijing Normal University, Beijing 100875, China; 202221470031@mail.bnu.edu.cn (S.Z.); emmawang@mail.bnu.edu.cn (L.W.); zhangxj19@mail.bnu.edu.cn (X.Z.); zhuyi@bnu.edu.cn (Y.Z.)
[2] State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China; ytwang@bnu.edu.cn
[3] Innovation Research Center of Satellite Application, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[4] China National Environmental Monitoring Centre, State Environment Protection Key Laboratory of Environmental Monitoring Quality Control, Beijing 100012, China; magw@cnemc.cn
* Correspondence: ayl@bnu.edu.cn; Tel.: +86-13426127458

Abstract: China's northern Hulun Lake is a significant body of water internationally. The issue of eutrophication has gained prominence in recent years. The achievement of precise chlorophyll-a (Chl-a) monitoring is crucial for safeguarding Hulun Lake's ecosystem. The machine learning-based remote sensing inversion method has been shown to be effective in capturing the intricate relationship between independent and dependent variables; however, it lacks a priori knowledge and is limited by the quality of remote sensing data sources. The relationship between independent and dependent variables can be more accurately simulated with the use of suitable auxiliary variables. Therefore, three machine learning models—random forest (RF), adaptive boosting (AdaBoost), and extreme gradient boosting (XGBoost)—were established in this study using meteorological observation parameters as auxiliary variables combined with Sentinel-2 satellite image remote sensing band combinations as independent variables and measured Chl-a data as dependent variables. The estimation effects before and after the fusion of meteorological ground observation data were compared, and the best model was used to estimate the spatial–temporal variation trend of Chl-a in the regional water body. The results show that (1) the addition of meteorological parameters as auxiliary variables improved the precision of the three machine models; the decision coefficient (R2) rose by 7.25%, 5.71%, and 7.20%, respectively, to 0.76, 0.66, and 0.73. (2) The concentration of Chl-a in the lake region was projected from June to October 2019 to October 2021 using the RF optimal estimating model of meteorological fusion. The northeast, southwest, and south of the lake were where the comparatively high concentration values of Chl-a were located, whereas the lake's center had a generally low concentration of the substance. Chromatically, Chl-a typically peaked in August after initially increasing and then declining. (3) The three rivers that feed into the river have varying levels of water pollution, with chemical oxygen demand (COD) and total nitrogen (TN) pollution being the most severe. This is what primarily caused the higher levels of Chl-a in the northeast, southwest, and south. This study is crucial for the preservation and restoration of Hulun Lake's natural ecosystem and offers some technical support for the monitoring of the lake's concentration of Chl-a.

Keywords: chlorophyll-a; machine learning; remote sensing inversion; meteorological fusion

Citation: Zhang, S.; A, Y.; Wang, L.; Wang, Y.; Zhang, X.; Zhu, Y.; Ma, G. Monitoring of Low Chl-a Concentration in Hulun Lake Based on Fusion of Remote Sensing Satellite and Ground Observation Data. *Remote Sens.* **2024**, *16*, 1811. https://doi.org/10.3390/rs16101811

Academic Editor: Gabriel Senay

Received: 6 March 2024
Revised: 29 April 2024
Accepted: 16 May 2024
Published: 20 May 2024

1. Introduction

Lake eutrophication is a major challenge in the modern water environment problem, which is referred to as excessively high nutrient salt content in lakes (such as nitrogen

and phosphorus) that causes excessive growth of algae and aquatic plants, deteriorating the lake's ecological environment and leading to the loss of ecological service function [1]. According to the bulletin of the Ministry of Ecological Environment of China for 2022, out of the 204 key lakes and reservoirs, 61 (29.9%) are eutrophic. Increased runoff, precipitation, and other nutrient input into the lakes as a result of intensifying human activity and the expansion of the watershed industry have made lake eutrophication worse in recent years. Hulun Lake, a national nature reserve, is known as the mother lake of the Hulun Buir grassland, which is essential to the social and economic growth of the area as well as the security of the ecological environment. However, Hulun Lake has been experiencing severe eutrophication in recent years as a result of the escalation of upstream non-point source pollution and the impact of nearby industrial activities [2]. The water quality monitoring data shows that the nutrient contents of nitrogen and phosphorus in Hulun Lake continuously increase, and the water eutrophication index remain high [3], which has a negative impact on the lake fishery resources and water landscape. As the primary pigment for photosynthesis in algae, chlorophyll-a (Chl-a) is one of the most important indicators to evaluate the level of eutrophication in lakes among many other monitoring indicators. A change in Chl-a concentration directly corresponds to the biomass of algae in the water body. Therefore, it is essential to monitor changes in Chl-a concentration in order to formulate scientific strategies for lake management and treatment, as well as to timely reflect the trend of lake eutrophication.

Remote sensing monitoring offers advantages over typical artificial sampling for monitoring Chl-a concentration due to its low cost, high efficiency, and continuous space–time scale. Researchers have successfully inverted Chl-a in several regions using different remote sensing satellites and sensors, including Landsat [4–6], Modis [7,8], Sentinel-2 [9–11], and others. The Sentinel-2 satellite is crucial in water quality inversion studies because of its high resolution and frequent revisit cycle compared to other satellite platforms. The satellite collects data in 13 spectral bands, ranging from visible to near-infrared, enabling precise monitoring of changes in Chl-a concentrations in water [12]. In the inversion method of Chl-a, the traditional empirical model based on spectral characteristics has been widely used. The empirical model [13,14] estimating Chl-a concentration by correlating remote sensing reflectance with Chl-a concentration is straightforward but does not include an optical mechanism. Its precision is constrained by particular conditions and the type of water. The semi-analytical model [15,16] integrates empirical models and radiation transfer principles, incorporating specific physical concepts. The analysis model [17] effectively replicates the light propagation in water, establishes the correlation between Chl-a and optical coefficients, and successfully achieves Chl-a concentration inversion. Yet, it requires a broad variety of input parameters concerning the optical characteristics of water, leading to limited practical utility. In recent years, there has been an increasing trend in utilizing machine learning technology to enhance the accuracy of Chl-a concentration inversion in various research [10,18–20]. Machine learning techniques, like support vector machines (SVMs) and random forests (RFs), are effective in handling intricate water body conditions and diverse data sources. The remote sensing inversion model based on machine learning solely considers optical features as independent variables, lacking pre-existing knowledge. The uncertainty is significant, and its accuracy needs to be enhanced. Introducing suitable auxiliary variables helps enhance the simulation of the connection between independent and dependent variables [21]. Yang et al. found that using vegetation data and a drought index as auxiliary data can improve the accuracy of the remote sensing inversion of ground temperature [22]. Other scholars have also found that surface temperature and a vegetation index can assist the production of soil products [23]. Meteorological elements, like temperature, precipitation, wind speed, and sun radiation, have been proven to greatly impact Chl-a concentrations [24–27]. Increased water temperature accelerates algae growth, which, in turn, increases the concentration of Chl-a, which is especially noticeable during seasons with notable temperature fluctuations [24]. The fluctuation in precipitation impacts the nutrient levels in the water, which, in turn, influence the photosynthetic processes of

algae. Changes in wind speed impact the mixing and stratification of the water, thereby influencing the distribution and proliferation of algae [26]. Solar radiation is a source of energy for photosynthesis, and it also impacts algae growth and Chl-a synthesis [27]. Considering the ease of accessing meteorological data and the extensive use of high-precision meteorological data, can meteorological data and remote sensing band combinations be utilized as input elements to enhance the accuracy of Chl-a inversion?

Based on the above thinking, this study focuses on studying Hulun Lake in Inner Mongolia, China. It utilizes meteorological data as auxiliary variables, incorporates remote sensing band combinations from Sentinel-2, and applies various machine learning algorithms to develop a Chl-a estimation model that integrates sky and earth data. The optimal machine learning algorithm was chosen to obtain the temporal and spatial distribution of Chl-a in Hulun Lake from June to October, spanning 2019 to 2021, and identify the key prevention areas for algal bloom outbreaks. This study delves deeper into the water quality of rivers flowing into the lake, analyzes the causes affecting Chl-a levels in Hulun Lake, and offers scientific support for controlling algal blooms and safeguarding the environment in Hulun Lake.

2. Materials and Methods

2.1. Study Area and Data

2.1.1. Study Area

Hulun Lake (Figure 1) is located in the western part of the Hulun Buir grassland in Inner Mongolia, China, with geographic coordinates ranging from 48.51°N to 49.34°N and 117°E to 117.69°E. It is the fifth largest freshwater lake in China and is known as the mother lake of Hulun Buir grassland. It covers an area of approximately 2000 square kilometers and has an uneven oblique rectangular shape. Hulun Lake is a semi-arid grassland lake with a temperate continental climate characterized by short, cool summers and long, cold winters. The average annual temperature ranges from −0.50 to 2.67 °C, with yearly precipitation between 247 and 319 mm. Annual evaporation is approximately 1400 mm, far exceeding precipitation levels. The prevailing wind direction is northwest year-round, with an average wind speed of approximately 5 m per second. Hulun Lake is a key component of the Erguna drainage system, receiving its primary sources of replenishment from atmospheric precipitation, surface runoff, and groundwater. The Krulun, Wuerxun, and Xinkai Rivers are the main sources of surface runoff. The lake plays a crucial role in maintaining regional ecological balance and driving economic development. However, the water quality of the lake is deteriorating due to strong evaporation, reduced rainfall, and human activities, leading to frequent algal blooms that pose a significant threat to human health and economic development.

