remote sensing

MDPI

Article

Retrieval of Chlorophyll-a Concentrations Using Sentinel-2 MSI Imagery in Lake Chagan Based on Assessments with Machine Learning Models

Xuming Shi [1], Lingjia Gu [1,*], Tao Jiang [2], Xingming Zheng [2], Wen Dong [3] and Zui Tao [3]

[1] College of Electronic Science & Engineering, Jilin University, Changchun 130012, China
[2] Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130102, China
[3] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China
[*] Correspondence: gulingjia@jlu.edu.cn

Abstract: Chlorophyll-a (Chl-a) is an important characterized parameter of lakes. Monitoring it accurately through remote sensing is thus of great significance for early warnings of water eutrophication. Sentinel Multispectral Imager (MSI) images from May to September between 2020 and 2021 were used along with in-situ measurements to estimate Chl-a in Lake Chagan, which is located in Jilin Province, Northeast China. In this study, the extreme gradient boosting (XGBoost) and Random Forest (RF) models, which had similar performances, were generated by six single bands and six band combinations. The RF model was then selected based on the assessments ($R^2 = 0.79$, RMSE = 2.51 $\mu g\ L^{-1}$, MAPE = 9.86%), since its learning of the input features in the model conformed to the bio-optical properties of Case 2 waters. The study considered Chl-a concentrations in Lake Chagan as a seasonal pattern according to the K-Nearest-Neighbors (KNN) classification. The RF model also showed relatively stable performance for three seasons (spring, summer and autumn) and it was applied to map Chl-a in the whole lake. The research presents a more reliable machine learning (ML) model with higher precision than previous empirical models, as shown by the effects of the input features linked with the biological mechanisms of Chl-a. Its robustness was revealed by the temporal and spatial distributions of Chl-a concentrations, which were consistent with in-situ measurements in the map. This research was capable of revealing the current ecological situation in Lake Chagan and can serve as a reference in remote sensing of inland lakes.

Keywords: Chlorophyll-a; Lake Chagan; Sentinel-2; seasonal pattern; machine learning

check for updates

Citation: Shi, X.; Gu, L.; Jiang, T.; Zheng, X.; Dong, W.; Tao, Z. Retrieval of Chlorophyll-a Concentrations Using Sentinel-2 MSI Imagery in Lake Chagan Based on Assessments with Machine Learning Models. *Remote Sens.* **2022**, *14*, 4924. https://doi.org/10.3390/rs14194924

Academic Editors: Pingping Luo, Xindong Wei, Kanhua Yu, Bin Guo and Joshua Viers

Received: 6 September 2022
Accepted: 27 September 2022
Published: 1 October 2022

1. Introduction

Gross primary productivity (GPP) is closely linked to various aspects of the global carbon cycle [1]. Solar-induced chlorophyll-a (Chl-a) fluorescence (SIF), which is emitted by Chl-a, has been beneficial for monitoring crop productivity, an important part of GPP [2]. Moreover, Chl-a is crucial for modeling plant health in the terrestrial carbon cycle [3]. In another major area, Chl-a is one of the primary pigments in the phytoplankton of water bodies and can be used to describe phytoplankton biomass [4]. Accordingly, the ecological condition of lakes, which is strongly linked to GPP and the global carbon cycle, can be observed through the Chl-a concentration.

With the emergence of remote sensing, the measurement of water quality parameters has been greatly simplified [5]. Compared with traditional methods, the retrieval of Chl-a by remote sensing has more temporal and spatial advantages [6]. Early studies using ocean satellites found the band ratios of reflectance in the blue and green bands were useful for the ocean [7,8]; these were mainly determined by Chl-a. Some inland and coastal waters (Case 2 waters), however, have more complex bio-optical properties because of the interaction caused by Chl-a, total suspended sediment (TSS) and colored dissolved organic matter (CDOM) [9]. In addition, Chl-a retrieval is also influenced by cloud cover. It is necessary to improve atmospheric correction algorithms specifically for Case 2 waters [10]. All these factors make the retrieval of Chl-a a challenging task [11].

Many efforts have been made to retrieve Chl-a, including selecting the appropriate sensor or satellite and optimizing the model based on band combinations. The temporal resolution of the Moderate Resolution Imaging Spectroradiometer (MODIS) is 1 day, which makes long-term observations feasible [12]. However, its spatial resolution of 250–1000 m makes observations difficult in some regional lakes, since medium or small lakes require a higher spatial resolution (10–30 m). With a medium spatial resolution (30 m), Landsat's Thematic Mapper (TM) and Operational Land Imager (OLI) sensors have been applied for Chl-a retrieval [13,14]. However, considering the limited spectral bands of the sensors onboard Landsat and the longer revisit period, it is difficult to retrieve Chl-a accurately [15]. The Sentinel Multispectral Imager (MSI) has a good overall temporal resolution of 5 days and a spatial resolution of 10–60 m [16]. It contains three bands in the red-edge range (705 nm, 740 nm and 783 nm), which are able to capture the subtle Chl-a information [17]. Therefore, the Sentinel Multispectral Imager (MSI) can be a more suitable sensor for Chl-a estimations.

On the other hand, numerous algorithms have been proposed for estimating Chl-a, including empirical algorithms and analytical algorithms. Analytical algorithms focus on the mechanism of radiation in water, making the model physically explainable [18,19]. However, the cumbersome analysis of complex formulas requires a higher level of computation and model derivation. Empirical algorithms include the band ratio algorithm [20], the three-band algorithm [21], the four-band algorithm [22] and the enhanced three-band index [23]. These algorithms choose appropriate bands from different satellites to develop models in the form of a regression based on in-situ measurements. Since different waters always have various parameters, the interactions among TSS, CDOM and Chl-a cause difficulties for large-scale applications [24]. As a branch of artificial intelligence, Machine Learning (ML) algorithms have been applied to retrieve Chl-a [25].

ML algorithms can have multiple input variables and have achieved great robustness in estimation [26,27]. These ML models include convolutional neural networks [28] (CNN), neural networks (CNN-LTSM) [29], Random Forest (RF) [10], extreme gradient boosting (XGBoost) [30], support vector machine (SVM) [24], and the light gradient boosting machine (LGBM) [1]. However, it is hard to observe the learning process of ML models due to their opaque structures. Analyzing the features' importance is useful [15], but the explanations are still limited. The recently developed Shapley additive explanation (SHAP) is able to provide further visualization in ML models [31,32]. SHAP is capable of interpreting complex ML models by observing how the features influence the predicted values in the learning process and assessing the interactive effects of the input features with a series of graphs. Thus, the validity of model can be evaluated with high ecological relevance for Chl-a retrieval [33].

This study combined Sentinel-2 images with in-situ measurements to develop a reliable ML model for Chl-a retrieval with good performance, as determined through an assessment of the ML models. Specifically, the major contributions of this study included (1) developing XGBoost and RF models for Chl-a retrieval with higher accuracy compared with empirical algorithms; (2) assessing the two models on the basis of the bio-optical properties of Case 2 waters through SHAP, after which the RF model was selected to estimate Chl-a concentration; and (3) applying the K-nearest neighbors (KNN) method to observe the seasonal patterns in Lake Chagan and analyze the temporal and spatial variations in the Chl-a concentration between 2020 and 2021. The article is structured as follows. The datasets are described in Section 2, RF model is presented and its seasonal performances are analyzed in Section 3, and Section 4 presents a discussion on the experimental results, followed by the conclusion in Section 5.

2. Study Area and Data

2.1. Study Area

Lake Chagan (45°09′–45°30′N, 124°03′–124°34′E) is one of the largest freshwater lakes in northeast China [34]. It is located in the hinterland of the Northeast Songnen Plain, most of which is located in the former Golros Mongolian Autonomous County in the northwest of Jilin Province, as shown on the left of Figure 1. Lake Chagan spans 38 km from east to west and 14 km from north to south. Since Lake Chagan is a National Natural Reserve, the lake and water network provide protection for many large fish breeding bases. As a typical Case 2 water, its eutrophication needs to be monitored. With the continuous development of tourism in recent years, the environment is being threatened. In this study, 5 periodic field measurements in Lake Chagan were conducted and a total of 98 samples were collected between 2020 and 2021, with a seasonal coverage from spring to autumn (May to September). The entire area of Lake Chagan can be divided into two parts. Alongside the main lake area, Lake Xinmiao is located to the southeast of the main lake area, as shown on the right of Figure 1.

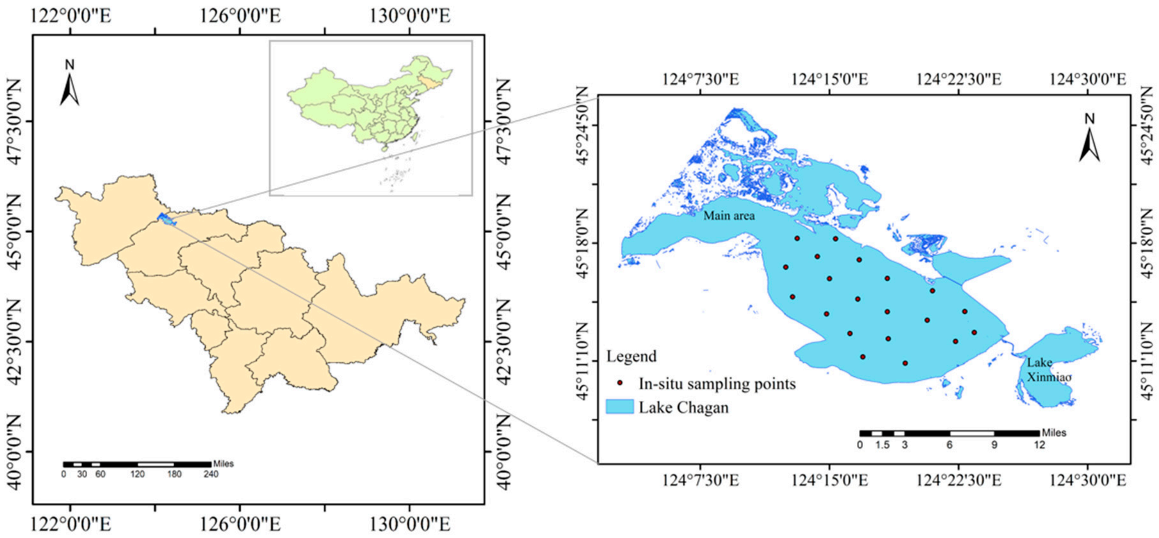


Figure 1. Study area and field measurements. The left-hand part shows the geographical location and the right-hand part shows the composition of the lake and distribution of sampling points in the field.

