applied sciences

MDPI

Article

Retrieval of Chlorophyll-a Concentrations in the Coastal Waters of the Beibu Gulf in Guangxi Using a Gradient-Boosting Decision Tree Model

Huanmei Yao [1,*], Yi Huang [1], Yiming Wei [1], Weiping Zhong [2] and Ke Wen [1]

[1] School of Resources, Environment and Materials, Guangxi University, Nanning 530004, China; 1915303004@st.gxu.edu.cn (Y.H.); yimingwei621@gmail.com (Y.W.); wenke1996@163.com (K.W.)
[2] Automatic Monitoring Office, Marine Environmental Monitoring Center of Guangxi Zhuang Autonomous Region, Beihai 536000, China; oceannicole@163.com
* Correspondence: yaohuanmei@gxu.edu.cn

Abstract: Remote sensing for the monitoring of chlorophyll-a (Chl-a) is essential to compensate for the shortcomings of traditional water quality monitoring, strengthen red tide disaster monitoring and early warnings, and reduce marine environmental risks. In this study, a machine learning approach called the Gradient-Boosting Decision Tree (GBDT) was employed to develop an algorithm for estimating the Chl-a concentrations of the coastal waters of the Beibu Gulf in Guangxi, using Landsat 8 OLI image data as the image source in combination with field measurements of Chl-a concentrations. The GBDT model with B4, B3 + B4, B3, B1 − B4, B2 + B4, B1 + B4, and B2 − B4 as input features exhibited higher accuracy (MAE = 0.998 μg/L, MAPE = 19.413%, and RMSE = 1.626 μg/L) compared with different physics models, providing a new method for remote sensing inversion of water quality parameters. The GBDT model was used to study the spatial distribution and temporal variation of Chl-a concentrations in the coastal sea surface of the Beibu Gulf of Guangxi from 2013 to 2020. The results showed a spatial distribution with high concentrations in nearshore waters and low concentrations in offshore waters. The Chl-a concentration exhibited seasonal changes (concentration in summer > autumn > spring ≈ winter).

Keywords: chlorophyll-a; gbdt model; Guangxi Beibu Gulf; remote sensing inversion

Citation: Yao, H.; Huang, Y.; Wei, Y.; Zhong, W.; Wen, K. Retrieval of Chlorophyll-a Concentrations in the Coastal Waters of the Beibu Gulf in Guangxi Using a Gradient-Boosting Decision Tree Model. *Appl. Sci.* **2021**, *11*, 7855. https://doi.org/10.3390/app11177855

Academic Editor: Lorena Parra

Received: 20 July 2021
Accepted: 23 August 2021
Published: 26 August 2021

1. Introduction

Chlorophyll-a (Chl-a) is an important index that can reflect phytoplankton biomass and the state of eutrophication. The concentration of Chl-a increases as the phytoplankton biomass increases, and an increase in phytoplankton may cause red tides, with the potential to threaten public health [1] and wildlife [2] and being harmful to the environment [3,4]. The Beibu Gulf in Guangxi receives more than 120 small- and medium-sized inflows carrying a large amount of organic matter and inorganic salts, and it is likely to experience eutrophication and red tides. From 2014 to 2017, several abnormal water quality events occurred in the Beibu Gulf of Guangxi, including two red tides. Consequently, water quality monitoring in the Guangxi Beibu Gulf is of significance for protecting the water quality and environment of the Beibu Gulf and ensuring the health of its residents [5].

Traditional monitoring methods cannot completely reflect the spatial and temporal distribution of the water quality because of the small coverage of such methods, which are limited by high costs. However, satellite remote sensing enables automated monitoring of water quality parameters, including the Chl-a and total suspended solids. Such techniques benefit from lower costs and greater spatial and temporal coverage [6], and such studies have been undertaken overseas and in China [7–10].

In recent years, machine learning has been used to retrieve water quality. Machine learning enhances the accuracy of inversion and the generalization ability through model

training and, in turn, predicting and analyzing test data [11]. This method uses internal implicit networks and structures to determine the complex characteristics of input data and obtain explicit relationships among the output variables [12,13]. Several approaches, including the decision tree [12,14], BP neural network [15,16], support vector machine model (SVM) [17,18], and extreme machine learning approaches [19], have strong adaptability, fault tolerance, and organization of data.

The optical radiometric measurement of the Chl-a concentration in coastal waters remains a challenge due to the presence of phytoplankton, suspended matter, and colored, dissolved organic matter. Some studies have revealed the advantages associated with the machine learning method applied in this field. In Galician Rias (northwest Spain), the neural network techniques were applied to estimate the Chl-a concentrations in three different water types. The results showed the capacity of the neural network to predict the Chl-a concentrations in coastal waters [20]. The Mixture Density Network (MDN) was induced for seamless retrieval of Chl-a data records in inland and coastal waters in the study of Pahlevan et al. As evidenced through image and satellite matchup analyses, the model generated realistic spatial distributions and provided a more accurate Chl-a map [21]. The Gradient-Boosting Decision Tree (GBDT) is a machine learning technique for regression, classification, and other tasks, using a decision tree flowchart approach combined with the boosting ensemble technique. The GBDT improves the capacity of the decision tree by reducing the residuals generated during the training procedure [22,23]. It has been widely applied in social science research [24–28] and gradually introduced into the field of natural science [1–7,29–35]. The GBDT exhibits much better performance in the retrieval of water depth compared with the single-band, dual-band, and BP neural network models [36]. Some studies have shown that the GBDT model can achieve higher simulation accuracy to the random forest (RF) algorithm and regression tree input with the same meteorological factors [37]. By constantly calculating the best fitting value and updating the classifier, the GBDT algorithm can obtain explicit relationships and features between different types of data with little prior knowledge. Considering the complex relationship between the water quality parameters and spectral characteristics, the GBDT model has the potential for faster and more accurate application of remote sensing retrieval of Chl-a concentrations in coastal waters.