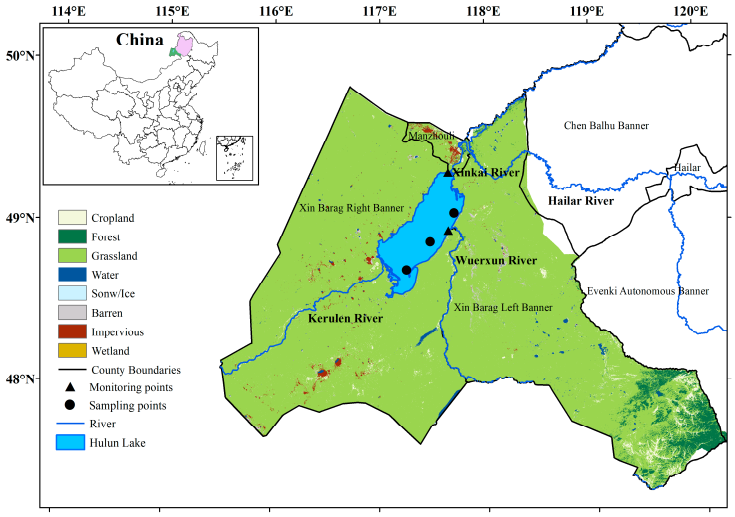

Figure 1. Schematic of study area.

2.1.2. Data Collection

(1) water quality data

The time scale for Chl-a data used to establish the machine learning model in this paper ranges from June to October, spanning 2016 to 2021. Data were collected once a month at five different points: Ganzhuhua, Xiaohekou, Chuhu, Ruhu, and Huxin. (Some months lack data.) The first two points are national monitoring stations for Hulun Lake, while the latter three are field sampling points. The time scale for water quality data from rivers flowing into the lake is from June to October, spanning 2019 to 2021, with monthly measurements (though some months lack data). The three water quality sampling sites include Xiaohekou (Xinkai River), Gibhurangtu (Wuerxun River), and Alatan Ermolle (Krulun River);

(2) meteorological data

The raw meteorological data used in this paper are derived from the China Meteorological Data Set V3.0, which contains daily meteorological data from January 1951 onward at foundational meteorological stations across China. In this paper, we extracted data from meteorological stations around Hulun Lake (Manzhouli, Hailar, Xin Barag Right Banner, and Xin Barag Left Banner) for daily average temperature, daily precipitation, daily average relative humidity, average wind speed, and sunshine hours from June to October, spanning 2016 to 2021;

(3) remote sensing data

The Sentinel-2 series is part of the European Space Agency (ESA) Earth observation program, the Copernicus Program. It includes two similar satellites, Sentinel-2A and Sentinel-2B, launched in 2015 and 2017, respectively. When working in tandem, these satellites provide global coverage with a frequency of every five days. Sentinel-2 is equipped with a Multi-Spectral Imager (MSI) that captures images in 13 spectral bands ranging from 443 nm to 2190 nm. Visible and near-infrared bands (B2, B3, B4, and B8) have 10 m resolution; red-edge and short-wave infrared bands (B5, B6, B7, B8A, B11, and B12) have 20 m resolution; and three atmospheric correction bands (B1, B9, and B11) have 60 m resolution. The study acquired L1C-level image data (radiometrically calibrated) with cloud cover less than 20% from the Copernicus Open Access Hub, synchronized or quasi-synchronized with Chl-a sampling times. The images were then subjected to atmospheric correction using the Sen2Cor plug-in provided by the ESA, and the visible and near-infrared bands were resampled to 10 m.

2.2. Methods

According to the workflow chart (Figure 2), the article uses the following main methods.

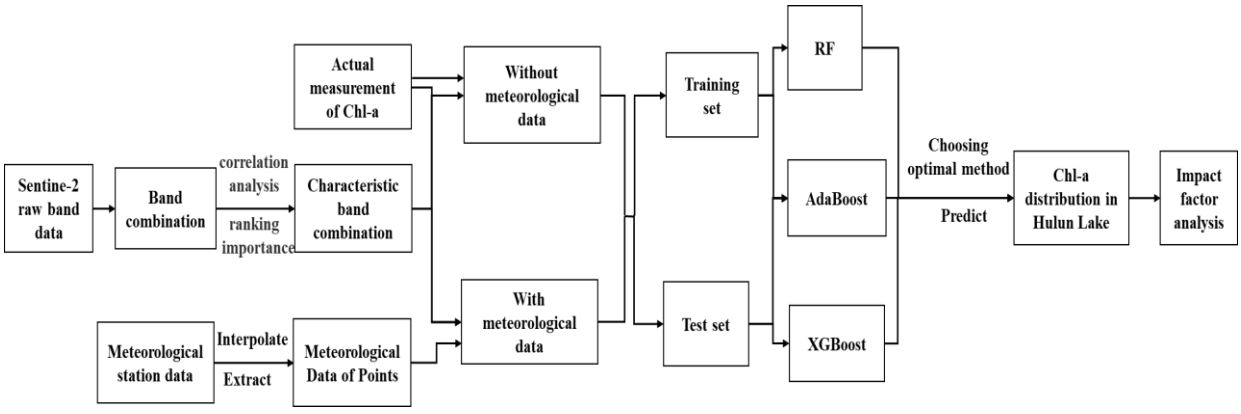

Figure 2. Workflow chart of study.

2.2.1. Remote Sensing Feature Selection

In this paper, Pearson correlation analysis was used to screen for band combinations with high correlation to chlorophyll-a concentration, and then, the selected combinations were ranked by RF-based feature importance to further filter the selection of characteristic remote sensing bands.

(1) Pearson correlation analysis

Pearson correlation analysis is a statistical method used to quantify the strength of a linear relationship between two continuous variables. The Pearson correlation coefficient (r) ranges from -1 to 1, where 0 indicates no linear association, -1 indicates perfect negative correlation, and 1 indicates perfect positive correlation. The following is the calculating formula:

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \quad (1)$$

In the formula, $X(x_1, x_2, \ldots, x_n)$ and $Y(y_1, y_2, \ldots, y_n)$ are two consecutive sequences of related variables, and r is the correlation coefficient;

(2) RF-based importance ordering

RF is an integrated learning method that builds several decision trees to perform classification or regression analysis. By examining the decision trees' node splits, RF may determine how important a feature is.

2.2.2. Machine Learning Methods

In this paper, three popular machine learning methods—RF, AdaBoost, and XGBoost—were used to construct the Chl-a inversion model. RF is an integrated algorithm based on decision trees proposed by Breiman in 2001 [28]. It constructs multiple decision trees by bagging at random and makes a final prediction by averaging or voting the results [29]. AdaBoost [30] is an adaptive boosting algorithm, which can iteratively adjust the weight of training data so that the data that were wrongly predicted in the previous iteration will receive more attention in the subsequent iterations. XGBoost is an efficient algorithm based on gradient lifting [31], which corrects the prediction error of the previous step by adding new weak learners incrementally, and uses a gradient descent algorithm to minimize the loss function.

2.2.3. Accuracy Evaluation Index

In this study, three common evaluation indexes—the coefficient of determination (R^2), the mean absolute error (MAE), and the root mean square error (RMSE)—were selected to assess the performance of the machine learning model. R^2 quantifies the model's ability to account for the variability in response variables, with a higher value indicating better model performance. The MAE calculates the mean of the absolute differences between predicted and actual values, while the RMSE calculates the square root of the average of the squared prediction errors. Lower values for both indicate greater accuracy. The formula is as follows:

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

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \hat{y}_i \right| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^{N} \left(y_i - \hat{y}_i\right)^2} \quad (4)$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the average value of the actual values.

3. Results

3.1. Spatial–Temporal Analysis of Measured Chl-a

Based on the available measured sample point data, we plotted the mean Chl-a concentration at each sample point from June to October, spanning 2016 to 2021, and the multi-year monthly mean concentrations at five sample points (Figure 3). The results show that among the five sample points, the concentration values at two sample points located at the edge of the lake were relatively high, 14.4 μg/L and 11.28 μg/L, for Ganzhuhua and Xiaohekou, respectively, while the concentration values at the other three sample points were relatively low, with Chuhu, Ruhu, and Huxin at 10.22, 8.63, and 7.70 μg/L, respectively. The monthly mean concentration for the five sample sites was highest in September at 13 μg/L, followed by October at 11.4 μg/L, and lowest in June at 8.71 μg/L.