2.2. In-Situ Data Collection

This study used an irradiance sensor and two radiance sensors which can measure wavelengths ranging from 320 nm to 950 nm with a spectral sampling interval of 3.3 nm to measure the in-situ hyperspectral data. One radiance sensor pointing to the sky was set at 40° from the normal direction of the surface to measure the sky's radiance (L_{sky}), one radiance sensor pointing to the water's surface was set at 40° from the normal direction of the surface to measure the total radiance leaving the water (L_{sw}) and one irradiance sensor pointed vertically to the sky to measure the total downwelling irradiance (E_d (0^+)). Remote sensing reflectance (R_{rs}) was then obtained as shown in Equation (1) [35]

$$R_{rs} = (L_{sw} - r \times L_{sky})/E_d(0^+) \tag{1}$$

where r refers to the wind speed, which was assumed to be 0.025. The water surface spectrometer was about 1 m over the hull and the benchmark was projected plus or minus 15 degrees. Each point was measured 10 times. Considering the extent and water quality of the area, 20 sample points were evenly selected, which were able to cover the main area of the entire lake (Figure 1).

An EXO multi-parameter water quality meter from YSI was used to measure six parameters simultaneously including the water surface temperature, the Chl-a concentration, the pH value, depth, etc. The sky was clear, and the sampling time was mostly between 10:00 and 14:30 each day, which is near the overpass time of the Sentinel-2 satellite. Each sampling point covered an interval longer than 3 min and was measured two times then averaged (Table 1). Water samples collected in situ were analyzed for TSS content (mg L^{-1}) in the laboratory (Equation (2)). A certain volume of the samples was filtered through a filter membrane and dried at 103–105 °C for 1 h. The sample was then cooled at room temperature and weighed by repeated drying. When the weight difference between the latest two times of drying was less than 0.4 mg, Weight A was recorded. The filter membrane was pre-processed in a drying oven at 103–105 °C for one hour, and was also cooled at room temperature, weighed and dried repeatedly. When the weight difference between the latest two times of drying was less than 0.2 mg, Weight B was recorded.

$$TSS = \frac{(A - B) \times 10^6}{V} \tag{2}$$

where V refers to the volume of the samples; A refers to the total weight of the TSS, the filter membrane and the weighing bottle; and B refers to the weight of the filter membrane and the weighing bottle.

Table 1. In-situ measurements of Chl-a.

In-Situ Date	Number of Points	Average (μg L^{-1})
22 July 2020	20	20.78
21 August 2020	18	19.61
26 September 2020	20	24.46
26 May 2021	20	15.60
18 July 2021	20	28.47

2.3. Satellite Data

Sentinel-2 offers valuable information for remote sensing of inland water [36]. Remote sensing images of Sentinel-2 (Sentinel-2A and -2B) were accessed from the Sentinel data distribution system (https://scihub.copernicus.eu./). The parameters of Sentinel-2 are shown in Table 2. Since the Sentinel-2A and Sentinel-2B satellites have the same sensor on board, the revisit period of five days has advantages for long-term observations because of the collaboration of the two satellites. Bands in the MSI images consist of visible light, near infrared (VNIR) and short-wave infrared (SWIR), which can achieve a high level of monitoring. ESA provides Level 2A products that include atmospheric corrected bottom layer reflectance data, known as Bottom-Of-Atmosphere (BOA) reflectance. In this study, 20 cloud free Level-2A images were downloaded and matched with the in-situ measurements. The matched dataset considered a time window with a maximum of 8 days, and the average was 3.2 days (Table 3). All bands of the images were resampled to a 10 m spatial resolution through the bilinear sampling method.

Table 2. Bands of Sentinel-2 used in this study and their parameters.

Sentinel-2 Bands	Central Wavelength (nm)	Resolution (m)
Band 1—Coastal aerosol	443	60
Band 2—Blue	490	10
Band3—Green	560	10
Band 4—Red	665	10
Band 5—Vegetation red edge	705	20
Band 6—Vegetation red edge	740	20
Band 8—NIR	842	10

Table 3. Comparison of selected Sentinel-2 images and in situ measurements.

Year	In-Situ Date	Sentinel-2 Date
2020	22 July	22 July
2020	21 August	21 August
2020	26 September	30 September
2021	26 May	18 May
2021	18 July	22 July

3. Methodology

All input variables were used for training the XGBoost and RF models, and their performance on the validation set was evaluated (Figure 2). An assessment based on SHAP was used to observe the learning of each point in these two models, and the more reliable model was selected in combination with the bio-optical properties of Case 2 waters. KNN analysis was used to classify the Chl-a concentrations and obtain the seasonal patterns. RF was further applied in different seasons, and, finally, the Chl-a was mapped across the whole lake.

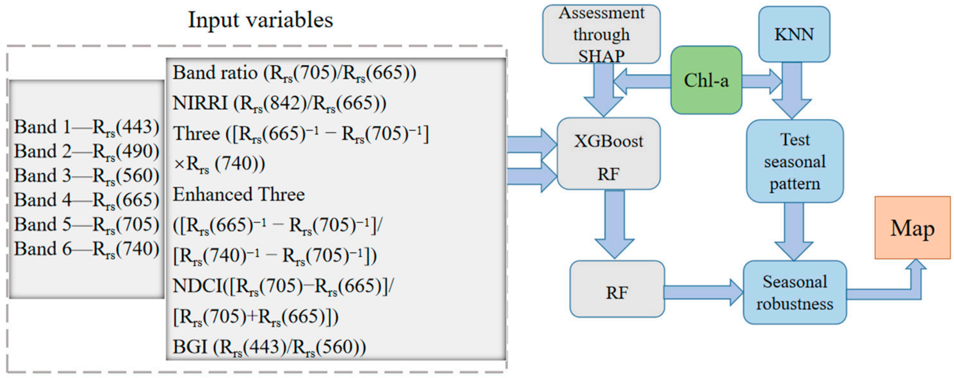

Input variables
Band 1—$R_{rs}(443)$
Band 2—$R_{rs}(490)$
Band 3—$R_{rs}(560)$
Band 4—$R_{rs}(665)$
Band 5—$R_{rs}(705)$
Band 6—$R_{rs}(740)$
Band ratio ($R_{rs}(705)/R_{rs}(665)$)
NIRRI ($R_{rs}(842)/R_{rs}(665)$)
Three ($[R_{rs}(665)^{-1} - R_{rs}(705)^{-1}]$
$\times R_{rs}(740)$)
Enhanced Three
($[R_{rs}(665)^{-1} - R_{rs}(705)^{-1}]/$
$[R_{rs}(740)^{-1} - R_{rs}(705)^{-1}]$)
NDCI($[R_{rs}(705)-R_{rs}(665)]/$
$[R_{rs}(705)+R_{rs}(665)]$)
BGI ($R_{rs}(443)/R_{rs}(560)$)
Assessment through SHAP
Chl-a
KNN
XGBoost
RF
Test seasonal pattern
Map
RF
Seasonal robustness

Figure 2. Flowchart of the proposed RF model for estimating Chl-a. The left-hand panel shows the input variables and the right-hand panel shows the evaluation and application processes.

3.1. Model Structure

In this study, models were developed using field measurements and Sentinel-2 BOA reflectance to estimate the Chl-a concentration. The selection of input variables included the first six bands of Sentinel-2 (443 nm, 490 nm, 560 nm, 665 nm, 705 nm, 740 nm) and six band combinations based on empirical algorithms that have been applied in inland lakes. The bands listed in the reference model were replaced with the similar bands from Sentinel-2 (Table 4). The output was the Chl-a concentration (Figure 2). In total, the 98 points included the BOA reflectance and the relative Chl-a concentration measured in situ. The training set and the validation set were at a ratio of 7:3. The average Chl-a concentration in the training set (N = 68) and validation set (N = 30) was similar, 21.87 μg L^{-1} and 21.72 μg L^{-1}, respectively. Before we used SHAP to explain the model, it was necessary to develop a ML model with relatively strong robustness and good performance [37]. A grid search strategy combined with 10-fold cross-validation was used to tune the hyperparameters of the XGBoost and RF models.

As introduced in [38,39], XGBoost and RF are both tree models, which are both made up of a number of trees and finally produce an abundance of trees. RF constructs a decision tree by randomly adding back the extracted sampling set and randomly unpacking the extracted feature set. Via the method of bootstrap resampling technology, the final prediction results of the model are determined by voting on a number of key factors. XGBoost explicitly adds regularization to the object function items to prevent the model from overfitting and to control the complexity of the model.

Table 4. Sources of the band combinations used in the existing algorithms.

Algorithm	Index	Reference
Band ratio	$R_{rs}(709)/R_{rs}(665)$	Duan et al. [20]
Three	$[R_{rs}(671)^{-1} - R_{rs}(710)^{-1}] \times R_{rs}(740)$	Gitelson et al. [21]
Enhanced Three	$[R_{rs\,(}665)^{-1} - Rrs(705)^{-1}]/$ $[R_{rs}\ (740)^{-1} - R_{rs}(705)^{-1}]$	Yang et al. [23]
NIRRI	$R_{rs}(865)/R_{rs}(655)$	Duan et al. [13]
NDCI	$[R_{rs}\ (708)^{-1} - R_{rs}(665)^{-1}]/$ $[R_{rs}\ (708)^{-1} + R_{rs}(665)^{-1}]$	Mishra et al. [40]
BGI	$R_{rs}(443)/R_{rs}(561)$	Nguyen et al. [41]

3.2. SHAP Method

SHAP (SHapley Additive exPlanations) was applied to assess the ML models [32]. It can be used for interpreting the predictions of ML models by calculating the contribution of each feature to the prediction. The proposed SHAP value was used as a united approach to explain the output of the ML model, including the benefits both in terms of global interpretability and local interpretability. SHAP originates from the Shapley value, which was initially a method of allocating expenditure according to individual contributions through the average of the marginal contributions across all permutations. Interpretations of the Shapley value are expressed as an additive feature attribution method by SHAP and the predicted value of the model is interpreted as the sum of the attribution values of each input feature in SHAP. SHAP thus actually attributes the output value to the Shapley value of each feature. The calculation principle of the Shapley value can be expressed as Equation (3)

$$\Phi_j = \sum_{S \subseteq \{x_1, \cdots, x_p\} \backslash \{x_j\}} \frac{|S|!(p - |S| - 1)!}{p!} \left(f_x\left(S \cup \{x_j\}\right) - f_x(S)\right) \quad (3)$$

where $\{x_1, \cdots, x_p\}$ is the set of all input features, p is the number of features, $\{x_1, \cdots, x_p\} \backslash \{x_j\}$ is the feature set except for $\{x_j\}$ and $f_x(S)$ is the prediction for feature subset S. When the relationships among all the input features are nonlinear or the features are not independent, the SHAP values calculate the weighted average value for all possible features. SHAP combines these conditional expectations with the values from the classical Shapley value of game theory to sum each attribution value to Φ_j.