In light of the above considerations, the main aim of this study was to (1) develop a machine learning algorithm for Chl-a estimation in the coastal region of the Beibu Gulf in Guangxi, exploring the potential of GBDT model in the retrieval of water quality parameters in coastal waters, (2) compare the performance of the GBDT with that of conventional models, and (3) analyze the factors determining the temporal–spatial distribution of the Chl-a concentration in the Beibu Gulf in Guangxi.

2. Materials and Methods

2.1. Study Area

The Beibu Gulf (107°57′ E~109°48′ E, 21°00′ N~22°15′ N, Figure 1) is a coastal region in Guangxi Province, included in the administrative region of Qinzhou, Beihai, and Fangchenggang. The study area can be roughly divided into the Qinzhou Bay, Fangcheng Bay, Dafeng Estuary Bay, Nanliu Estuary Bay, Tieshan Port Bay, and Pearl Bay.

Located south of the Tropic of Cancer, the Beibu Gulf of Guangxi is dominated by a subtropical climate with an oceanic monsoon, exhibiting transitional characteristics from subtropical to tropical [38]. The average sea surface temperature in the Beibu Gulf of Guangxi is approximately 22.6 °C, with high temperatures experienced from June to August and low temperatures occurring from December to March. Rainfall is cyclical, with the wet season ranging from May to October and the dry season ranging from November to the following April. The annual average rainfall is approximately 1500 mm, and the rivers in Guangxi are mainly replenished by rainwater, particularly during the wet season and a small amount in the dry season.

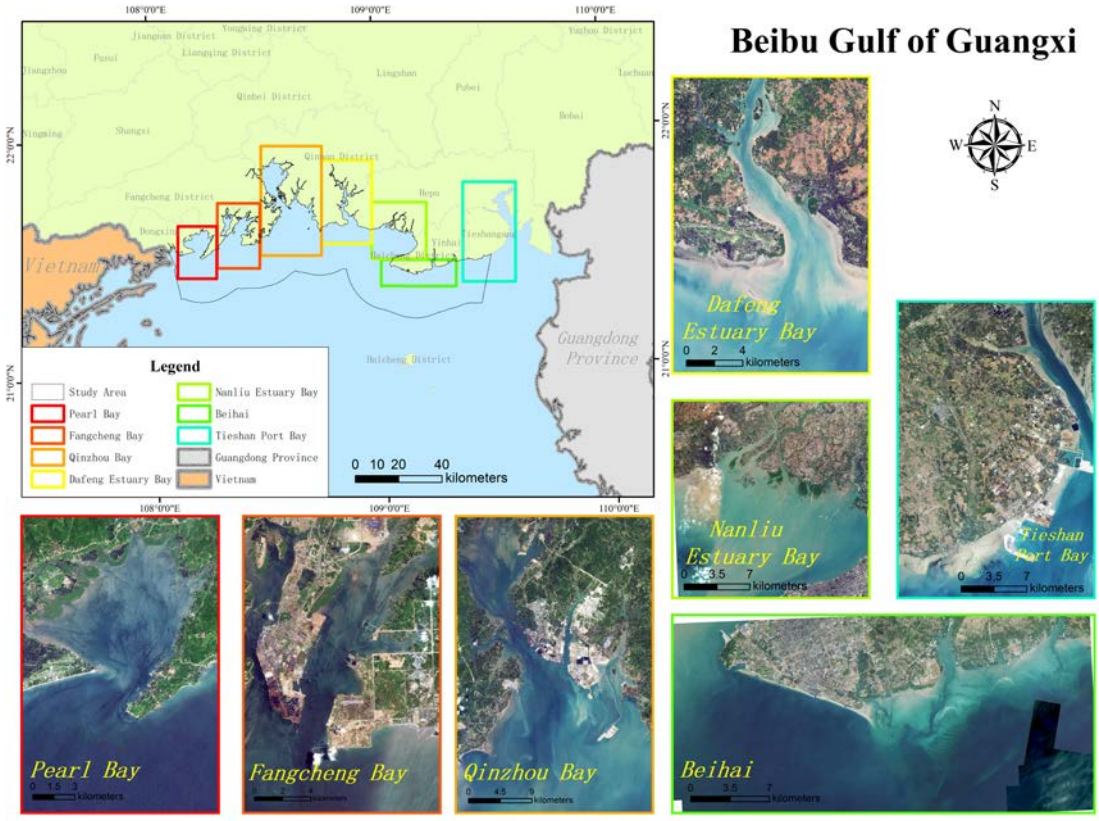

Figure 1. Study region. The colored rectangles represent the mouths of the bays in the Beibu Gulf.

2.2. Dataset

2.2.1. In Situ Data

Chl-a concentrations were measured every 30 min at the automatic monitoring station of the Marine Environmental Monitoring Center of the Guangxi Zhuang Autonomous Region [39] (Figure 2). The quality of the coastal water at a depth of 0.5 m was monitored every half an hour by a multi-parameter probe (6600V2-4, YSI, Yellow Springs, OH, USA) produced by Xylem, and the concentration of Chl-a was determined by a in vivo fluorescence method.

The instantaneous value of the Chl-a concentration closest to the transit time of Landsat 8 (11:11 a.m.) was selected as the model input dataset.

2.2.2. Satellite Data Acquisition and Pre-Processing

The Landsat 8 OLI satellite data with a nominal 30-m spatial resolution used in this study were accessed from the United States Geological Survey (USGS) portal (https://glovis.usgs.gov (accessed on 1 January 2021)). The images (path-row: 125-045) covered the coastal waters of the Beibu Gulf of Guangxi to the maximum extent, and the 13 automatic monitoring stations established by the Marine Environmental Monitoring Center of Guangxi are in the coverage. 34 scene images with low cloud cover were selected, ranging from 2013 to 2020 (Table 1).