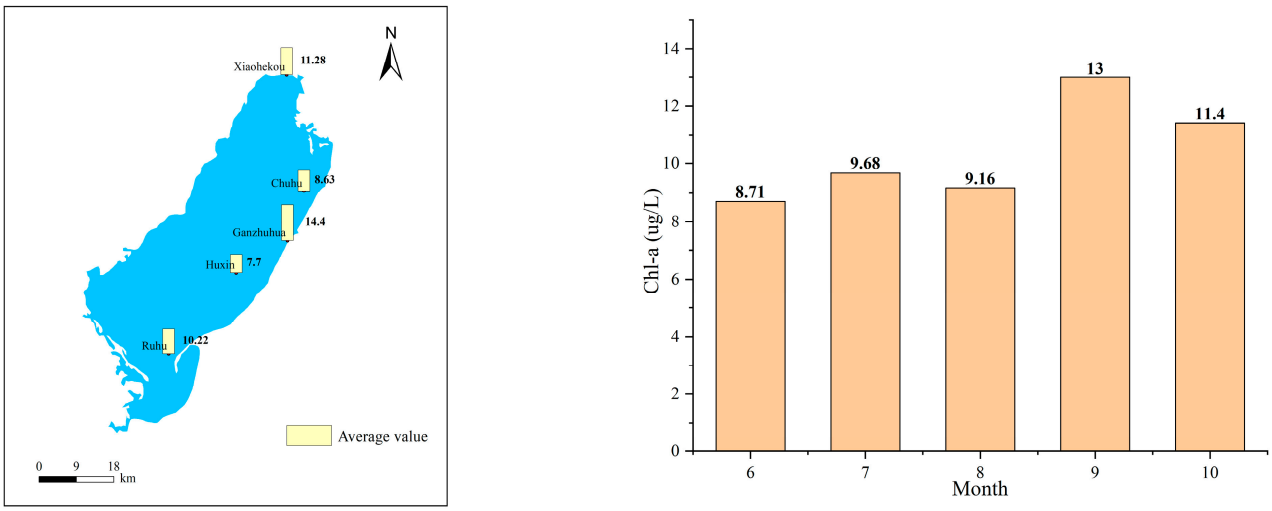

Figure 3. Distribution of mean Chl-a concentrations at each sample point (**left**) and multi-year monthly mean concentrations of Chl-a for the 5 samples (**right**).

3.2. Machine Learning Model Construction

3.2.1. Selection of Characteristic Meteorological Factors

Based on previous research, we selected the precipitation, average temperature, sunshine time, average wind speed, and relative humidity of the day at the sampling time as the characteristic meteorological factors for constructing the machine learning model. In order to match the pixel size of remote sensing images, Kriging interpolation was used to interpolate the data of four weather stations around Hulun Lake (Figure 4) to 10 m accuracy. Then, the corresponding sampling site meteorological data on the day of sampling was extracted.

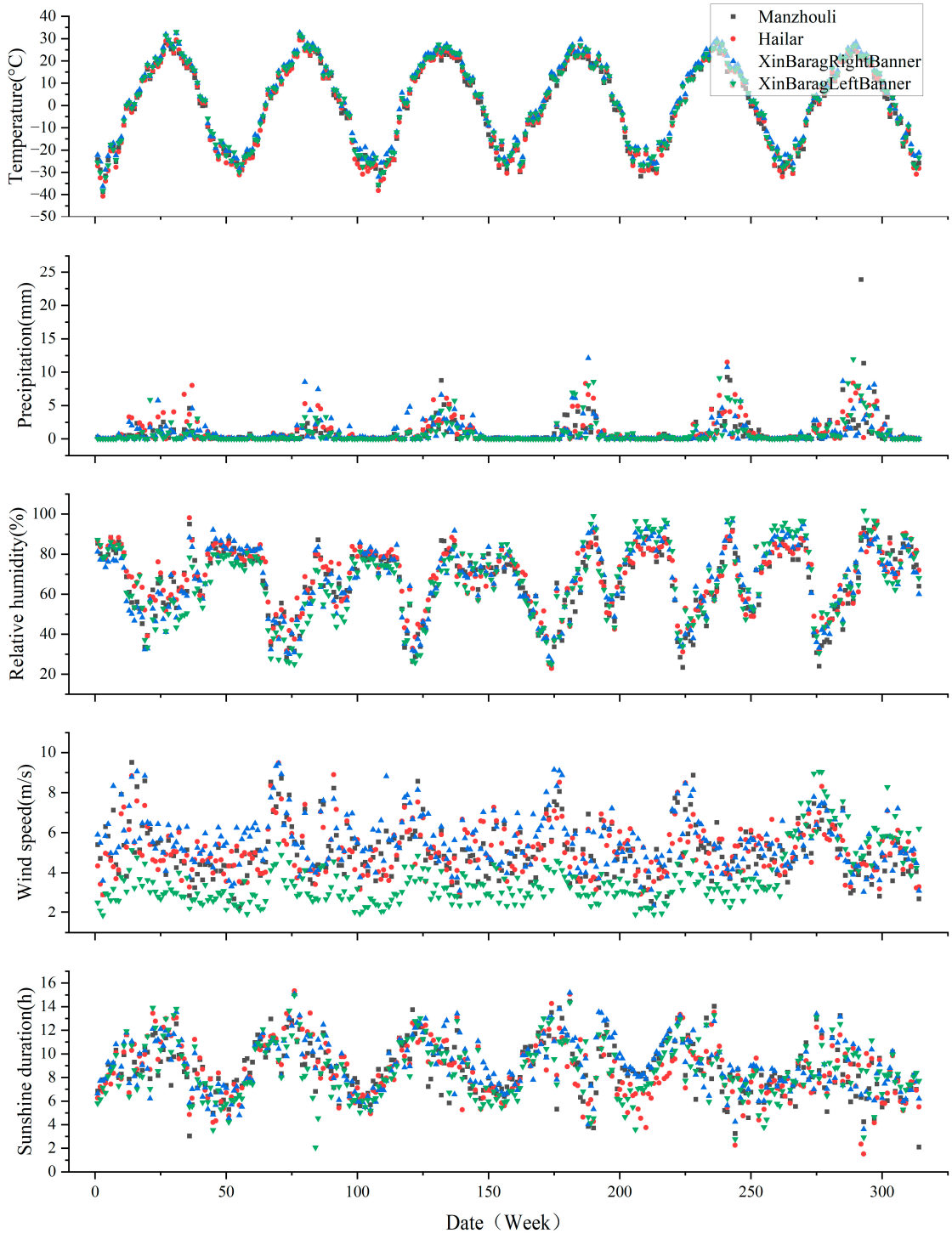

Figure 4. Weekly average values of temperature, precipitation, relative humidity, wind speed, and sunshine duration at four meteorological stations from 2016 to 2021.

3.2.2. Selection of Remote Sensing Feature Bands

The spectral characteristics of the water concentrate primarily in the visible and near-infrared bands. Chl-a is a crucial factor in water bodies and has a notable impact on the spectral characteristics of water. Distinct absorption peaks of Chl-a were observed near blue-violet and red light, with noticeable reflection peaks near green light and the red edge. The reflection peak of the latter shifted toward longer wavelengths as the concentration

increased. We used the B2-B8A band of Sentinel-2 to construct the model, based on the spectral properties provided and other studies. Typical band models consist of single band, band sum, band ratio, band difference, three-band, and normalized ratio models. This study combined B2-B8A bands in pairs to create the band sum, band difference, band ratio, and normalized index models. Considering the difference between bands as numerators and denominators, the band ratio model was constructed with bands as the numerator and denominator one by one. The three-band concept is derived from the bio-optical principle and necessitates the identification of certain bands: the red absorption peak band (λ1), the adjacent fluorescence peak band (λ2), and the absorption peak band mostly influenced by pure water (λ3). For the configuration of the Sentinel-2 remote sensing bands, B4 and B5 were chosen as λ1 and λ2, while B6, B7, B8, and B8a bands were chosen as λ3 to create a three-band model. Ultimately, a total of 152 band combinations were created, consisting of 8 single bands, 28 band sums, 28 band differences, 28 normalized indexes, 52 band ratios, and 4 three-band models (Table 1). We then examined the Pearson correlation coefficient between Chl-a concentration and various band combinations (Table 2).

Table 1. Remote sensing band combination prediction data set.

Feature Category	Formula	Quantity
Single band	B_i	8
Triple band	$[1/(\lambda 1) - 1/(\lambda 2)] \times \lambda 3$	4
Band sum	$(B_i + B_j)$	28
Band difference	$B_i - B_j$	28
Normalized index	$(B_i - B_j)/(B_i + B_j)$	28
Band ratio	(B_i/B_j)	52

Note: In the table, B represents the reflectance of bands in remote sensing images, and i and j represent band numbers. For the purpose of this study, the Sentinel-2 B2-B8A waveband was used.