3.3. KNN for the Seasonal Pattern Test

Chl-a concentrations showed a seasonal distribution in Lake Taihu [10]. After the Chl-a concentrations had been labeled, KNN classification was used to analyze whether Lake Chagan also had a seasonal pattern. The K-nearest neighbors (KNN) algorithm [42] was successfully applied for this classification [43]. KNN decides the class which the sample to be classified belongs to according to the class of the nearest sample or several samples. Its performance thus largely depends on the distance metric used to identify the nearest neighbors. The value of K is also of great importance. If a larger value of K is selected, it is equivalent to using the training examples in a larger neighborhood for prediction, and the approximate error of learning will increase. In this case, training instances far away from the input instance would make the prediction wrong and prone to underfitting.

3.4. Accuracy Assessment

As is shown in Equations (4)–(6), the coefficient of determination (R^2), the root mean square error ($RMSE$) and the mean absolute percentage error ($MAPE$) were used to evaluate the models' performance. The F1 score was applied for evaluating the results of the KNN classification.

$$R^2 = \frac{\sum\limits_{i=1}^{N} \left(E_i - \overline{M}\right)^2}{\sum\limits_{i=1}^{N} \left(M_i - \overline{M}\right)^2} \tag{4}$$

$$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N} (M_i - E_i)^2} \tag{5}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} \frac{|M_i - E_i|}{M_i} \tag{6}$$

where N is the total number of points, i represents each sample, and M and E refer to the measured and estimated Chl-a concentrations, respectively.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

where Precision represents the proportion that is actually positive in all cases where the prediction is positive and Recall represents the proportion that has been predicted correctly in all actual positive examples.

4. Results

4.1. Development and Assessment of the Models

4.1.1. Performance of XGBoost and RF

Both the XGBoost model and the RF model performed well on the validation set, as shown by the statistical metrics. The XGBoost model (R^2 = 0.80, RMSE = 2.42 μg L^{-1}, MAPE = 9.55%) and the RF model (R^2 = 0.79, RMSE = 2.51 μg L^{-1}, MAPE = 9.86%) are shown in Figure 3. The retrieved value and the real value were evenly distributed on both sides of the line. In the XGBoost model, 13 points of the retrieved Chl-a concentrations were smaller than the measured values and 17 points for the Chl-a concentration were larger than the measured values, while the RF model showed a similar performance. At the same time, both models had a tendency to overestimate the Chl-a values with a low concentration (<15 μg L^{-1}) and slightly underestimate the Chl-a values with a high concentration (>30 μg L^{-1}). The performances of three evaluation indicators were close, and XGBoost performed better. After this analysis, it was reasonable to select the XGBoost model to estimate the Chl-a concentrations.

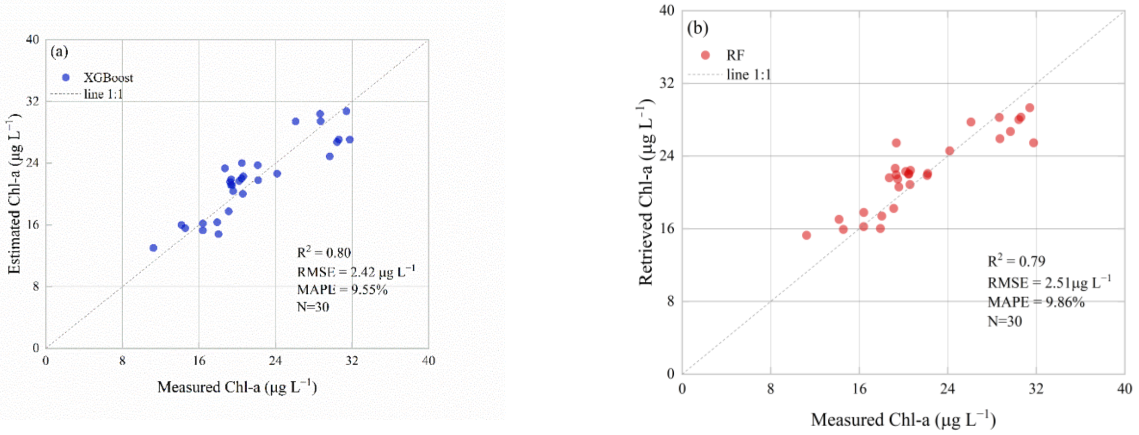
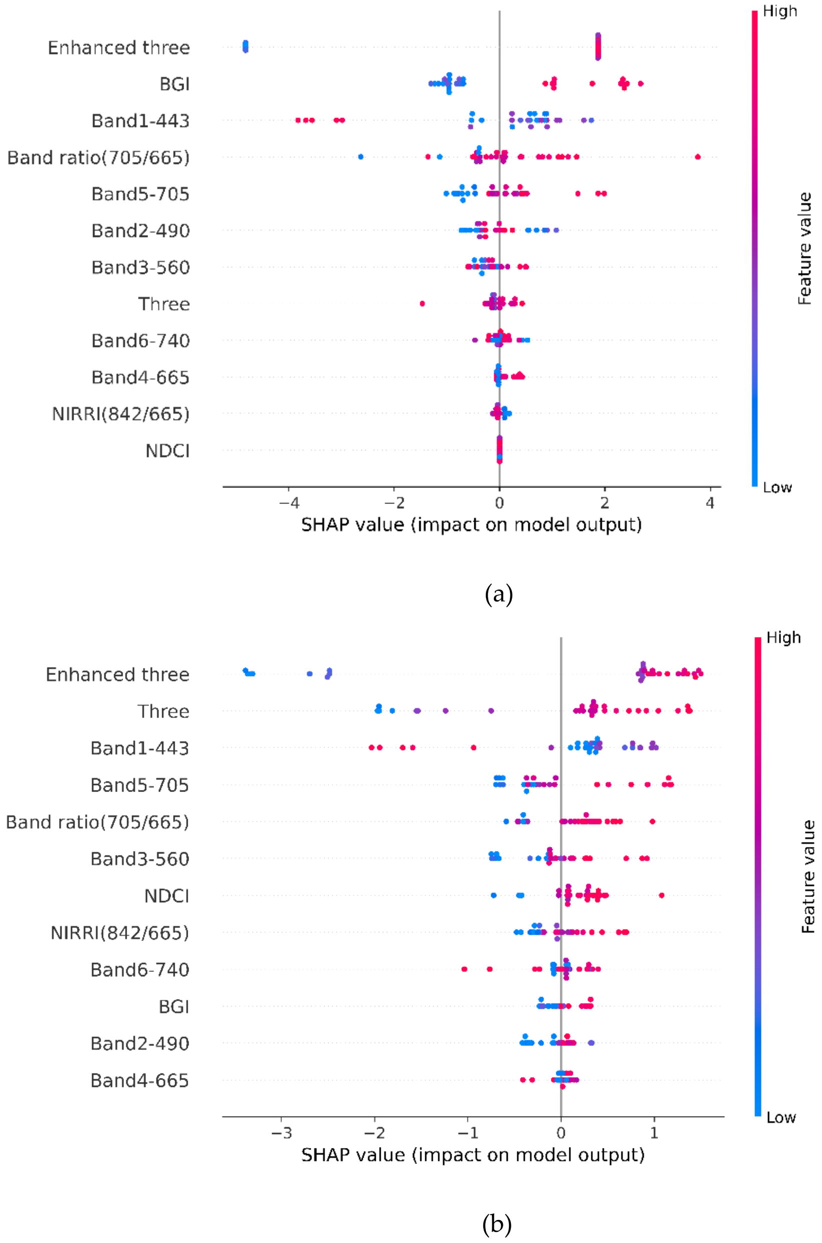


Figure 3. Performance of the validation results of the XGBoost and RF models for the Chl-a concentrations. (**a**) XGBoost model; (**b**) RF model. The straight line is 1:1, the horizontal axis is the measured value and the vertical axis is the retrieved value.

4.1.2. Assessment through SHAP

The XGBoost model had better performance; however, the internal learning process of the two ML models was unknown. The retrieval of Chl-a is mainly based on two absorption peaks in the spectrum, usually located near 440 nm and 675 nm. The area around 440 nm is greatly affected by TSS and CDOM, while the area around 675 nm is less affected by other water factors. Therefore, the area around 675 nm is usually selected as the spectrum for retrieving the concentration of Chl-a [44]. In the RF model (Figure 4), for the feature of Band 1—443 nm and the feature of Band 4—665 nm, if the value was lower, the output had a higher value of Chl-a, which was a very obviously negative correlation. This was consistent with the principle of retrieving Chl-a in most inland lakes due to the strong absorption of Chl-a at around 443 nm and 675 nm.



(a)

(b)

Figure 4. Figure of the features of density scatter: (**a**) XGBoost; (**b**) RF. The features are ranked by importance from top to bottom on the vertical axis. On the horizontal axis, red represents a high correlation, and blue represents a low correlation. This illustrates the relationship between the input features and the output Chl-a values.

For the XGBoost model, however, the feature of Band 1—443 nm also showed the same negative correlation while the feature of Band 4—665 nm and the Chl-a values showed a positive correlation, which was a violation of the bio-optical properties of Case 2 waters. It should be noted that a few samples in the RF model appeared on the right half of the RF model and seemed to be positively correlated. This may have been due to atmospheric refraction. By contrast, almost all of the points in the XGBoost model appeared on the right half, which was obviously abnormal for this band. Besides, the feature of NDCI should have a single correlation with Chl-a that is positive. Conversely, the SHAP value in the XGBoost model was zero, which meant this input variable was learned in such a way that half had a positive correlation and half had a negative correlation. In the RF model, it was positive.