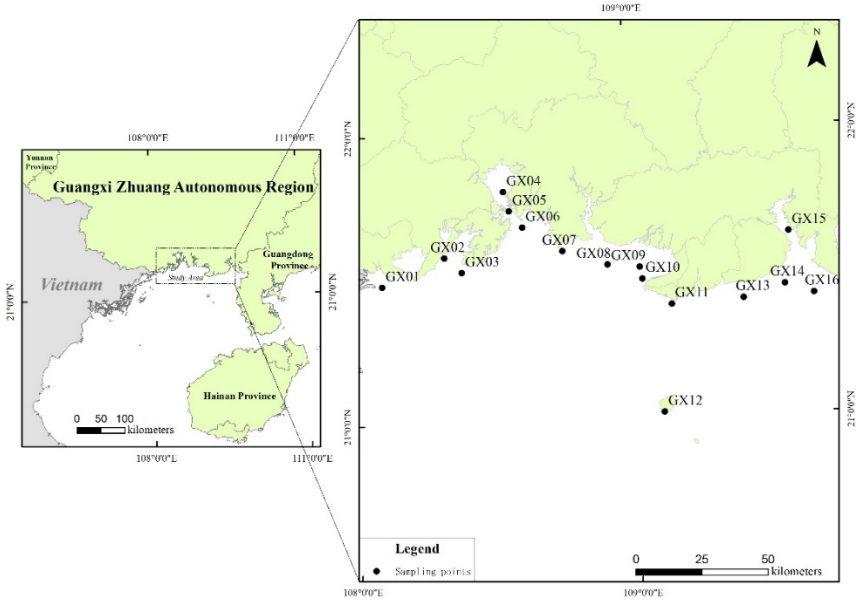

Figure 2. Distribution of sampling points.

Table 1. Dates of remote sensing images.

Date	Cloud Cover	Date	Cloud Cover
7 December 2020	22.26	28 October 2017	0.13
5 November 2020	12.66	2 March 2017	0.09
2 September 2020	18.40	14 February 2017	0.28
27 April 2020	14.75	28 December 2016	1.07
23 February 2020	9.22	9 October 2016	3.60
5 December 2019	9.63	3 June 2016	15.18
18 October 2019	15.99	23 October 2015	0.64
2 October 2019	5.56	7 October 2015	11.35
15 August 2019	15.80	1 June 2015	21.82
11 May 2019	34.00	14 April 2015	1.01
20 February 2019	33.79	1 August 2014	16.62
18 December 2018	26.23	14 June 2014	14.28
31 October 2018	0.03	21 January 2014	0.05
29 September 2018	4.98	5 January 2014	1.24
11 July 2018	19.78	20 December 2013	0.42
9 June 2018	1.89	4 December 2013	0.03
1 February 2018	1.89	2 November 2013	5.83

The Landsat 8 OLI satellite images were radiologically calibrated and atmospherically corrected before further processing. By radiometric calibration, the digital number (DN) recorded by the sensor could be converted into the spectral radiance and to the Top of the Atmosphere (TOA) reflectance. The surface reflectance may have changed after the atmospheric transmission, so the atmospheric correction was required. The error of reflectance reduced after atmospheric correction and could be used for the retrieval of Chl-a concentration.

In this study, all of the images were processed for radiometric calibration and atmospheric correction using the FLAASH model in the ENVI 5.3.1 software package. The corrected image reduced the influence of water vapor particles in the air and was clearer than the image before correction. The spectral curve of the pixel after atmospheric correction was closer to the actual spectral curve of the ground object and more in line with the requirements of inversion as shown in Figure 3.

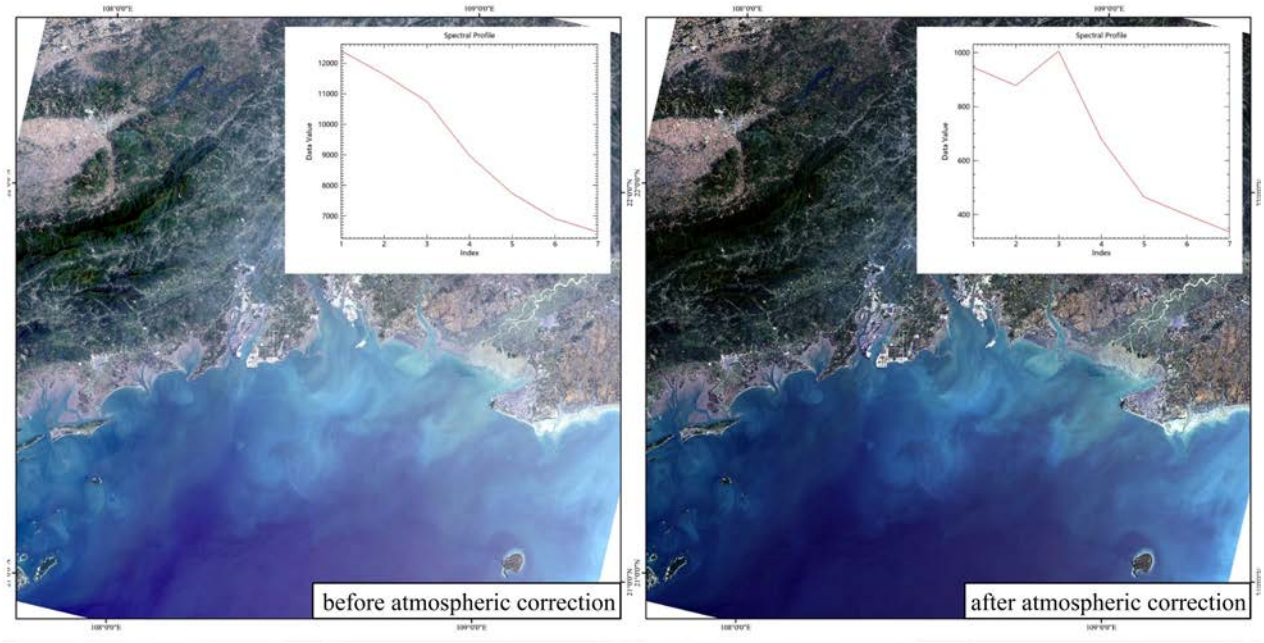

Figure 3. Comparison before and after atmospheric correction.