Table 2. Correlation coefficient table for combination of Chl-a and remote sensing wavebands by Pearson selection.

Band Combination	Correlation Coefficient	Band Combination	Correlation Coefficient
B5−B6	−0.549 ***	B7/B5	0.398 ***
B4−B6	−0.545 ***	B7/B2	0.398 ***
B4−B7	−0.535 ***	B8A/B5	0.394 ***
B4−B8A	−0.531 ***	B7−B8A	−0.390 ***
B3−B8A	−0.514 ***	B6−B8A	−0.383 ***
B3−B6	−0.512 ***	B3−B7/B3 + B7	−0.380 ***
B5−B7	−0.494 ***	B6−B8	0.372 **
B3−B7	−0.493 ***	B6/B2	0.371 **
B5−B8A	−0.483 ***	B3−B6/B3 + B6	−0.370 **
B4−B8	−0.460 ***	B6/B5	0.369 **
B8A/B3	0.459 ***	B3−B5	−0.369 **
B7/B3	0.455 ***	B5/B3	0.365 **
B8−B8A	−0.440 ***	B2−B8A	−0.364 **
B6/B3	0.434 ***	B8/B3	0.357 **
B3−B8	−0.432 ***	B3−B8A/B3 + B8A	−0.356 **
B8A/B4	0.426 ***	B3−B5/B3 + B5	−0.350 **
B8A/B2	0.423 ***	B4−B7/B4 + B7	−0.350 **
B7/B4	0.422 ***	B2−B7/B2 + B7	−0.348 **
B7−B8	0.407 ***	B2−B8A/B2 + B8A	−0.341 **
B6/B4	0.399 ***	B8A/B8	0.340 **

Note: ** $p < 0.01$, *** $p < 0.001$.

The Pearson coefficient indicated a weak linear link between the single band and Chl-a concentration, with correlation coefficients less than or equal to 0.2. None of the correlations were significant, with the strongest correlation between B8A and Chl-a at 0.2. Furthermore, the linear association between bands and Chl-a was low. Table 2 lists the first 40 bands with a strong correlation. It is found that the correlation between band difference, band ratio, the normalized difference index, and Chl-a is high, showing a very significant or extremely significant correlation. B5-B6 has the highest correlation, and the absolute value of the correlation coefficient is 0.549.

In this study, we ranked the importance of random forest (RF) to further screen remote sensing bands with a significant $p < 0.05$, as the Pearson correlation coefficient only accounts for linear correlation, and the relationship between the remote sensing band and Chl-a concentration may be nonlinear. According to the importance ranking of RF, the remote sensing bands after the first six have lower importance and tend to converge. In order to prevent data redundancy, the top six band combinations selected for machine learning based on their importance were B3-B8, B3-B5/B3 + B5, B3-B6, B5/B3, B8/B3, and B3-B7.

3.2.3. Comparison of Different Machine Learning Models

Three machine learning models (random forest, AdaBoost, and XGBoost) were built using the Sklearn machine learning database. The models used selected meteorological feature factors and remote sensing feature bands as input variables and the measured Chl-a concentration as the response variable. In total, 70% of the data was chosen randomly for the training set, while 30% was allocated to the verification set.

The remote sensing band was initially utilized as the input variable. The study trained three models with the number of n_estimators (decision trees) set at intervals of 50 starting from 50 while keeping the other parameter at default settings. The test set's performance was analyzed across various numbers of decision trees to determine the most optimal number. Next, the meteorological factors were included in the input variables. The prediction performance of the three machine learning models was then compared using only remote sensing characteristic wave bands as input variables and both remote sensing characteristic wave bands and meteorological factors as input variables under the original parameter settings (Figure 5).

According to the comparison diagram (Figure 5), the RF model showed the highest performance among the three machine learning models when no meteorological data were included, with an R^2 of 0.708, MAE of 2.999 μg/L, and RMSE of 4.703 μg/L. In contrast, the AdaBoost model had the lowest performance, with an R^2 of 0.621, MAE of 3.759 μg/L, and RMSE of 5.099 μg/L on the test set. The meteorological data integration enhanced the precision of the test sets for the three machine learning models. The R^2 of the three machine learning models increased by 0.051, 0.035, and 0.049, respectively. The enhanced precision levels were 7.25%, 5.71%, and 7.20%. Following the enhancement, RF's R^2 was 0.759, the highest among the models, while AdaBoost had the lowest precision with an R^2 of 0.657. In addition, the precision of the MAE and RMSE were improved to different degrees, which means that the performance effect of the machine learning model was enhanced to a certain extent after the meteorological data auxiliary variables were added.

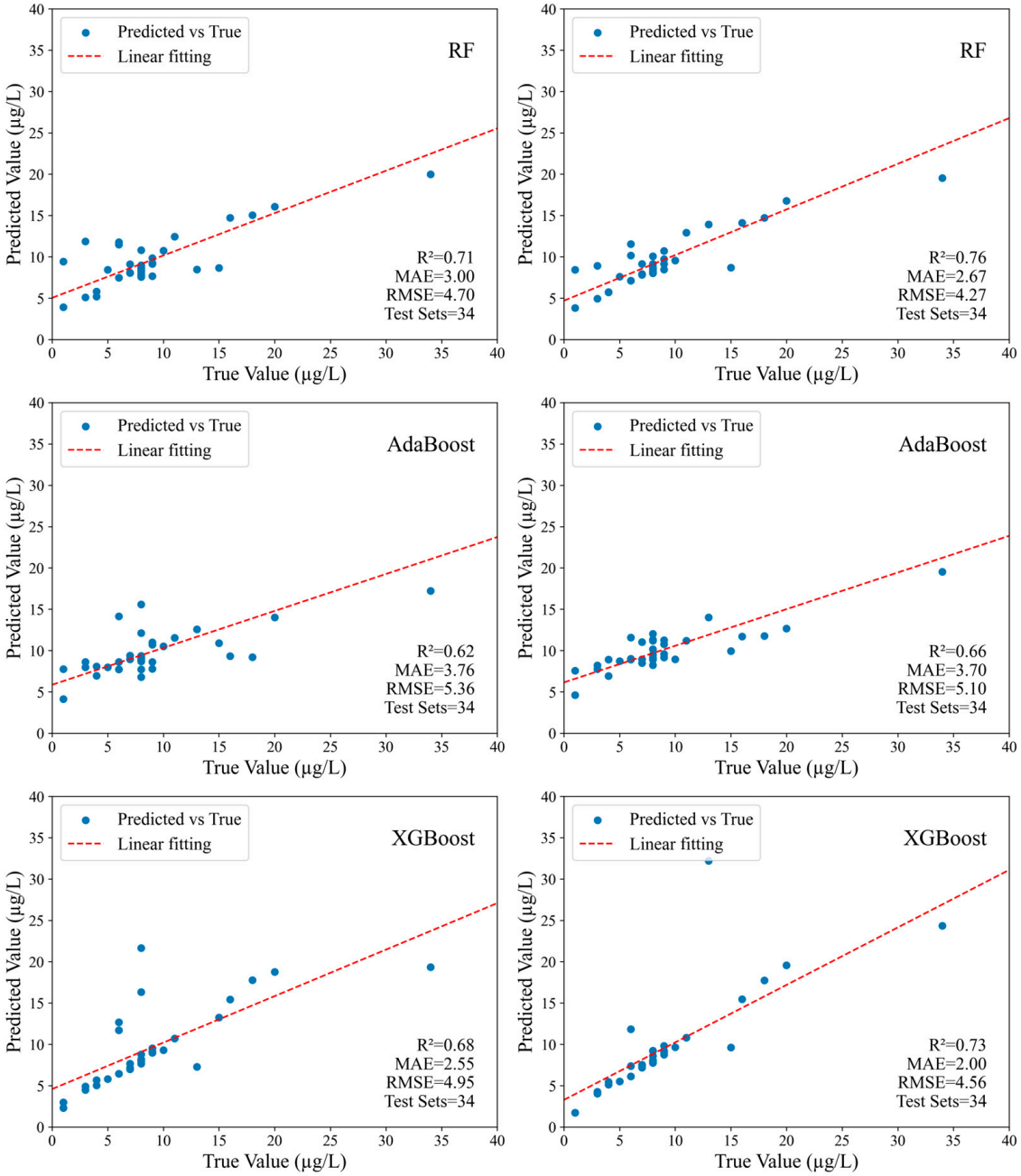

Figure 5. Comparison of predicted Chl-a concentration with measured Chl-a concentration by the three machine models (without meteorological auxiliary data on the left side and with meteorological auxiliary data on the right side).