Furthermore, the learning process of the other input features was also observed. The features of Band 1—443 nm, Band Ratio (705/665), NIRRI (842/665), Enhanced Three, Three and BGI were all consistent with the correlation of the corresponding empirical algorithm in the RF model. Enhanced Three, Band Ratio (705/665), NIRRI (842/665) and BGI all showed a positive correlation, which were also consistent with functional forms of the related empirical algorithms. Thus, the XGBoost model showed incorrect learning of the input features, especially for the features of Band 4—665 nm and NDCI. Accordingly, the RF model was developed and selected for estimating Chl-a.

4.2. Spatial and Temporal Analysis of Chl-a

4.2.1. Test for Seasonal Patterns of Chl-a Concentrations

In order to verify the seasonal patterns of Chl-a concentrations in Lake Chagan, the KNN method was applied to identify the Chl-a. During the process, each sample of Chl-a concentration was labeled for the corresponding season and then analyzed to determine whether the Chl-a concentrations had a seasonal pattern by considering the classification results of the model (Figure 5). Since the observation period of the field measurements was from May to September, the seasonal observations included three months in summer and only one month each in spring and autumn. In the classification model, by taking the F1 score (Equation (7)) as the index [45], a 6:4 division of the training and validation set was generated. The concentration of measured Chl-a and retrieved Chl-a in each season could be observed. The same method of k-fold (k = 10) cross-validation was used and a F1 score of 0.81 was finally obtained. A few points in summer were misidentified as spring and autumn, which may have been due to the range of Chl-a concentrations in summer being larger and because there were more data points (N = 58). According to the results, the Chl-a concentration for September in autumn was close to that of July and August in summer, which caused a few misclassified points. However, the predicted values for the three seasons are close to the average level of the true values. In addition, the points in KNN are divided into groups according to the distance and weights [46]. Therefore, the process of classification in the validation set refers to concentrations that were adjacent to those presented in the training set, and these points were prone to be divided into labels corresponding to the training set. In general, this result at least illustrates that the Chl-a concentrations of Lake Chagan show a relatively seasonal trend and that it is necessary to test the performance of RF in different seasons.

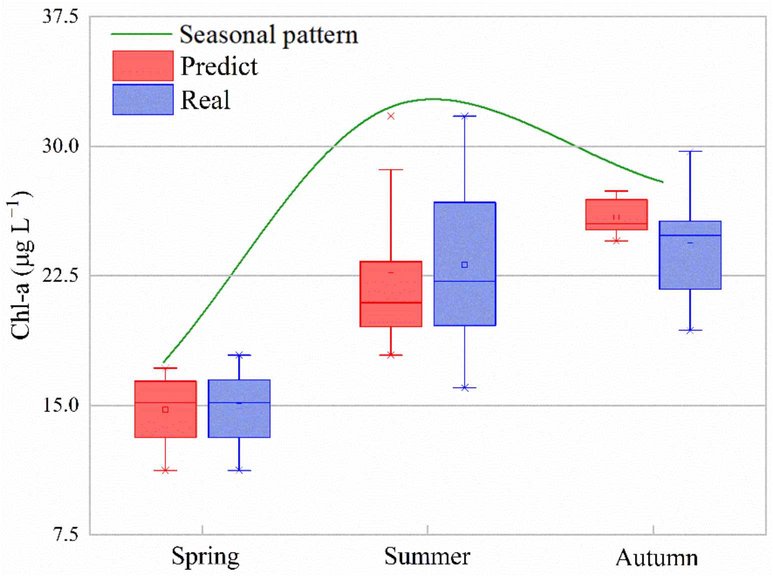


Figure 5. Box plot of seasonal classification results of the validation set achieved by KNN.

4.2.2. Robustness Analysis of RF in Three Seasons

The retrieved Chl-a concentrations from spring to autumn in 2020 and 2021 were compared with the measured Chl-a concentrations to observe the performance of the RF model in each season. The RF model performed well overall in these three seasons (Figure 6). The performance for summer is the best (R^2 = 0.86, RMSE = 1.75 μg L^{-1}, MAPE = 5.75%), followed by spring (R^2 = 0.62, RMSE = 1.25 μg L^{-1}, MAPE = 6.40%) and autumn (R^2 = 0.56, RMSE = 1.74 μg L^{-1}, MAPE = 5.22%). The performances for summer and autumn were slightly lower in terms of R^2, but the model also performed well in terms of RMSE and MAPE. The reason for this may be that spring and autumn had fewer sampling points compared with summer. On the other hand, the Chl-a concentrations at each point within the months were close and there was no sudden increase or decrease in the concentration. In summary, considering the comprehensive performance based on the three evaluation indicators for the three seasons and all points (Figure 3b), the RF model had relatively robustness and it is feasible for long-term retrieval of data for the whole lake.

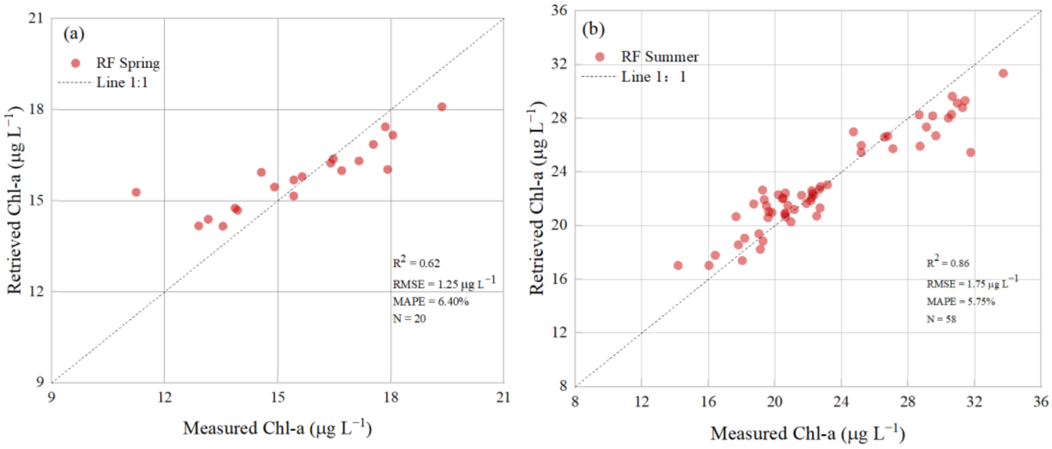


Figure 6. *Cont.*

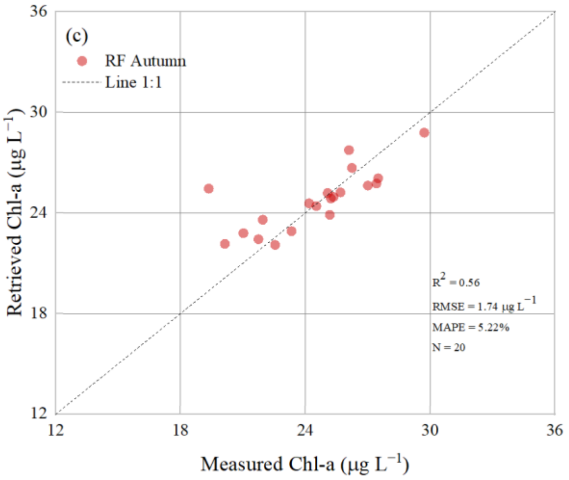


Figure 6. The RF model's performance on validation set in different seasons: (**a**) spring, (**b**) summer and (**c**) autumn.

4.2.3. Seasonal and Spatial Analysis

To observe the temporal and spatial distribution of Chl-a in Lake Chagan, the RF model was applied to retrieve the Chl-a concentrations based on the field sampling points and the Sentinel-2 images. The dates of the Sentinel-2 images used for seasonal averages of Chl-a were close to those of the in situ measurements. Because of cloud and the limited number of Sentinel-2 images, the annual seasonal average Chl-a concentrations may not fully reflect the actual situation, especially in spring and autumn. In general, however, a seasonal pattern of Chl-a was observed. Initially, the retrieved Chl-a concentrations of RF related to the in-situ sampling points in each season were averaged for 2020 and 2021 (Figure 7).

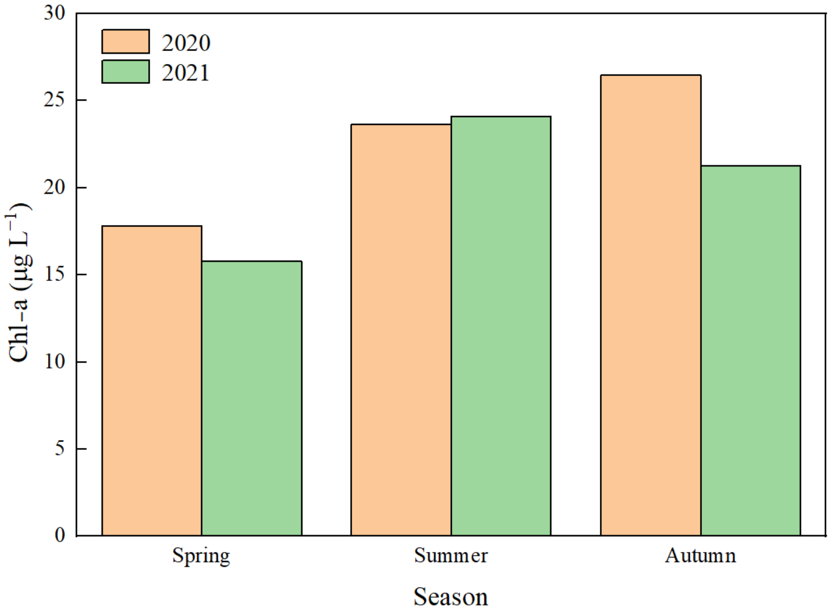


Figure 7. Average Chl-a concentrations retrieved by the RF model between 2020 and 2021 for sampling points in the field.

It can be observed that the Chl-a concentration was lowest in spring during these two years (16.05 ± 0.28 μg L^{-1}) and it was close in summer (22.87 ± 1.21 μg L^{-1}) and autumn (22.75 ± 1.51 μg L^{-1}). The retrieved Chl-a values were obtained pixel by pixel to give a general view of the whole lake (Figure 8). MSI-derived Chl-a had distinct seasonal patterns across Lake Chagan from 2020 to 2021. Chl-a concentrations were substantially higher during summer (June–August) and autumn (September) in comparison with spring (May). Among the in situ points, the Chl-a concentration in spring and autumn in 2020 was higher than that it was in 2021, while that in summer was lower than and nearly equal to that in 2021. The maximum value appeared in September 2020, which was consistent with the results of the in situ measurements. Summer, a special season for the growth of Chl-a, has some changes in the climate and environment. For lakes in northeastern China, only the months from May to September are suitable for in situ sampling. More points of data in summer are not only conducive to a comprehensive assessment of the growth of Chl-a but also for monitoring the overall changes in the lake ecosystem, combined with observations in spring and autumn.