2.2.3. Calibration Dataset

With the pre-processing of remote sensing images accomplished by radiometric calibration and atmospheric correction of the image, the measured data of the Chl-a concentration were matched with the spectral data of the monitoring sampling points, producing 117 samples in total (Table 2).

Table 2. Chl-a concentration and reflectance of some samples.

Dates	Site Number	Concentration (μg/L)	Reflectance						
			B1	B2	B3	B4	B5	B6	B7
14 April 2015	GX05	5.30	0.077	0.070	0.084	0.059	0.038	0.029	0.024
23 October 2015	GX04	2.80	0.085	0.075	0.083	0.054	0.030	0.012	0.006
28 December 2016	GX02	3.20	0.070	0.061	0.074	0.048	0.017	0.005	0.003
14 February 2017	GX02	2.00	0.078	0.068	0.070	0.040	0.023	0.008	0.005
11 July 2018	GX13	8.80	0.119	0.116	0.127	0.108	0.108	0.108	0.088

Clouds and shadows on the fog surface may have caused data anomalies, resulting in deviations between the inversion results and field data. To improve the accuracy of the inversion model, outliers needed to be removed.

A boxplot was used to filter abnormal data. The reflectance of each band for the 117 samples was calculated (Figure 4), 7 abnormal data points in the study samples were identified and removed, and 110 samples were used in the study thereafter.

2.3. GBDT Model

The Gradient-Boosting Decision Tree (GBDT) algorithm was proposed by Friedman in 1999 [40]. The algorithm restricts weak learners from using only the Classification and Regression Tree (CART) model, a widely used model for constructing decision trees for both classification and regression problems. When building a regression tree using the CART model, the feature selection index generally uses the node minimum sample variance. The larger the sample variance, the greater the node data scatter with low purity. The CART branches through the variance threshold of each node. When all the variance is lower than the threshold value of each node, or after reaching the set stop conditions, the CART decision tree is completed. In the GBDT algorithm, the CART decision tree is associated

with the boosting algorithm. In general, the residuals are calculated and evaluated after each iteration and processed as the input by the next iteration, thus minimizing the loss function and improving the fitting accuracy of the model. When the residuals reach a lowest value, or the setting termination condition has been reached, the model is constructed, and the regression result will be exported. The specific step of the GBDT regression algorithm is shown in the following sequence of equations and in Figure 5.

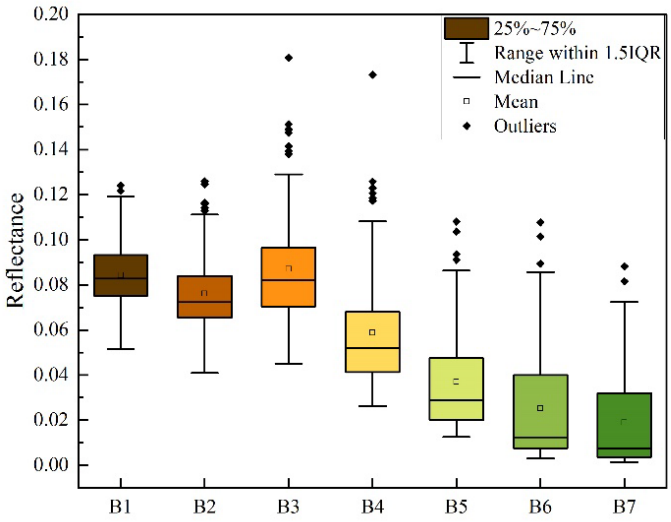

Figure 4. Reflectance boxplot of each band.

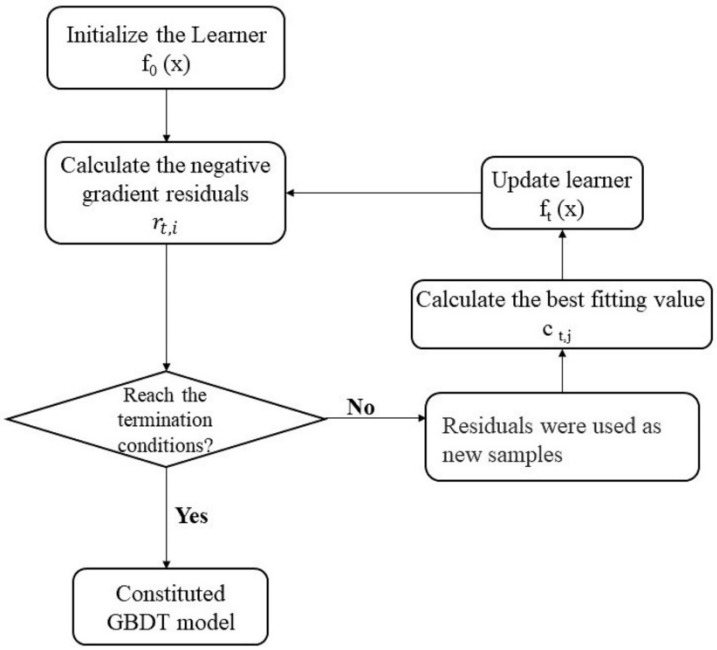

Figure 5. Flow chart of the Gradient Descent Boosting Decision Tree algorithm.