3.3. Estimation of Chl-a Concentration in Hulun Lake

3.3.1. Spatial Variation

Based on the above research, the completely trained model of meteorological fusion was then used over the entire lake surface. The monthly estimation map of chlorophyll-a in Hulun Lake for one pair is shown in Figure 6. Spatially, in most images, the Chl-a concentration in the middle part of Hulun Lake was generally low, whereas relatively high Chl-a values were most frequently recorded near the northeast side of the lake, followed by the south and southwest. Among them, the concentrations in the relatively high Chl-a regions in July, August, and September of 2019, August and September of 2020, and July and August 2021 were all above 15 μg/L and reached 20 μg/L in August of 2020, which

might mean a high risk of algal bloom explosion. In terms of spatial distribution changes during the year, the spatial distribution of Chl-a concentration saw significant alterations in 2019 and 2020. The region with high Chl-a values was unstable, and the change was minimal in 2021. The high-value sector was primarily located on the southwest bank and northeast side of the lake.

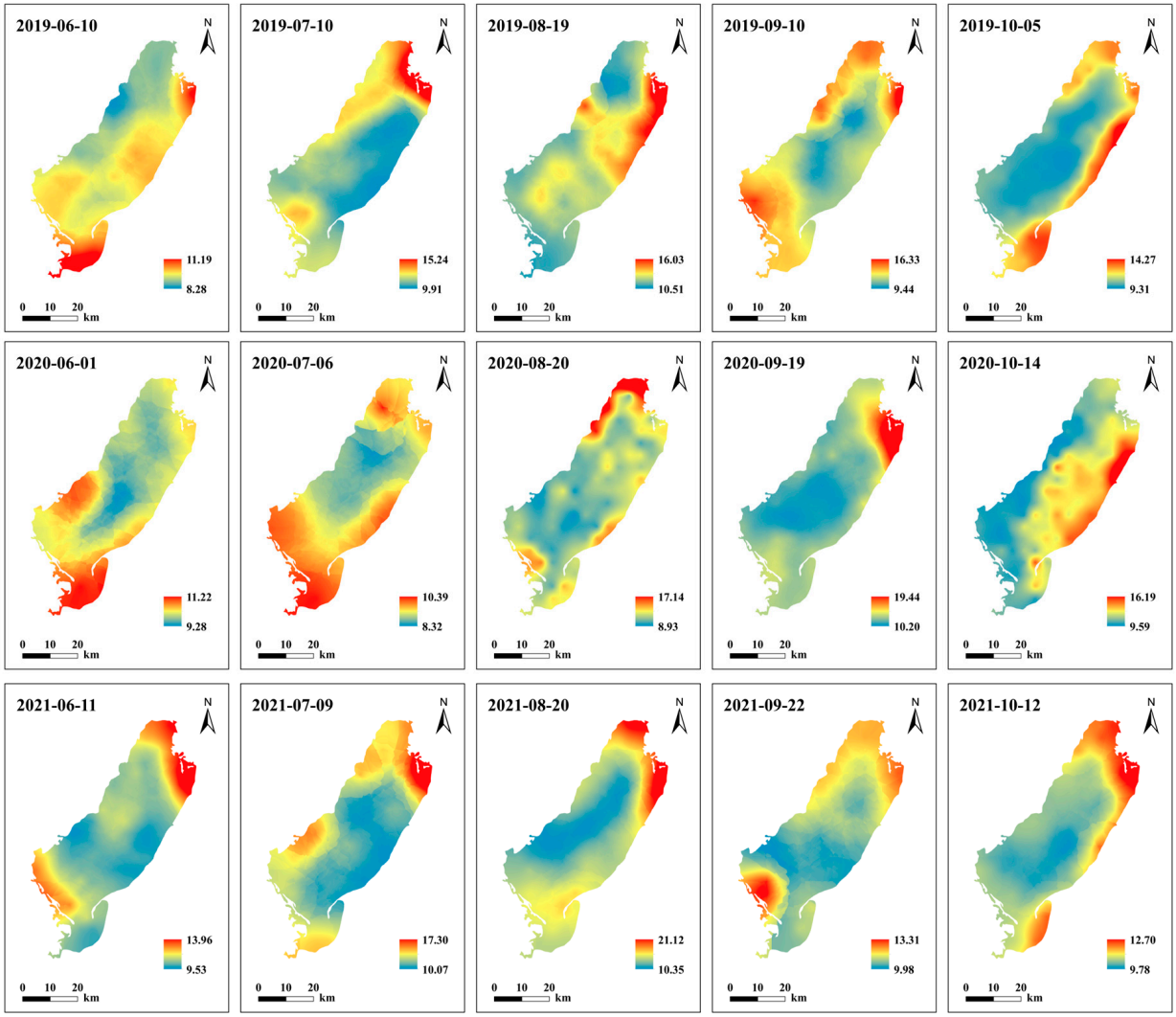

Figure 6. Estimation map of Chl-a concentration in Hulun Lake.

The river flowing into the lake serves as a vital conduit for energy transfer between the lake and its surroundings, and the nutrients brought by the river play a significant role in stimulating algal growth. We noticed that the southwest and northeast parts of the lake are located near the three rivers that flow into the lake: the Xinkai River, Urxun River, and Krulun River. Figure 7 displays the variations in six key water quality parameters (pH, DO, COD, NH_3-N, TN, and TP) of the three rivers that flow into the lake from June to October, spanning 2019 to 2021. The results indicate that the chemical oxygen demand (COD) levels of the three rivers flowing into the lake are high, with average values of 30.62, 25.73, and 17.24 mg/L, respectively. Compared to the surface water quality level (Table 3), the Xinkai River has a COD concentration level classified as Class V, indicating the highest degree of pollution. The Urxun River is classified as Class IV, whereas the Krulun River is classified as Class III. The overall nitrogen concentration in the three rivers was around 1.5 mg/L, meeting Class V standards. The average TP concentrations were 0.09, 0.06, and 0.12 mg/L, respectively, with TP levels in the Krulun River surpassing Class III standards. The NH3-N

concentrations in all three rivers exceeded 0.5 mg, reaching Class III limits. Higher nutrient levels were an important factor in the frequent occurrence of high-value regions of Chl-a concentration near the three inlets.

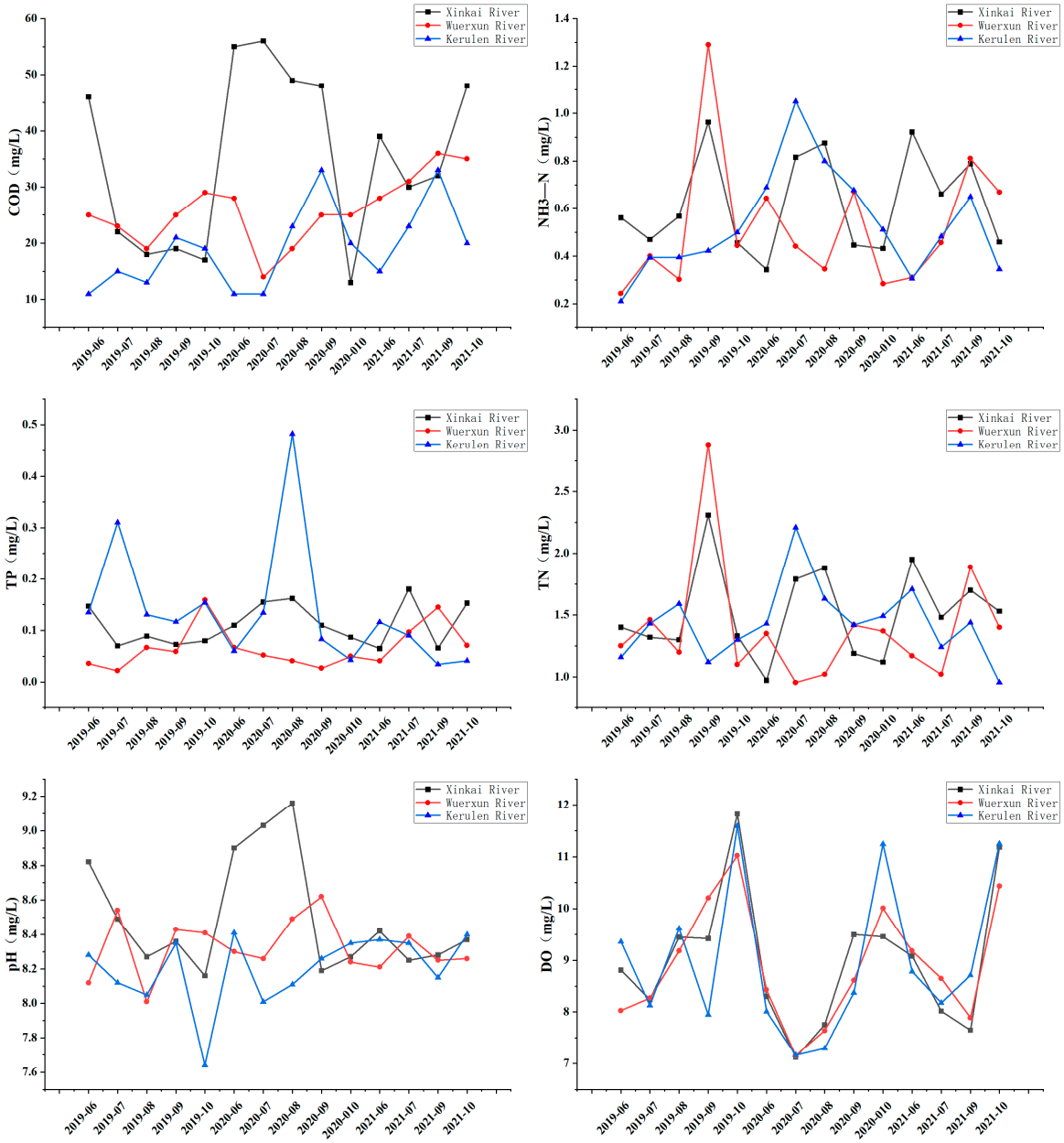

Figure 7. Monthly concentration change in main water quality parameters of rivers entering lake.