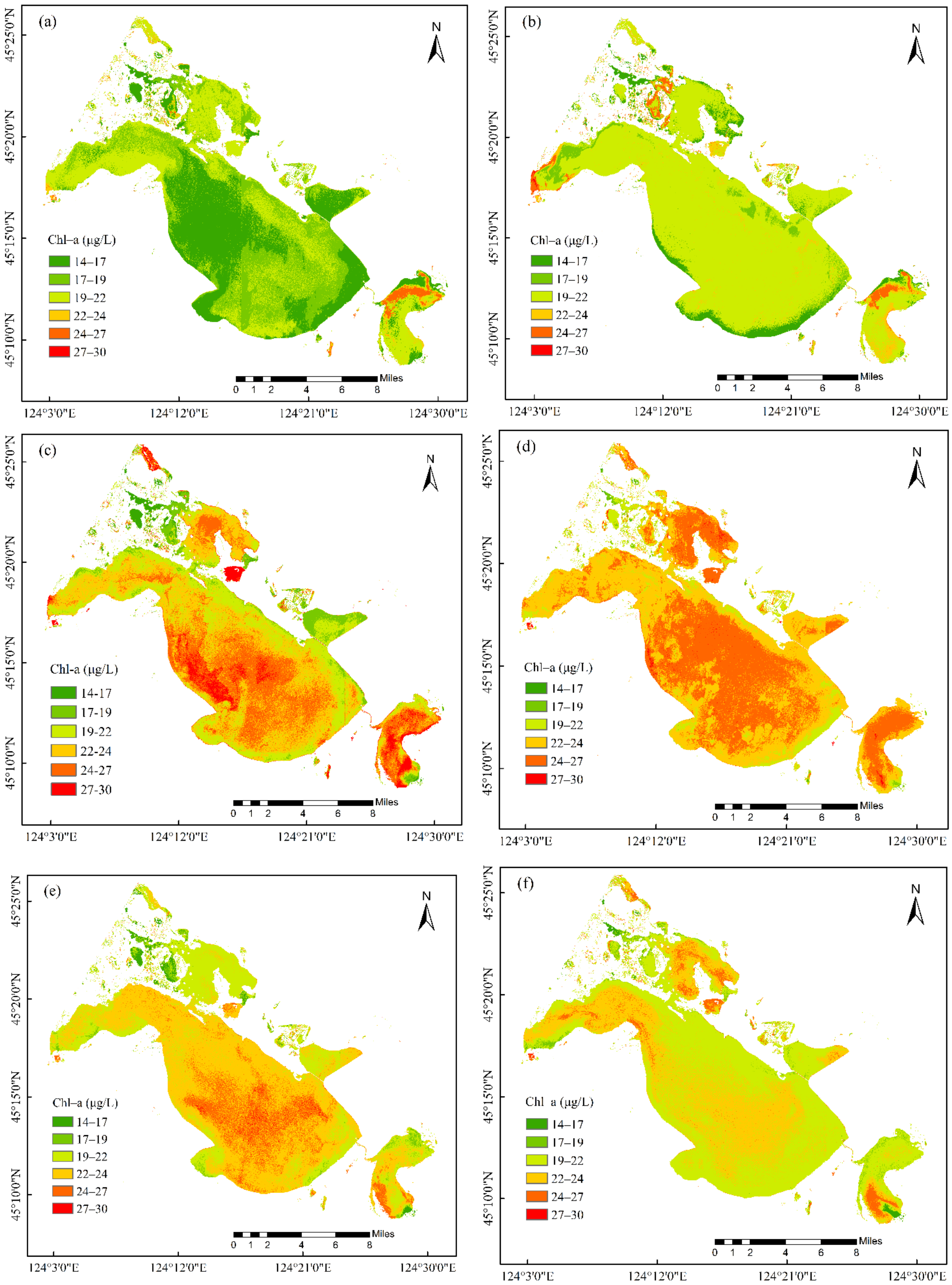


Figure 8. Spatial distribution of Chl-a concentrations retrieved by the RF model between 2020 and 2021 for the whole lake: (**a**) 2020 spring; (**b**) 2021 spring; (**c**) 2020 summer; (**d**) 2021 summer; (**e**) 2020 autumn; (**f**) 2021 autumn.

For the spatial distribution, the southwestern part of the lake was always the area with the greatest average Chl-a concentration over these three seasons on the whole. The concentration of Chl-a in the main lake area in the summer of 2021 was significantly higher than that in 2020, while the concentration of Chl-a in the summer of 2021 was slightly higher than that of 2020. Special temperature and climate conditions can cause this [47]. Therefore, the importance of large-scale and long-term application has been highlighted [48]. The Chl-a concentration in the autumn of 2020 was higher than that in 2021, but both had higher distributions in the south of the main lake area, showing a similar spatial distribution in different years and the same season. In situ sampling and measurements were concentrated in the main lake area, but from the perspective of the satellite images, when a more generalized model was developed, the distribution could also be observed.

5. Discussion

5.1. Significance of the Input Features in ML

SHAP values can explain ML models by visualizing the learning of each feature and sample in the model. The characteristics of the black box have always been a major problem in ML algorithms. It appears that these models perform well for selected metrics, but the learning process is difficult to observe, which may lead to misleading results. Relative feature importance can reveal the importance of the features to the predictions [15]. In this study, the results of relative feature importance for the two ML models (Figure 9) were similar to those of the SHAP bar plots (Figure 10), which showed SHAP values on the right-hand side of the bee swarm figure of the features' density scatter (Figure 4). It can be clearly seen that almost every feature made a contributions, with nonzero SHAP values except for NDCI in the XGBoost model (Figure 10b), which showed the same result as the evaluation of the relative feature importance (Figure 9b). This suggests that this input feature has no effect on the Chl-a concentration. In addition, each feature was learned in the RF model to a certain extent, while XGBoost focused mostly on the feature of Enhanced Three, resulting in uneven learning of the features.

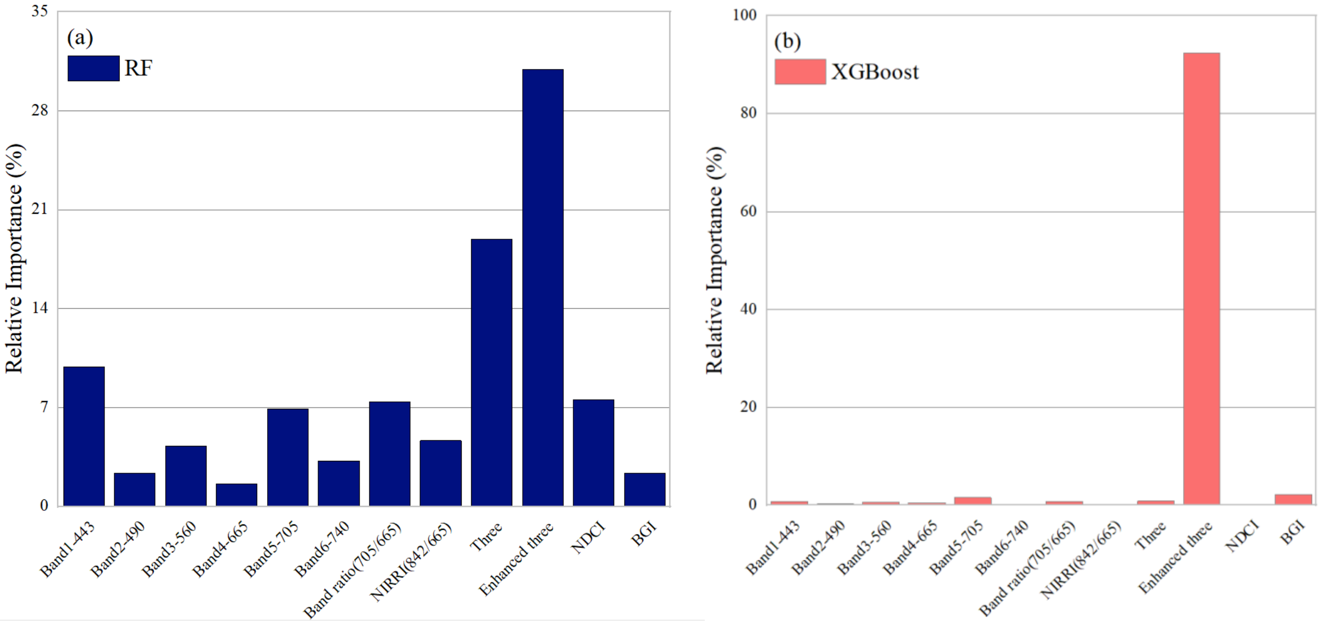


Figure 9. Relative feature importance of each input feature: (**a**) RF; (**b**) XGBoost model.

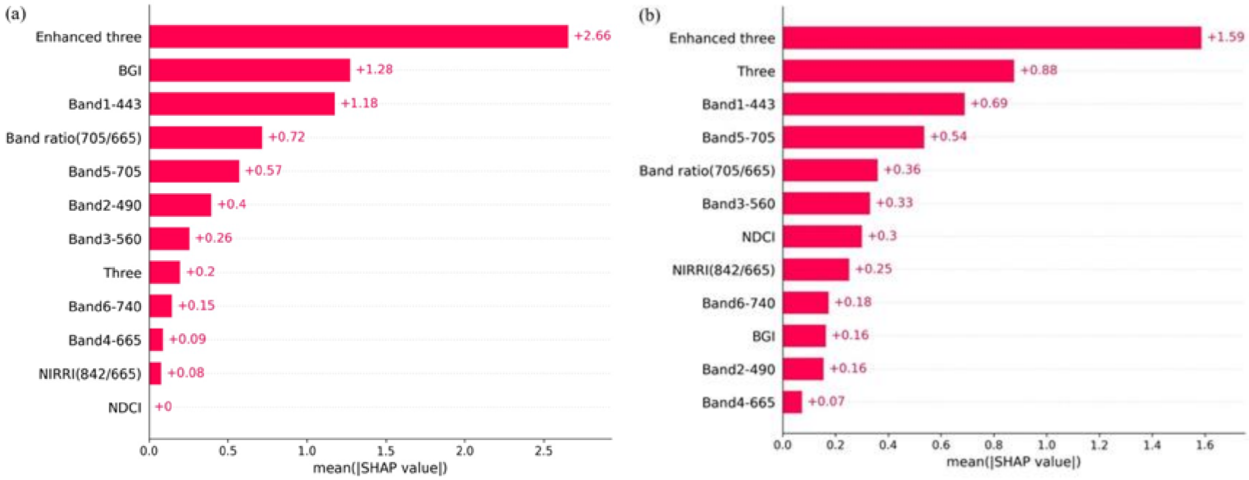


Figure 10. SHAP summary bar plot: (**a**) RF model; (**b**) XGBoost model.