1. Initialize the cart learner:

$$f_0(x) = argmin_c \sum_{i=1}^{n} L(y_i, c) \quad (1)$$

2. In round t, the negative gradient of each sample is calculated:

$$r_{t,i} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f_{(x)}=f_{t-1}(x)} \quad (2)$$

3. The CART regression tree T_t is obtained by fitting $(x_i, r_{t,i})$, $i = 1, 2, \ldots, m$, and the leaf node region is divided into $R_{t,j}$, $j = 1, 2, \ldots, J$;

4. Traverse (referring to one visit to each node in the tree (or graph) along a certain search route) the node region, and calculate the output value of each leaf node R_t, namely the best fitting value $c_{t,j}$:

$$c_{t,j} = argmin_c \sum_{x_i \in R_{t,j}} L(y_i, f_{t-1}(x_i) + c) \quad (3)$$

5. Update the learner:

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{t,j} I(x \in R) \quad (4)$$

6. Repeat these steps until the termination condition is reached, and the final strong learner expression is obtained by adding the weak learners as follows:

$$f(x) = f_0(x) + \sum_{t=1}^{T} \sum_{j=1}^{J} c_{t,j} I(x \in R) \quad (5)$$

To avoid over- or underfitting of the model, the setting of the GBDT model was determined by grid searching with cross-validation. By 10-fold cross-validation, the method of cross validation used in this study, the dataset was separated into 10 subsamples. After one of the subsamples was randomly used as a testing set with the rest as a training set, the GBDT model was constructed, and the performance was evaluated. This process was repeated 10 times. By grid research, all the potential setting was traversed with 10-fold cross-validation, and the parameters with best performance were given. In this study, the mean square error (MSE) was set as a loss function, and the learning rate was set as 0.1. The number of the CART decision trees was set as 100, and the max depth of the decision tree was set as 3.

The inversion accuracy of the remote sensing inversion model was assessed using the mean absolute error (*MAE*), mean absolute percentage error (*MAPE*), and root mean square error (*RMSE*), which are defined as follows:

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |(y_i - f_i)| \quad (6)$$

$$MAPE(\%) = \frac{1}{n} \times \sum_{i=1}^{n} \left| \frac{(y_i - f_i)}{y_i} \times 100\% \right| \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (y_i - f_i)^2} \quad (8)$$

where n is the number of data pairs, the subscript i denotes individual data points, and y and f represent the measured and estimated values, respectively.

The correlation coefficient (R^2) was also measured to show how well the variation of one model explained the variation in the concentration of Chl-a. Generally, the largest R^2 with the smallest RMSE gives the best prediction model. In this study, the models with correlation coefficients (R^2) exceeding 0.7 would be selected to verify the inversion accuracy.

Theil–Sen and Mann-Kendall trend analysis were used to analyze the variation of Chl-a concentrations in the coastal sea surface of the Beibu Gulf of Guangxi. Theil–Sen and

Mann-Kendall trend analysis includes Theil–Sen slope estimation and the Mann-Kendall significance test, which does not require the dataset to meet the normal distribution on the time series, nor does it require a dataset correlation between time series, which is insensitive to outliers in time series and has a strong ability to avoid measurement errors in datasets or discrete data.

The remote sensing images with sensing dates ranging from 2013 to 2020 were converted into the trained GBDT model, and the output results were imported into ArcGIS 10.5 for raster processing. After elimination of the outlier values and Inverse Distance Weighted (IDW) interpolation, the spatial and temporal distributions of chlorophyll were visualized for analysis.

3. Results

3.1. Performance Assessment

In this study, single bands and the single-band ratio, band combination, and water index of the Landsta8 OLI images in the study area were used to establish the feature library of the GBDT model. The single-sample Kolmogorov–Smirnov test (K–S test) was used in SPSS software to determine whether these variables were in line with the normal distribution. The results showed that the value of progressive significance of the test samples was greater than 0.05, proving that the variables were in line with the normal distribution. In this case, Pearson's correlation analysis was used to test the importance of the features.

Features with high importance (correlation coefficient higher than 0.6) are shown in Table 3. The features with correlation coefficients higher than 0.7 (B4, B3 + B4, B3, B1 − B4, B2 + B4, B1 + B4, and B2 − B4) were selected as the input variables.

Table 3. Importance of modeling features.

Feature	Correlation Coefficient	Feature	Correlation Coefficient	Feature	Correlation Coefficient
B4	0.763 **	B2 − B3	−0.694 **	B4 + B7	0.674 **
B3 + B4	0.751 **	B4/B1	0.691 **	B4 + B6	0.668 **
B3	0.725 **	B4 + B5	0.689 **	B3 + B7	0.664 **
B1 − B4	−0.724 **	B1 − B3	−0.686 **	B3 + B6	0.660 **
B2 + B4	0.717 **	B2 + B3	0.686 **	B4/B2	0.647 **
B1 + B4	0.706 **	B3 + B5	0.680 **	FAI	−0.614 **
B2 − B4	−0.704 **	B1 + B3	0.675 **	B1/B4	−0.609 **

** Significant correlation at the 0.01 level (bilateral).

The input features of the GBDT model were added successively, and the accuracy of the inversion results was evaluated and compared (Table 4). The results demonstrated that the inversion accuracy was enhanced as more variables were added, suggesting that additional variables could significantly improve the performance of the GBDT model for the retrieval of the Chl-a concentration, and the GBDT model with all the selected variables performed with a higher accuracy (R^2 = 0.778).