Table 3. Limit values of surface water environmental quality standards (unit: mg/L).

Items	Class I	Class II	Class III	Class IV	Class V
pH (dimensionless)			6–9		
DO (≥)	7.5	6.0	5.0	3.0	2.0
COD (≤)	15.0	15.0	20.0	30.0	40.0
NH3-N (≤)	0.15	0.5	1.0	1.5	2.0
TP (≤)	0.0	0.1	0.2	0.3	0.4
TN (≤)	0.2	0.5	1.0	1.5	2.0

3.3.2. Temporal Variation

According to Figure 8, in 2019, the Chl-a concentration in Hulun Lake followed an upward trend initially, reaching its highest in September at 12.11 μg/L, and subsequently fell. The lowest concentration was recorded in June at 9.39 μg/L, as shown in Figure 8. In 2020, the Chl-a concentration showed a slight change, reaching its highest average value of 11.62 μg/L in September and lowest of 9.35 μg/L in July. In 2021, Chl-a levels followed a similar pattern as in 2019. The average value reached its peak in August at 13.00 μg/L, and the lowest in June was 10.65 μg/L. The Chl-a concentration in Hulun Lake during non-ice periods typically followed an increasing trend, peaking in August. Algal blooms in the lake usually occur in July and August, aligning with peak Chl-a concentration.

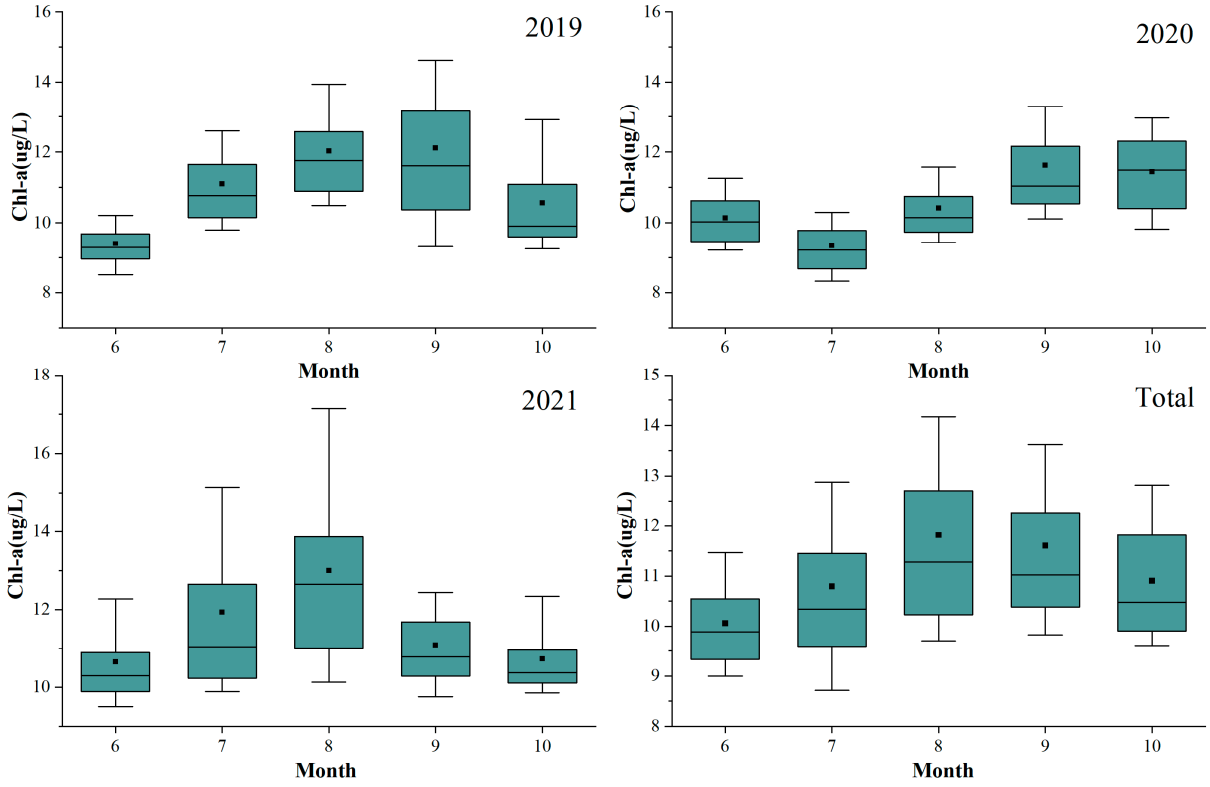

Figure 8. Hulun Lake Chl-a concentration monthly change box diagram from June to October. (The center line of the box represents the median value; the dots represent the average value; the bottom and top of the box represent the upper and lower quartiles, respectively; and the bottom and top lines represent the 10th and 90th percentiles, respectively.)

4. Discussion

The machine learning model in this work was built using Chl-a as the response factor and a combination of meteorological factors and remote sensing bands as input variables. The outcomes demonstrate that the machine learning model's precision increased, with the RF model's precision being the highest at 0.76 and the improvement amplitude at 7.25% (Figure 5). This shows that using meteorological factors as auxiliary variables can improve the accuracy of Chl-a inversion. Beyond water quality, other researchers have conducted similar research in areas like precipitation and soil moisture [32–34]. Ji et al. predicted precipitation by combining a vegetation index, topographic factors, and other auxiliary variables [32]. When Wang et al. studied albedo, they found that the model's accuracy improved after adding snow as an auxiliary variable [33]. In studies predicting wheat biomass, some scholars combined several vegetation indices, improving the accuracy of the RF model [34]. These results suggest that using auxiliary variables closely related to the target variables is an important method for improving model accuracy.

The selection of feature bands plays a key role in building a model [35]. In this study, the characteristic bands of Chl-a were progressively chosen using RF importance ranking and Pearson correlation analysis. Certain bands (like B5, B6, and B7) showed up in multiple combinations with strong correlation, suggesting that the red band of Sentinel-2 showed a lot of potential for monitoring Chl-a concentration [36]. Numerous studies have demonstrated how well machine learning models are able to grasp nonlinear interactions. This work effectively predicted the concentration of chlorophyll-a using a machine learning approach, thus validating the usefulness of artificial intelligence in monitoring water quality. The RF model outperforms the other two machine learning methods before and after, while Wu [19] et al. and Xie et al. [37] also concluded that the RF model had the best phenotype when conducting remote sensing inversion using multiple machine learning methods. Simultaneously, we observed that while the overall accuracy increased, the prediction effect was not very good for large concentrations of Chl-a before and after. On the one hand, this was mainly because the small amount of data with high concentrations of Chl-a in the limited dataset, which makes the machine learning model insensitive to the case of high Chl-a concentration, and the prediction effect is poor [38–40]. In future studies, the use of longer sequences of Chl-a observations to build models will improve this situation. On the other hand, it is due to the fact that integrated learning methods such as RF lead to a weakening for extreme values by averaging the predicted values [23,41].