Chl-a, TSS and CDOM jointly affect the optical signals of Case 2 waters. The components are complex and changeable [49]. In line with the bio-optical properties of Case 2 waters, Chl-a exhibits strong absorption at 443 nm and 665 nm. Moreover, the band at 705 nm and the other six empirical algorithms also had relatively biological mechanisms of Chl-a in lakes. Therefore, it is beneficial for assessing the XGBoost and RF models. As discussed in Section 4.1.2, the inaccurate learning of the XGBoost model was because the reflectance at 665 nm should have a negative correlation with the Chl-a concentration, because of the strong absorption of Chl-a. In addition, the feature of NDCI and Chl-a concentration should have a positive correlation [40]. Moreover, the atmosphere had an impact on the process of retrieval [10]. Strong absorption at around 665 nm may be interfered with by clouds, resulting in the XGBoost model not catching the weak Chl-a information hidden in the input features. This also appeared in the RF model, but most samples in the RF model were learned normally. Furthermore, the selection of input features was also considered. The best results (Table 5) were obtained when 12 variables were used as input. The performance during the process fluctuated slightly, possibly due to positive synergies or negative suppression between bands. After the feature of Band 4—665 nm was added as the last feature, a sudden improvement was made. In fact, the feature of Band 4—665 nm contains abundant Chl-a information, as shown by absolute values in the plots (Figures 9 and 10), thus not indicating that this feature contributes the least.

Table 5. Performance of the RF models with different input features. Numbers from 3 to 12 refer to the first n ranks in the SHAP bar plot.

Model	R^2	RMSE (μg L^{-1})	MAPE (%)
RF-3 features	0.71	2.91	11.41
RF-4 features	0.75	2.74	10.63
RF-5 features	0.75	2.71	10.81
RF-6 features	0.72	2.90	11.08
RF-7 features	0.73	2.83	10.87
RF-8 features	0.70	2.99	11.27
RF-9 features	0.74	2.76	10.67
RF-10 features	0.72	2.87	11.14
RF-11 features	0.73	2.80	10.64
RF-12 features	0.79	2.51	9.86

5.2. Comparison with Other Empirical Algorithms

For the retrieval of Chl-a concentrations, representative empirical models have achieved good results for inland lakes (Table 4). However, these models depended on data from certain research regions but may not have good performance in other areas, partly because of the various inherent optical properties (IOPs). The ML models performed well in terms of accuracy, robustness and computational demand. Compared with the NIRRI algorithm, which was developed for Lake Chagan [20], and Enhanced Three [23], which scored the most for the SHAP values (Figure 11), the RF model in this study performed better on the validation set. Though the bands from different satellites may have some effects, RF conformed to the bio-optical properties of Case 2 waters in terms of the SHAP values, which helped mitigate the impact of the black box characteristics of ML models. The results demonstrated that ML models can improve their seasonal performances. Actually, it would be possible to develop an independent model for each month [10]. However, since the numbers of in situ points in spring, summer and autumn were not balanced, it was more meaningful to develop a model with relatively stronger generalization. In addition, ML models take the relationship between the bands and the Chl-a values into account (Figure 10).

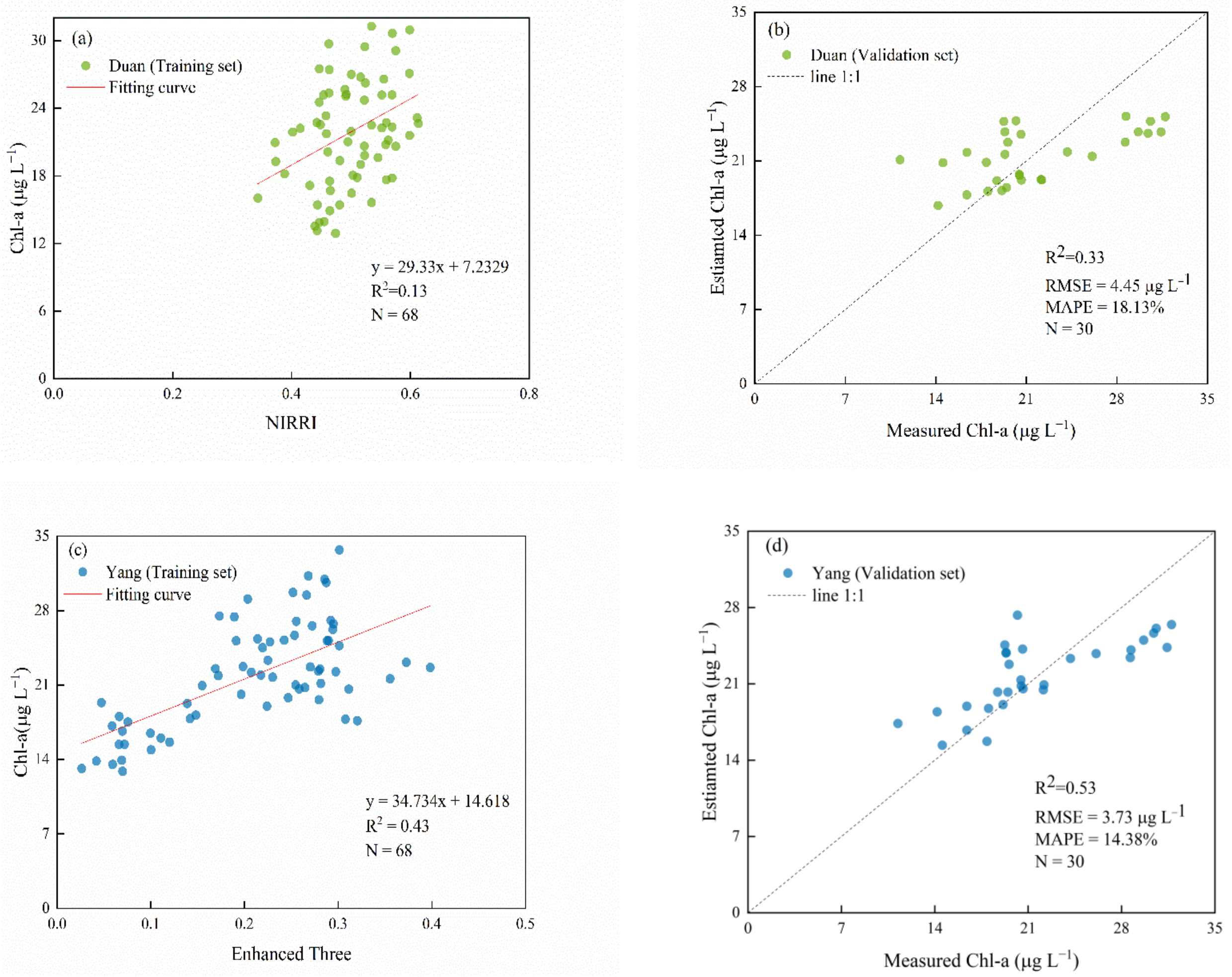


Figure 11. Algorithms' performances on (**a**) the training set of the NIRRI algorithm, (**b**) the validation set of the NIRRI algorithm, (**c**) the training set of the Enhanced Three algorithm and (**d**) the validation set of the Enhanced Three algorithm.

The RF model had the ability to capture the relationships among the variables and the nonlinear relationship between the input and the output, showing the marginal effect of a feature on the predictions of ML models [50]. Each feature contributed the strengths of the related empirical algorithms to the model, and may be beneficial by reducing impacts of NAP and CDOM on Chl-a retrieval [15]. In addition, the red edge bands of Sentinel-2 data were more effective for monitoring Chl-a in lakes. Due to the continuous renovation and development of the spectral bands of satellites, more information hidden in the pivotal bands may be gradually revealed, and the models with focused algorithms are being developed for lakes of various sizes [51].

5.3. Validation of In Situ and Sentinel-2 Reflectance

In fact, there are two alternative products of Sentinel-2. The Level-1C products result from using a digital elevation model to project the image onto a cartographic geometry. Per-pixel radiometric measurements are provided in the Top-Of-Atmosphere (TOA) reflectance data, along with the parameters used to transform them into radiance. Level-2A products derived from the associated Level-1C products have been systematically generated by the ground segment over Europe since March 2018, and their production was extended to the global scale in December 2018. L1C products can be processed by the Sen2cor [33] and CR2CC processors [24], a process which is somewhat complicated. TOA reflectance has been applied for Chl-a retrieval [10]. To verify the availability of directly using the Sentinel-2 BOA reflectance data published by the ESA, one group of in situ hyperspectral measurements for 21 August 2020 simultaneously with the time of overpass in the cloud-free Sentinel-2 images was selected for observation. Since the interval of the in situ hyperspectral data was 3.3 nm, the bands were selected according to relative bands of the Sentinel-2 data (Table 6).

Table 6. Sentinel-2 MSI and selected relative in situ hyperspectral bands.

Sentinel-2 MSI (nm)	In Situ Hyperspectral (nm)
443	442.1
490	491.6
560	560.9
665	666.5
705	702.8
740	739.1
783	782
842	841.4
865	864.5

The reflectance of 18 in situ points and the relative Sentinel-2 BOA reflectance was averaged and plotted (Figure 12). Clouds in the sky, wind speed and the time of observation between the in situ and satellite values were considered to have an influence. Though reflectance appeared to have a numerical influence, from another perspective, the polygonal trend implied that BOA reflectance had a similar changing tendency to the in situ hyperspectral data. According to the results of the three correlations, the two curves had a strong correlation (Table 7). In situ hyperspectral data and MSI data are fundamentally in agreement [29]. This reveals the feasibility of using Sentinel-2 L2A reflectance data for Chl-a retrieval.