Table 4. Results of assessing the accuracy of multiple feature variables.

Feature Variables	MAE (μg/L)	MAPE (%)	RMSE (μg/L)	R^2
B4	2.641	51.365	3.616	0.043
B4, B3 + B4	1.416	27.539	1.970	0.685
B4, B3 + B4, B3	1.387	26.988	1.912	0.695
B4, B3 + B4, B3, B1 − B4	1.284	24.968	1.793	0.729
B4, B3 + B4, B3, B1 − B4, B2 + B4	1.247	24.250	1.731	0.755
B4, B3 + B4, B3, B1 − B4, B2 + B4, B1 + B4	1.303	25.355	1.752	0.752
B4, B3 + B4, B3, B1 − B4, B2 + B4, B1 + B4, B2 − B4	0.998	19.414	1.626	0.778

The inversion results and the fitting of the measured values of the GBDT model constructed with B4, B3 + B4, B3, B1 − B4, B2 + B4, B1 + B4, and B2 − B4 as the input features are shown in Figure 6. The GBDT model performed well, as indicated (MAE = 0.998 μg/L, MAPE = 19.414%, RMSE = 1.626 μg/L, and R^2 = 0.778).

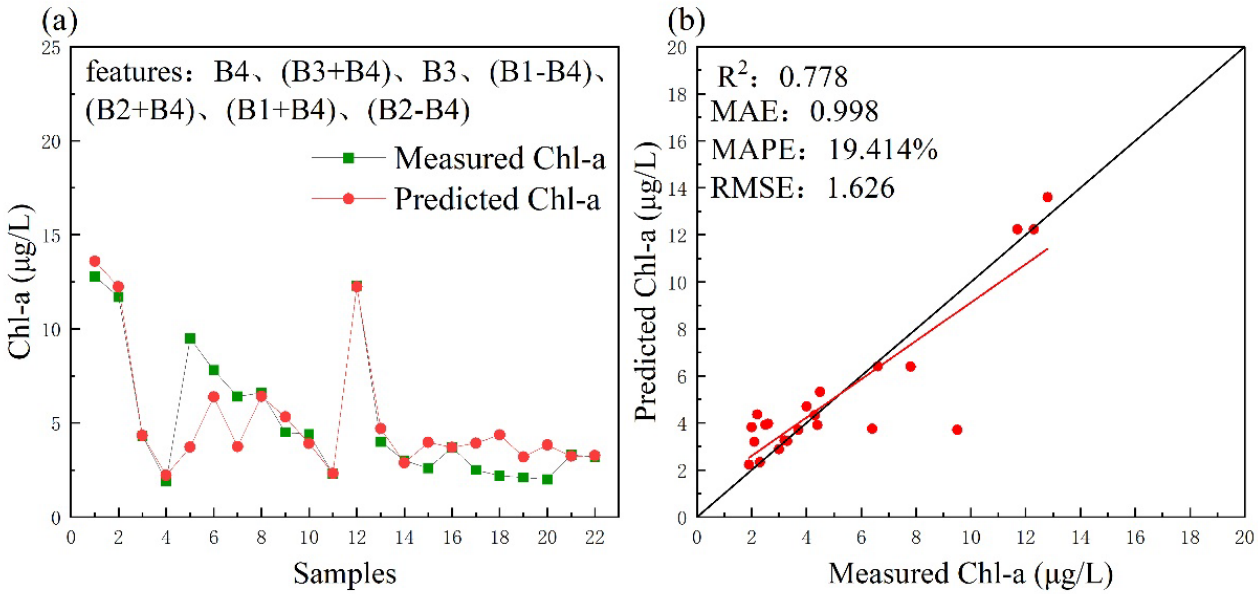

Figure 6. Accuracy verification results of seven characteristic variables. (**a**) Inversion results. (**b**) Fitting results.

3.2. Spatial–Temporal Distribution of Chl-a

3.2.1. Spatial Variations of Chl-a

Based on the inversion results of the Chl-a concentrations from 2013 to 2020, the distribution of the Chl-a concentrations in the coastal waters of the Guangxi Beibu Gulf was obtained by taking the average value (Figure 7). The average Chl-a concentration in each bay in the coastal waters of the Guangxi Beibu Gulf from 2013 to 2020 is shown in Table 5.

Table 5. Details of Chl-a concentration distribution of 2013~2020.

Bay	Minimum Value (μg/L)	Maximum Value (μg/L)	Average (μg/L)
Pearl Bay	1.283	10.082	5.031
Fangcheng Bay	1.571	5.034	3.372
Qinzhou Bay	1.508	13.003	6.600
Dafeng Estuary Bay	1.570	12.410	8.198
Nanliu Estuary Bay	0.836	19.703	11.469
Beihai	2.883	13.131	7.461

The concentration of Chl-a in the coastal sea surface of the Beibu Gulf of Guangxi was higher in the nearshore coastal waters and lower in the offshore waters, and it gradually decreased from north to south. The concentration of Chl-a in the Nanliu Estuary Bay was the highest, and the average of the whole region was 11.469 μg/L. The means of the Dafeng Estuary Bay, Beihai, Qinzhou Bay, Pearl Bay, and Fangcheng Bay were 8.198, 7.461, 6.600, 5.031, and 3.372 μg/L, respectively.