The spatial–temporal distribution of Chl-a in Hulun Lake from 2019 to 2021 was obtained based on the RF model (Figure 6), and the generated Chl-a somewhat displayed a consistent pattern. The results show that the lake's center had a generally lower Chl-a level than its surroundings, and in line with the conclusions of Qian et al. [42] and Cao et al. [43] on phytoplankton and algal blooms in Hulun Lake, the regions with relatively high Chl-a concentration were mainly distributed in the northeast, south, and southwest of the lake, which indicates that the distribution of algal blooms in Hulun Lake has a high consistency with Chl-a concentration. In addition to meteorological conditions, nutrients are necessary for the growth of algae. The northeastern side of the lake is located near the Xinkai River and the Wuerxun River, and the southwestern side is located near the Krulun River. The nutrients N and P in the soil of the river carriers and the partial injection of human domestic water and livestock emissions caused the accumulation of nutrients. The results show that nutrient components, such as COD, TN, TP, and NH3-N, in the three rivers entering the lake, were polluted to different degrees (Figure 7), which were important factors for the distribution of the Chl-a concentration in Hulun Lake. There are also small estuarine scenic spots on the northeastern side. Human activities, including fossil fuel combustion and automobile exhaust emissions, have exacerbated the increase in nutrient concentration [44]. Furthermore, Hulun Lake has a significant fish farming base, and the additional pollution brought on by fish farming exacerbates the decline in water quality. The northwest wind of Hulun Lake [45] is the predominant wind direction, and the landscape on the southwest side of the lake is blocked, which may have caused phytoplankton to accumulate, with higher concentrations of Chl-a. Certain laws were also demonstrated by the variations in Chl-a in the time series. The results show that Chl-a in Hulun Lake during the non-ice summer often increased initially before declining and peaking in August (Figure 8). Other scholars have also found similar phenomena in Taihu Lake [46,47], Chaohu Lake [48], and other regions in China, mostly as a result of the strong relationship between temperature and sunshine hours and the growth of algae and phytoplankton. Algal photosynthesis quickened and increased in number in July and August as temperatures rose and the amount of sunshine increased.

There are still shortcomings and a need for development in this study as follows: (1) The distribution of Chl-a throughout the entire Hulun Lake could not be accurately represented by the data gathered from the five sampling locations employed in this investigation. In addition, the monthly data frequency from June to October in 2016–2021 makes the modeling data small. (2) With the meteorological factors as auxiliary variables, in order to match the accuracy of the Sentinel-2 image, the Kriging interpolation method

was only used to interpolate the whole lake, and there was uncertainty. Furthermore, the meteorological elements disregarded the impact of accumulated variables in favor of merely taking into account the current index. In the future, auxiliary variable data sets with higher precision could be collected to improve the precision of the model. Achieving perfect synchronization between the measured data and remote sensing bands is challenging due to the time-resolution limitations of remote sensing data sources. We will further improve the quality of remote sensing data in the future. (3) In analyzing the spatial–temporal variation of Chl-a in Hulun Lake, only the impacts of the water quality of three rivers flowing into the lake were considered, and no further analysis of the impacts of Chl-a was available. The driving factors of Chl-a variation in different regions and at different times need to be further analyzed in the future to better support the water quality and quantity management of Hulun Lake.

5. Conclusions

(1) In this study, meteorological factors were used as auxiliary variables combined with the Sentinel-2 remote sensing reflection band combination as the input variables, and three machine learning algorithms were used to establish the Hulun Lake Chl-a simulation model. The outcomes demonstrate that the model's precision increased with the addition of auxiliary variables. The RF, AdaBoost, and XGBoost models' R^2 saw increases in the precision of 7.25%, 5.71%, and 7.20% to 0.76, 0.73, and 0.66, respectively. Furthermore, there was some improvement in the precision of both the MAE and RMSE;

(2) In this study, we produced a relatively high-resolution Chl-a distribution map for 2019–2021 based on the most precise RF model and found that the spatial distribution map reflected a similar situation in most selected scenarios, with concentrations in the center of the lake generally lower than those at the edge of the lake. Additionally, high values tend to be concentrated in the northeast, southwest, and south of the lake. Time series analysis shows that the overall concentration of Chl-a throughout the non-ice period tended to increase initially, then decrease, and peak around August. Therefore, it is necessary to increase the monitoring frequency of algal blooms in Hulun Lake during this period;

(3) After more research and analysis into the water quality of the rivers entering the river, it was discovered that the Xinkai River's COD level had reached Class V, and the three other rivers entering the river had varying degrees of pollution in terms of COD, TN, TP, and NH3-N. As a result, we need to monitor the water quality of the rivers that feed into the lake and restrict the amount of contamination that flows downstream.

Author Contributions: Conceptualization, S.Z. and Y.A.; methodology, S.Z.; validation, S.Z. and Y.A.; software, L.W.; formal analysis, S.Z.; investigation, S.Z. and Y.W.; resources, G.M. and Y.A.; data curation, X.Z.; visualization, L.W. and S.Z.; writing—original draft preparation, S.Z.; writing—review and editing, Y.A. and S.Z.; supervision, Y.W. and Y.W.; funding acquisition, Y.Z. and Y.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study is supported by the National Science Fund for Distinguished Young Scholars, No. 52125901; the Key Program of National Natural Science Foundation of China, No. 52339002; and the Natural Science Foundation Youth Program of China, No. 52209041; Fundamental Research Funds for the Central Universities, No. 2233100026.

Data Availability Statement: Sentinel-2 data are openly available for download from the Copernicus Open Access Hub, https://scihub.copernicus.eu/ (accessed on 21 December 2023).

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

References

1. Qin, B.; Gao, G.; Zhu, G.; Zhang, Y.; Song, Y.; Tang, X.; Xu, H.; Deng, J. Lake eutrophication and its ecosystem response. *Chin. Sci. Bull.* **2013**, *58*, 961–970. [CrossRef]
2. Chen, X.; Chuai, X.; Yang, L. Status Quo, Historical Evolution and Causes of Eutrophication in Lakes in Typical Lake Regions of China. *J. Ecol. Rural. Environ.* **2014**, *30*, 438–443.
3. Yu, H.; Shi, X.; Sun, B.; Zhao, S.; Liu, Y.; Zhao, M. Analysis of water quality and eutrophication changes in Hulun Lake from 2011 to 2020. *Arid. Zone Res.* **2021**, *38*, 1534–1545.
4. Song, W.; Yinglan, A.; Wang, Y.; Fang, Q.; Tang, R. Study on remote sensing inversion and temporal-spatial variation of Hulun lake water quality based on machine learning. *J. Contam. Hydrol.* **2024**, *260*, 104282. [CrossRef]
5. Jin, H.; Fang, S.; Chen, C. Mapping of the Spatial Scope and Water Quality of Surface Water Based on the Google Earth Engine Cloud Platform and Landsat Time Series. *Remote Sens.* **2023**, *15*, 4986. [CrossRef]
6. Zhang, C.; Han, M. Mapping Chlorophyll-a Concentration in Laizhou Bay Using Landsat 8 OLI data. In Proceedings of the 36th IAHR World Congress, The Hague, The Netherlands, 28 June–3 July 2015; pp. 12–17.
7. Doernhoefer, K.; Klinger, P.; Heege, T.; Oppelt, N. Multi-sensor satellite and in situ monitoring of phytoplankton development in a eutrophic-mesotrophic lake. *Sci. Total Environ.* **2018**, *612*, 1200–1214. [CrossRef]
8. Chen, Z.; Dou, M.; Xia, R.; Li, G.; Shen, L. Spatiotemporal evolution of chlorophyll-a concentration from MODIS data inversion in the middle and lower reaches of the Hanjiang River, China. *Environ. Sci. Pollut. Res.* **2022**, *29*, 38143–38160. [CrossRef] [PubMed]
9. Sun, S.; Xu, H.; Wu, Y.; Wu, P.; Yang, H. Estimation of Chlorophyll-a Concentration in Chaohu Lake and Nanfei River Based on Sentinel-2 Satellite Remote Sensing Imagery. *J. Hydroecology* **2023**, *44*, 58–66.
10. Ren, J.; Cui, J.; Dong, W.; Xiao, Y.; Xu, M.; Liu, S.; Wan, J.; Li, Z.; Zhang, J. Remote Sensing Inversion of Typical Offshore Water Quality Parameter Concentration Based on Improved SVR Algorithm. *Remote Sens.* **2023**, *15*, 2104. [CrossRef]
11. Virdis, S.G.P.; Xue, W.; Winijkul, E.; Nitivattananon, V.; Punpukdee, P. Remote sensing of tropical riverine water quality using sentinel-2 MSI and field observations. *Ecol. Indic.* **2022**, *144*, 109472. [CrossRef]
12. Cui, Z.; Kerekes, J.P. Potential of Red Edge Spectral Bands in Future Landsat Satellites on Agroecosystem Canopy Green Leaf Area Index Retrieval. *Remote Sens.* **2018**, *10*, 1458. [CrossRef]
13. Yin, Z.; Li, J.; Zhang, B.; Liu, Y.; Yan, K.; Gao, M.; Xie, Y.; Zhang, F.; Wang, S. Increase in chlorophyll-a concentration in Lake Taihu from 1984 to 2021 based on Landsat observations. *Sci. Total Environ.* **2023**, *873*, 162168. [CrossRef] [PubMed]
14. Wang, J.; Tang, J.; Wang, W.; Wang, Y.; Wang, Z. Quantitative Retrieval of Chlorophyll-a Concentrations in the Bohai-Yellow Sea Using GOCI Surface Reflectance Products. *Remote Sens.* **2023**, *15*, 5285. [CrossRef]
15. Yin, B.; Lu, H.; Li, Y.-M.; Wu, C.-Q.; Zhu, L.; Wang, Y.-F. Retrieve phycocyanin concentrations based on semi-analytical model in the Dianchi Lake, China. *Huan Jing Ke Xue Huanjing Kexue* **2011**, *32*, 472–478. [PubMed]
16. Zheng, Z.; Huang, C.; Li, Y.; Lyu, H.; Huang, C.; Chen, N.; Liu, G.; Guo, Y.; Lei, S.; Zhang, R.; et al. A semi-analytical model to estimate Chlorophyll-a spatial-temporal patterns from Orbita Hyperspectral image in inland eutrophic waters. *Sci. Total Environ.* **2023**, *904*, 166785. [CrossRef] [PubMed]
17. Yunmei, L.I.; Jiazhu, H.; Yuchun, W.E.I.; Wanning, L.U. Inversing Chlorophyll Concentration of Taihu Lake by Analytic Model. *J. Remote Sens.* **2006**, *10*, 169–175.
18. Chen, B.; Mu, X.; Chen, P.; Wang, B.; Choi, J.; Park, H.; Xu, S.; Wu, Y.; Yang, H. Machine learning-based inversion of water quality parameters in typical reach of the urban river by UAV multispectral data. *Ecol. Indic.* **2021**, *133*, 108434. [CrossRef]
19. Wu, Z.; Dong, S.; He, W.; Guo, Y.; Li, H.; Lu, R.; Xue, Y. Study on Remote Sensing Inversion of Chlorophyll A in Yangcheng Lake Based on Feature Selection and Machine Learning. *Adm. Tech. Environ. Monit.* **2022**, *34*, 27–32.
20. Karimian, H.; Huang, J.; Chen, Y.; Wang, Z.; Huang, J. A novel framework to predict chlorophyll-a concentrations in water bodies through multi-source big data and machine learning algorithms. *Environ. Sci. Pollut. Res.* **2023**, *30*, 79402–79422. [CrossRef]
21. Wang, J.; Wu, X.; Ma, D.; Wen, J.; Xiao, Q. Remote sensing retrieval based on machine learning algorithm: Uncertainty analysis. *J. Remote Sens.* **2023**, *27*, 790–801. [CrossRef]
22. Yang, Y.; Cao, C.; Pan, X.; Li, X.; Zhu, X. Downscaling Land Surface Temperature in an Arid Area by Using Multiple Remote Sensing Indices with Random Forest Regression. *Remote Sens.* **2017**, *9*, 789. [CrossRef]
23. Im, J.; Park, S.; Rhee, J.; Baik, J.; Choi, M. Downscaling of AMSR-E soil moisture with MODIS products using machine learning approaches. *Environ. Earth Sci.* **2016**, *75*, 1120. [CrossRef]
24. Wang, S.; Zhang, X.; Wang, C.; Chen, N. Multivariable integrated risk assessment for cyanobacterial blooms in eutrophic lakes and its spatiotemporal characteristics. *Water Res.* **2023**, *228*, 119367. [CrossRef] [PubMed]
25. Huisman, J.; Codd, G.A.; Paerl, H.W.; Ibelings, B.W.; Verspagen, J.M.H.; Visser, P.M. Cyanobacterial blooms. *Nat. Rev. Microbiol.* **2018**, *16*, 471–483. [CrossRef] [PubMed]
26. Li, J.; Luo, C.; Lu, H.; Xu, J.; Luo, L.; Pan, M.; He, F.; Man, X.; Zhang, R.; Gong, F.; et al. Spatio-temporal variation and driving factors of algal bloom at Lake Dianchi during 2002–2018. *Acta Ecol. Sin.* **2023**, *43*, 878–891.
27. Hu, L.; Shan, K.; Huang, L.; Li, Y.; Zhao, L.; Zhou, Q.; Song, L. Environmental factors associated with cyanobacterial assemblages in a mesotrophic subtropical plateau lake: A focus on bloom toxicity. *Sci. Total Environ.* **2021**, *777*, 146052. [CrossRef] [PubMed]
28. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
29. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
30. Freund, Y. Boosting a weak learning algorithm by majority. *Inf. Comput.* **1995**, *121*, 256–285. [CrossRef]