Table 7. Three methods of r for the polygonal trend between the in situ hyperspectral data and the ESA L2A products.

Method	r
Pearson	0.99
Spearman	0.98
Kendall	0.94

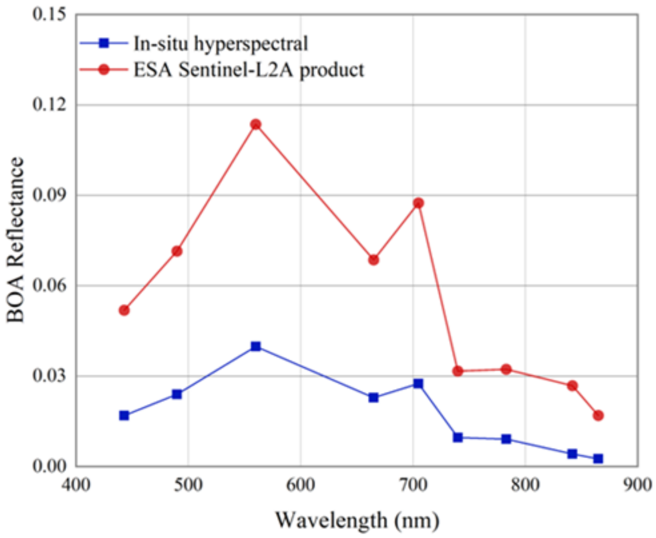


Figure 12. Polygonal trend between the in situ hyperspectral data and the ESA L2A products.

5.4. Ecological Variations across Lakes between 2020 and 2021

The type of land cover can influence variations in Chl-a [33]. As discussed in Section 4.2.3, the distribution of high Chl-a concentrations in the main lake area in three seasons was mainly located in the southeast part of the lake. In recent years, due to the need to ensure food security, areas around Lake Chagan have made great efforts to build saline alkali land irrigation areas. The water supply of Lake Chagan is relatively stable and sufficient, mainly because of farmland recession and loose water diversion. However, while ensuring the water supply of the lake, the backwater in the irrigation area has also caused an excessive load of nutrients in the lake, and a large amount of farmland backwater that is rich in a high concentration of nutrients, salinity and TSS was discharged, along with the continuous development of tourism, threatening the environmental safety of Lake Chagan’s water. The high nutrient concentration has led to eutrophication of the water body, which is conducive to the growth of algae [52].

From May to September between 2020 and 2021, the average, maximum and minimum Chl-a concentrations across the whole lake were retrieved by the RF model (Figure 13a). None of the three fluctuated much, which indicates that there was no obvious influence of anthropogenic activities as a whole between 2020 and 2021. On the other hand, Lake Chagan was greatly affected by wind speed and shore collapse, which produced suspended mineral and sediment particles [53]. A higher concentration of TSS will limit light transmission and thus restrict the growth of phytoplankton [54]. The TSS concentration of the in situ measurement points may reveal this phenomenon (Figure 13b). In the five groups of plots between August 2020 and July 2021, there was a clear increase in the concentration of TSS. This study undoubtedly provides a timely signal to local ecological management authorities. However, the measured Chl-a concentration did not increase significantly. This may indicate that TSS does not directly affect Chl-a unilaterally, and Chl-a concentration is related to the location of the sampling points and the orientation of the water connecting with the outside environment.

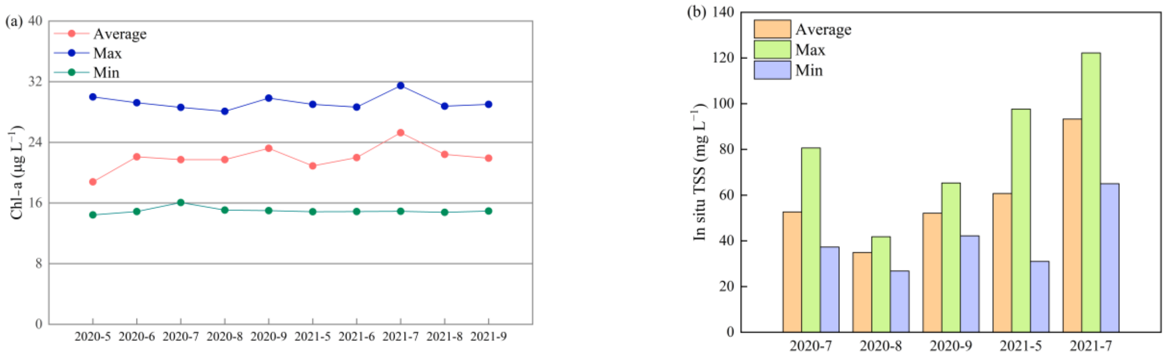


Figure 13. Time series of (**a**) averaged Chl-a concentrations retrieved by the RF model and (**b**) TSS concentrations of the in situ measurement points.

6. Conclusions

To estimate the Chl-a concentrations in Lake Chagan, a reliable RF model was developed based on in situ measurements and near-synchronous Sentinel-2 images. The model was assessed according to the bio-optical properties of Case 2 waters for Chl-a, and the disadvantages of XGBoost in this study were revealed. Compared with the empirical NIRRI algorithm (R^2 = 0.33, RMSE = 4.45 μg L^{-1}, MAPE = 18.13%) and the enhanced three-band index (R^2 = 0.53, RMSE = 3.73 μg L^{-1}, MAPE = 14.38%), the developed RF model showed good performance (R^2 = 0.79, RMSE = 2.51 μg L^{-1}, MAPE = 9.86%) on the validation set and conformed to the biological mechanisms of Chl-a in lakes. The model also proved to be relatively robust to the seasonal changes that were observed by KNN. The peak Chl-a concentration in the summer of 2021 was higher than that in 2020, but in the spring and autumn of 2020, the Chl-a concentration was higher than that in 2021. The Chl-a concentration fluctuated during the three seasons between 2020 and 2021, but remained stable as a whole. Our results demonstrate that it is necessary to assess ML models for retrieving Chl-a. The study also illustrates the feasibility of retrieving Chl-a concentrations at large scales via the reflectance data of on-board satellite sensors. In future studies, models with better precision will be developed based on more in situ measurements from different types of Case 2 waters.

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

Funding: This study was supported by the Land Observation Satellite Supporting Platform of the National Civil Space Infrastructure Project (CASPLOS-CCSI). The research is funded by the National Natural Science Foundation of China (Grant No. 41871225) and the Project of Jilin Province Development and Reform Commission (2021C044-7).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. Jian, J.; Bailey, V.; Dorheim, K.; Konings, A.G.; Hao, D.; Shiklomanov, A.N.; Snyder, A.; Steele, M.; Teramoto, M.; Vargas, R.; et al. Historically inconsistent productivity and respiration fluxes in the global terrestrial carbon cycle. *Nat. Commun.* **2022**, *13*, 1–9. [CrossRef]
2. Qiu, R.; Li, X.; Han, G.; Xiao, J.; Ma, X.; Gong, W. Monitoring drought impacts on crop productivity of the US Midwest with solar-induced fluorescence: GOSIF outperforms GOME-2 SIF and MODIS NDVI, EVI, and NIRv. *Agric. For. Meteorol.* **2022**, *323*, 109038. [CrossRef]