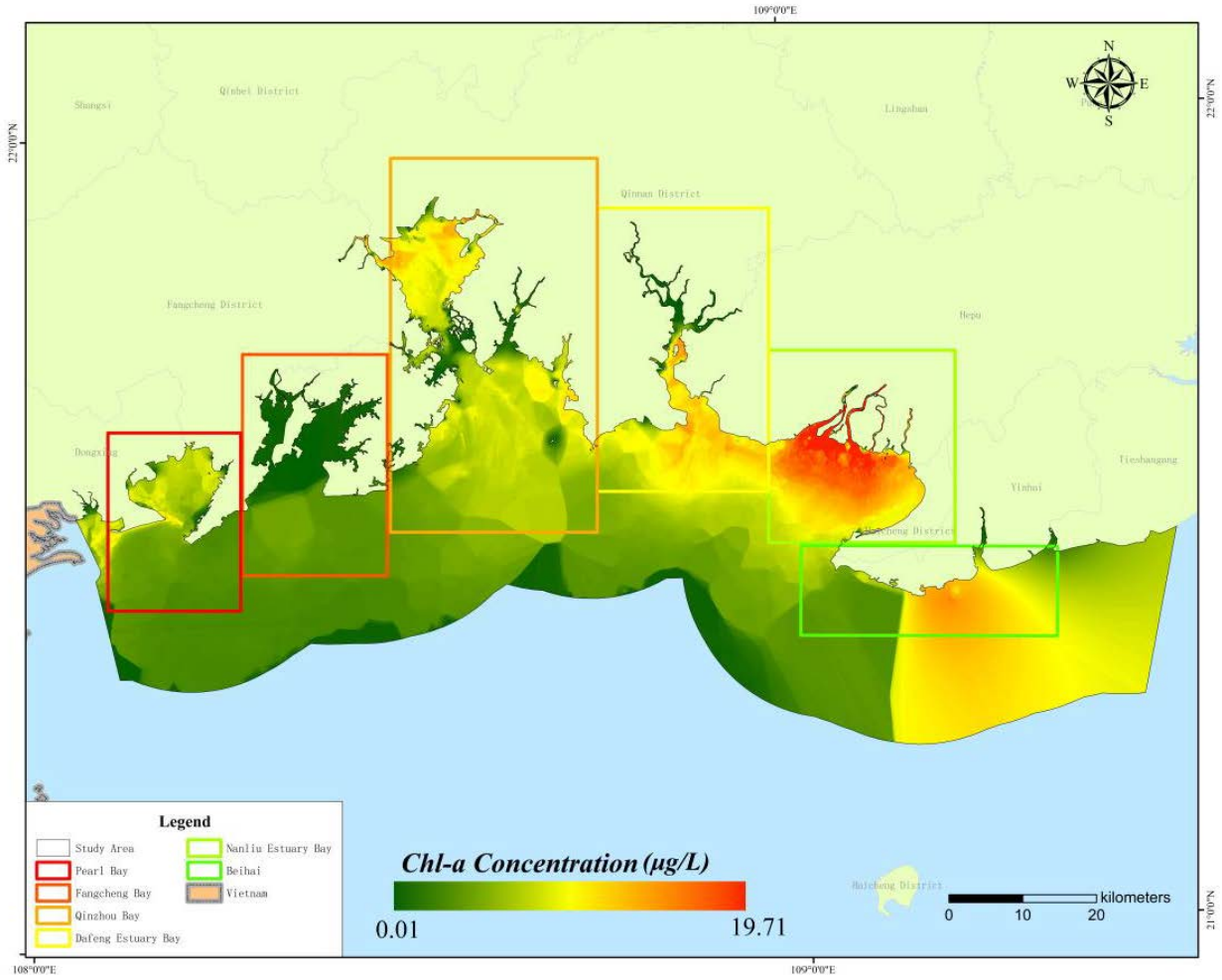

Figure 7. Distribution of Chl-a concentrations in Beibu Gulf in Guangxi.

3.2.2. Temporal Variations of Chl-a

The Chl-a concentrations in the four seasons in 2019 are presented in Figure 8.

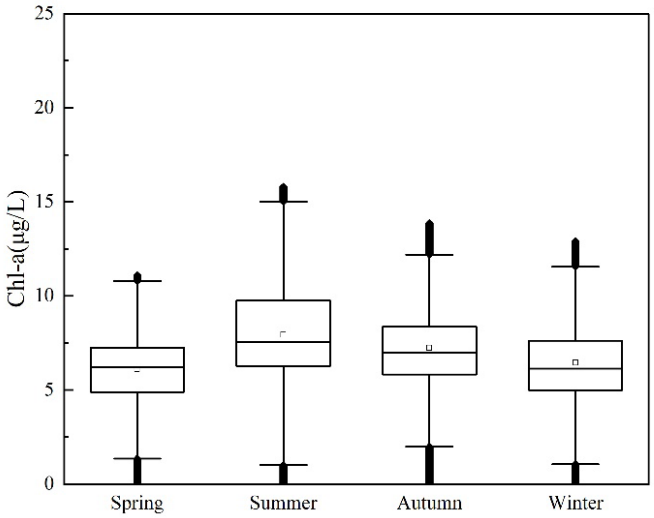

Figure 8. Seasonal changes of Chl-a.

The thermal and dynamic structures of the sea surfaces of the Beibu Gulf of Guangxi are affected by subtropical monsoons. The spatial and temporal differences in the Chl-a concentration in the Beibu Gulf were caused by the transport of nutrients. The concentration of Chl-a in the coastal sea surface exhibited clear seasonal changes. The average concentration of Chl-a in the summer was the highest (8.312 μg/L), while it was moderate during autumn (7.714 μg/L) and spring (6.954 μg/L) and lowest in winter (6.680 μg/L).

3.3. Theil–Sen and Mann-Kendall Trend Analysis

Theil–Sen and Mann-Kendall trend analysis demonstrated variations in the Chl-a concentrations in the coastal sea surfaces of the Guangxi Beibu Gulf (Figure 9), and the statistical results of each trend are shown in Table 6.