31. Chen, T.; Guestrin, C.; Assoc Comp, M. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
32. Ji, T.; Liu, R.; Yang, H.; He, T.; Wu, J. Spatial Downscaling of Precipitation Using Multi-source Remote Sensing Data: A Case Study of Sichuan-Chongqing Region. *J. Geo-Inf. Sci.* **2015**, *17*, 108–117.
33. Wang, J.; Wu, X.; Wen, J.; Xiao, Q.; Gong, B.; Ma, D.; Cui, Y.; Lin, X.; Bao, Y. Upscaling in Situ Site-Based Albedo Using Machine Learning Models: Main Controlling Factors on Results. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4403516. [CrossRef]
34. Wang, L.A.; Zhou, X.; Zhu, X.; Dong, Z.; Guo, W. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop J.* **2016**, *4*, 212–219. [CrossRef]
35. Jiang, G.-J.; Zhou, L.; Ma, R.-H.; Duan, H.-T.; Shang, L.-L.; Rao, J.-W.; Zhao, C.-L. Remote sensing retrieval for chlorophyll-a concentration in turbid case II waters (II): Application on MERIS image. *J. Infrared Millim. Waves* **2013**, *32*, 372–378. [CrossRef]
36. Dan, Y.; Zhou, Z.; Li, S.; Zhang, H.; Jiang, Y. Retrieval of chlorophyll-a concentration in pingzhai reservoir based on sentinel-2. *Environ. Eng.* **2020**, *38*, 180–185.
37. Xie, T.; Chen, Y.; Lu, W.; Wang, X. Comparison and analysis of chlorophyll-a retrieval model in the lower reaches of Minjiang River based on GF-1 WFV image. *Acta Sci. Circumstantiae* **2019**, *39*, 4276–4283.
38. El Din, E.S. A novel approach for surface water quality modelling based on Landsat-8 tasselled cap transformation. *Int. J. Remote Sens.* **2020**, *41*, 7186–7201. [CrossRef]
39. Pu, F.; Ding, C.; Chao, Z.; Yu, Y.; Xu, X. Water-Quality Classification of Inland Lakes Using Landsat8 Images by Convolutional Neural Networks. *Remote Sens.* **2019**, *11*, 1674. [CrossRef]
40. Zhang, H.; Xue, B.; Wang, G.; Zhang, X.; Zhang, Q. Deep Learning-Based Water Quality Retrieval in an Impounded Lake Using Landsat 8 Imagery: An Application in Dongping Lake. *Remote Sens.* **2022**, *14*, 4505. [CrossRef]
41. Meyer, H.; Kuehnlein, M.; Appelhans, T.; Nauss, T. Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. *Atmos. Res.* **2016**, *169*, 424–433. [CrossRef]
42. Qian, X.; Li, J.; Ao, W.; Pang, B.; Bao, S.; Wang, Q.; Liu, B.; Wang, Z. Seasonal dynamics of phytoplankton and its relationship with environmental factors in Lake Hulun. *J. Lake Sci.* **2022**, *34*, 1814–1827.
43. Cao, M.; Qing, S.; Du, Y.; Yuan, R.; Shun, B. Remote sensing monitoring of algal blooms in Hulun Lake based on SMDPSO algorithm. *J. Water Resour. Water Eng.* **2021**, *32*, 66–72, 80.
44. Guo, Z.; Li, C.; Shi, X.; Sun, B.; Zhao, S.; Quan, D.; Hou, B. Spatial and Temporal Distribution Characteristics of Chlorophyll A Content and Its Influencing Factor Analysis in Hulun Lake of Cold and Dry Areas. *Ecol. Environ. Sci.* **2019**, *28*, 1434–1442.
45. Liang, L.; Li, C.; Sun, B.; Wang, J.; Han, Z. Impact Factors and Characteristics of Water Quality Variation of Hulun Lake in Inner Mongolia Automous Region. *Bull. Soil Water Conserv.* **2017**, *37*, 102–106.
46. Li, N.; Li, Y.; Feng, J.-C.; Shan, Y.-J.; Qian, J.-N. Construction and Application Optimization of the Chl-a Forecast Model ARIMA for Lake Taihu. *Huan Jing Ke Xue Huanjing Kexue* **2021**, *42*, 2223–2231. [CrossRef] [PubMed]
47. Fu, X.; Zheng, M.; Su, J.; Xi, B.; Wei, D.; Wang, X. Spatiotemporal patterns and threshold of chlorophyll-a in Lake Taihu based on microcystins. *Environ. Sci. Pollut. Res.* **2023**, *30*, 49327–49338. [CrossRef]
48. Jiang, X.; Wang, S.-H.; Zhong, L.-X.; Jin, X.-C.; Sun, S.-Q. Seasonal variation characteristics of algae biomass in Chaohu Lake. *Huan Jing Ke Xue Huanjing Kexue* **2010**, *31*, 2056–2062.