3. Li, Y.; Ma, Q.; Chen, J.M.; Croft, H.; Luo, X.; Zheng, T.; Rogers, C.; Liu, J. Fine-scale leaf chlorophyll distribution across a deciduous forest through two-step model inversion from Sentinel-2 data. *Remote Sens. Environ.* **2021**, *264*, 112618. [CrossRef]
4. Brooks, B.W.; Lazorchak, J.M.; Howard, M.D.; Johnson, M.V.; Morton, S.L.; Perkins, D.A.; Reavie, E.D.; Scott, G.I.; Smith, S.A.; Steevens, J.A. Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? *Environ. Toxicol. Chem.* **2016**, *35*, 6–13. [CrossRef]
5. Mansaray, A.S.; Dzialowski, A.R.; Martin, M.E.; Wagner, K.L.; Stoodley, S.H. Comparing PlanetScope to Landsat-8 and Sentinel-2 for Sensing Water Quality in Reservoirs in Agricultural Watersheds. *Remote Sens.* **2021**, *13*, 1847. [CrossRef]
6. OReilly, J.E.; Werdell, P.J. Chlorophyll algorithms for ocean color sensors—OC4, OC5 & OC6. *Remote Sens. Environ.* **2019**, *229*, 32–47.
7. Antoine, D.; André, J.M.; Morel, A. Oceanic primary production: 2. Estimation at global scale from satellite (coastal zone color scanner) chlorophyll. *Global Bio-Geochem. Cycl.* **1996**, *10*, 57–69. [CrossRef]
8. OReilly, J.E.; Maritorena, S.; Mitchell, B.G.; Siegel, D.A.; Carder, K.L.; Garver, S.A.; Kahru, M.; McClain, C. Ocean color chlorophyll algorithms for SeaWiFS. *J. Geophys. Res. Ocean* **1998**, *103*, 24937. [CrossRef]
9. Chen, Z.; Hu, C.; Conmy, R.N.; Muller-Karger, F.; Swarzenski, P. Colored dissolved organic matter in Tampa Bay, Florida. *Mar. Chem.* **2007**, *104*, 98–109. [CrossRef]
10. Hang, X.; Li, Y.; Li, X.; Meng, X.; Sun, L. Estimation of Chlorophyll-a Concentration in Lake Taihu from Gaofen-1 Wide-Field-of-View Data through a Machine Learning Trained Algorithm. *J. Meteorol. Res.* **2022**, *36*, 208–226. [CrossRef]
11. Pahlevan, N.; Smith, B.; Schalles, J.; Binding, C.E.; Cao, Z.; Ma, R.; Alikas, K.; Kangro, K.; Gurlin, D.; Hà, N.; et al. Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. *Remote Sens. Environ.* **2020**, *240*, 111604. [CrossRef]
12. Gregg, W.W.; Casey, N.W. Sampling biases in MODIS and SeaWiFS ocean chlorophyll data. *Remote Sens. Environ.* **2007**, *111*, 25–35. [CrossRef]
13. Duan, H.; Zhang, Y.; Zhang, B.; Song, K.; Wang, Z. Assessment of chlorophyll-a concentration and trophic state for Lake Chagan using Landsat TM and field spectral data. *Environ. Monit. Assess* **2007**, *129*, 295–308. [CrossRef] [PubMed]
14. Bonansea, M.; Rodriguez, M.C.; Pinotti, L.; Ferrero, S. Using multi-temporal Landsat imagery and linear mixed models for assessing water quality parameters in Río Tercero reservoir (Argentina). *Remote Sens. Environ.* **2015**, *158*, 28–41. [CrossRef]
15. Cao, Z.; Ma, R.; Duan, H.; Pahlevan, N.; Melack, J.; Shen, M.; Xue, K. A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes. *Remote Sens. Environ.* **2020**, *248*, 111974. [CrossRef]
16. Toming, K.; Kutser, T.; Laas, A.; Sepp, M.; Paavel, B.; Nõges, T. First experiences in mapping lake water quality parameters with Sentinel-2 MSI imagery. *Remote Sens.* **2016**, *8*, 640. [CrossRef]
17. Buma, W.G.; Lee, S.I. Evaluation of sentinel-2 and landsat 8 images for estimating chlorophyll-a concentrations in lake Chad, Africa. *Remote Sens.* **2020**, *12*, 2437. [CrossRef]
18. Li, J.; Ma, R.; Xue, K.; Zhang, Y.; Loiselle, S. A remote sensing algorithm of column-integrated algal biomass covering algal bloom conditions in a shallow Eutrophic Lake. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 466. [CrossRef]
19. Lyu, H.; Li, X.; Wang, Y.; Jin, Q.; Cao, K.; Wang, Q.; Li, Y. Evaluation of chlorophyll-a retrieval algorithms based on MERIS bands for optically varying eutrophic inland lakes. *Sci. Total Environ.* **2015**, *530*, 373–382. [CrossRef] [PubMed]
20. Duan, H.; Ma, R.; Hu, C. Evaluation of remote sensing algorithms for cyanobacterial pigment retrievals during spring bloom formation in several lakes of East China. *Remote Sens. Environ.* **2012**, *126*, 126–135. [CrossRef]
21. Gitelson, A.A.; Gao, B.C.; Li, R.R.; Berdnikov, S.; Saprygin, V. Estimation of chlorophyll-a concentration in productive turbid waters using a Hyperspectral Imager for the Coastal Ocean—The Azov Sea case study. *Environ. Res. Lett.* **2011**, *6*, 024023. [CrossRef]
22. Le, C.; Hu, C.; Cannizzaro, J.; English, D.; Muller-Karger, F.; Lee, Z. Evaluation of chlorophyll-a remote sensing algorithms for an optically complex estuary. *Remote Sens. Environ.* **2013**, *129*, 75–89. [CrossRef]
23. Yang, W.; Matsushita, B.; Chen, J.; Fukushima, T.; Ma, R. An enhanced three-band index for estimating chlorophyll-a in turbid case-II waters: Case studies of Lake Kasumigaura, Japan, and Lake Dianchi, China. *IEEE Trans. Geosci. Remote Sens. Lett.* **2010**, *7*, 655–659. [CrossRef]
24. Li, S.; Song, K.; Wang, S.; Liu, G.; Wen, Z.; Shang, Y.; Lyu, L.; Chen, F.; Xu, S.; Tao, H.; et al. Quantification of chlorophyll-a in typical lakes across China using Sentinel-2 MSI imagery with machine learning algorithm. *Sci. Total Environ.* **2021**, *778*, 146271. [CrossRef]
25. Kwon, Y.S.; Baek, S.H.; Lim, Y.K.; Pyo, J.; Ligaray, M.; Park, Y.; Cho, K.H. Monitoring coastal chlorophyll-a concentrations in coastal areas using machine learning models. *Water* **2018**, *10*, 1020. [CrossRef]
26. Topp, S.N.; Pavelsky, T.M.; Jensen, D.; Simard, M.; Ross, M.R. Research trends in the use of remote sensing for inland water quality science: Moving towards multidisciplinary applications. *Water* **2020**, *12*, 169. [CrossRef]
27. Hafeez, S.; Wong, M.S.; Ho, H.C.; Nazeer, M.; Nichol, J.; Abbas, S.; Tang, D.; Lee, K.H.; Pun, L. Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: A case study of Hong Kong. *Remote Sens.* **2019**, *11*, 617. [CrossRef]
28. Yang, H.; Du, Y.; Zhao, H.; Chen, F. Water Quality Chl-a Inversion Based on Spatio-Temporal Fusion and Convolutional Neural Network. *Remote Sens.* **2022**, *14*, 1267. [CrossRef]

29. Liu, T.; Gu, Y.; Jia, X. Class-guided coupled dictionary learning for multispectral-hyperspectral remote sensing image collaborative classification. *Sci. China Technol. Sci.* **2022**, *65*, 744–758. [CrossRef]
30. Cui, Y.; Meng, F.; Fu, P.; Yang, X.; Zhang, Y.; Liu, P. Application of hyperspectral analysis of chlorophyll a concentration inversion in Nansi Lake. *Ecol. Inform.* **2021**, *64*, 101360. [CrossRef]
31. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Proc. Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–10.
32. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef]
33. Kim, Y.W.; Kim, T.; Shin, J.; Lee, D.S.; Park, Y.S.; Kim, Y.; Cha, Y. Validity evaluation of a machine-learning model for chlorophyll a retrieval using Sentinel-2 from inland and coastal waters. *Ecol. Indicat.* **2022**, *137*, 108737. [CrossRef]
34. Zhu, L.; Yan, B.; Wang, L.; Pan, X. Mercury concentration in the muscle of seven fish species from Chagan Lake, Northeast China. *Environ. Monit. Assess* **2012**, *184*, 1299–1310. [CrossRef]
35. Mobley, C.D. Estimation of the remote-sensing reflectance from above-surface measurements. *Appl. Opt.* **1999**, *38*, 7442–7455. [CrossRef] [PubMed]
36. Gascon, F.; Bouzinac, C.; Thépaut, O.; Jung, M.; Francesconi, B.; Louis, J.; Lonjou, V.; Lafrance, B.; Massera, S.; Gaudel-Vacaresse, A.; et al. Copernicus Sentinel-2A calibration and products validation status. *Remote Sens.* **2017**, *9*, 584. [CrossRef]
37. Rodríguez-Pérez, R.; Bajorath, J. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *J. Med. Chem.* **2019**, *63*, 8761–8777. [CrossRef]
38. Park, J.; Lee, W.H.; Kim, K.T.; Park, C.Y.; Lee, S.; Heo, T.Y. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. *Sci. Total Environ.* **2022**, *832*, 155070. [CrossRef]
39. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
40. Mishra, S.; Mishra, D.R. Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. *Remote Sens. Environ.* **2012**, *117*, 394–406. [CrossRef]
41. Ha, N.T.; Koike, K.; Nhuan, M.T.; Canh, B.D.; Thao, N.T.; Parsons, M. Landsat 8/OLI two bands ratio algorithm for chlorophyll-a concentration mapping in hypertrophic waters: An application to West Lake in Hanoi (Vietnam). *IEEE J. Select. Top. Appl. Earth Observat. Remote Sens.* **2017**, *10*, 4919–4929. [CrossRef]
42. Cover, T.; Hart, P. Nearest neighbour pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
43. Hwang, W.J.; Wen, K.W. Fast kNN classification algorithm based on partial distance search. *Electron. Lett.* **1998**, *34*, 2062–2063. [CrossRef]
44. Lee, Z.P.; Carder, K.L.; Peacock, T.G.; Davis, C.O.; Mueller, J.L. Method to derive ocean absorption coefficients from remote-sensing reflectance. *Appl. Opt.* **1996**, *35*, 453–462. [CrossRef]
45. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 1–13. [CrossRef]
46. Zhang, S. Cost-sensitive KNN classification. *Neurocomputing* **2020**, *391*, 234–242. [CrossRef]
47. Shi, K.; Zhang, Y.; Xu, H.; Zhu, G.; Qin, B.; Huang, C.; Liu, X.; Zhou, Y.; Lv, H. Long-term satellite observations of microcystin concentrations in Lake Taihu during cyanobacterial bloom periods. *Environ. Sci. Technol.* **2015**, *49*, 6448–6456. [CrossRef] [PubMed]
48. Ciancia, E.; Magalhães Loureiro, C.; Mendonça, A.; Coviello, I.; Di Polito, C.; Lacava, T.; Pergola, N.; Satriano, V.; Tramutoli, V.; Martins, A. On the potential of an RST-based analysis of the MODIS-derived chl-a product over Condor seamount and surrounding areas (Azores, NE Atlantic). *Ocean Dyn.* **2016**, *66*, 1165–1180. [CrossRef]
49. Odermatt, D.; Gitelson, A.; Brando, V.E.; Schaepman, M. Review of constituent retrieval in optically deep and complex waters from satellite imagery. *Remote Sens. Environ.* **2012**, *118*, 116–126. [CrossRef]
50. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Statist.* **2001**, *29*, 1189–1232. [CrossRef]
51. Downing, J.A.; Prairie, Y.T.; Cole, J.J.; Duarte, C.M.; Tranvik, L.J.; Striegl, R.G.; McDowell, W.H.; Kortelainen, P.; Caraco, N.F.; Melack, J.M.; et al. The global abundance and size distribution of lakes, ponds, and impoundments. *Limnol. Oceanogr.* **2006**, *51*, 2388–2397. [CrossRef]
52. Nguyen, H.Q.; Ha, N.T.; Nguyen-Ngoc, L.; Pham, T.L. Comparing the performance of machine learning algorithms for remote and in situ estimations of chlorophyll-a content: A case study in the Tri An Reservoir, Vietnam. *Water Environ. Res.* **2021**, *93*, 2941–2957. [CrossRef]
53. Li, S.; Song, K.; Zhao, Y.; Mu, G.; Shao, T.; Ma, J. Absorption characteristics of particulates and CDOM in waters of Chagan Lake and Xinlicheng Reservoir in autumn. *Huan Jing Ke Xue* **2016**, *37*, 112–122. [PubMed]
54. Feng, L.; Hu, C.; Chen, X.; Zhao, X. Dramatic inundation changes of China's two largest freshwater lakes linked to the Three Gorges Dam. *Environ. Sci. Technol.* **2013**, *47*, 9628–9634. [CrossRef] [PubMed]