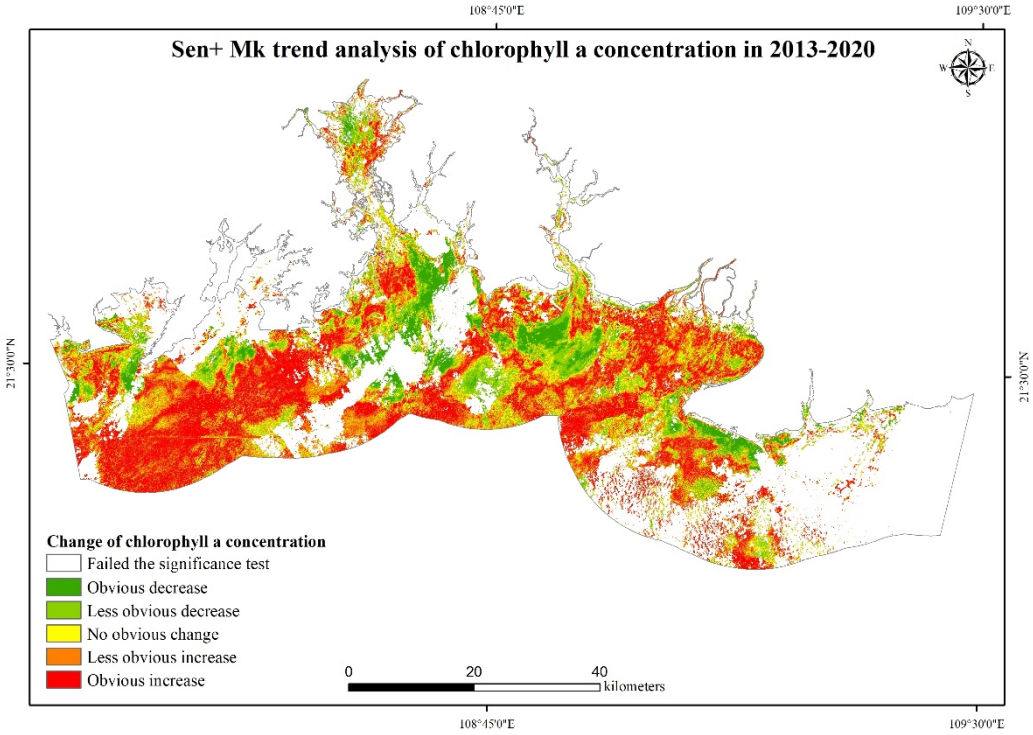

Figure 9. Trend analysis diagram of chlorophyll concentration.

Table 6. Statistical table of the area of each trend.

Trend of Chl-a Concentration	Area (km^2)
Obvious decrease	193.550
Less obvious decrease	356.720
No obvious change	383.690
Less obvious increase	724.450
Obvious increase	761.490
Failed the significance test	1721.900

The area of this study area was 4141.8 km^2, of which the area with an obvious decrease in its Chl-a concentration was 193.55 km^2, the area with a less obvious decrease was 356.72 km^2, the area with no obvious change was 383.69 km^2, the area with a less obvious increase was 724.45 km^2, and the area with an obvious increase was 761.49 km^2. The remaining 1721.900 km^2 in the study area showed no significant change. The concentration of Chl-a in the coastal sea surface has exhibited an increase in recent years.

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

4. Discussion

4.1. Comparison of Different Models

Two machine learning models were used for comparison in this study. An artificial neural network is a machine learning model that can explore the nonlinear relationship between the input variables and target data though training and adjusting the inside interconnected processing neurons [41]. The features of high prediction accuracy, self-adaptation, and robustness make it widely used for retrieval in waters with complex optical characteristics [42]. The support vector machine (SVM) is a useful tool for nonlinear statistical learning and regression analysis [43], whose training provides the support vectors to separate the classes in a multidimensional attribute space (the inland water's trophic status classification is based on machine learning and remote sensing data).

The artificial neural network and SVM model were trained in MATLAB software. Given the same input variables as the GBDT model, the performance of them for the estimation of the Chl-a concentration in the coastal waters of the Beibu Gulf in Guangxi were evaluated.

The results from verifying the accuracies of different models are compared in Figure 10 and Table 7.

Table 7. Comparison of model accuracies.

Model	Variables	MAE (μg/L)	MAPE (%)	RMSE (μg/L)	R^2
Single band	B3	3.381	65.758	3.705	0.563
	B4	3.006	58.470	2.903	0.719
Band ratio	B4/B1	1.967	38.261	1.935	0.706
Band combination	B2 + B3	2.035	39.582	2.248	0.637
	B2 + B4	1.898	36.927	2.029	0.671
	B3 + B4	1.795	34.913	1.959	0.698
Water index	FAI	1.843	35.884	2.226	0.591
SVM	B4, B3 + B4, B3, B1 − B4, B2 + B4, B1 + B4, B2 − B4	2.085	39.784	2.986	0.527
GBDT model	B4, B3 + B4, B3, B1 − B4, B2 + B4, B1 + B4, B2 − B4	0.998	19.414	1.626	0.778
Neural network	B4, B3 + B4, B3, B1 − B4, B2 + B4, B1 + B4, B2 − B4	1.492	28.472	1.974	0.714

The GBDT model performed well, as indicated in the statistical metrics. The results of the analyses of the accuracy demonstrated that the GBDT model exhibited the highest inversion accuracy for the Chl-a concentration, with an RMSE of 1.626; the remote sensing inversion model for the Chl-a concentration based on the GBDT algorithm exhibited advantages in the inversion accuracy.

The GBDT model performed better than the B4/B1 model and the other models tested. While the B4/B1 model reasonably estimated the Chl-a concentration (R^2 = 0.706; Figure 10b), it exhibited deviations at high and low Chl-a concentrations. The floating algae index (FAI), a water index strongly correlated with Chl-a which was proposed by Hu in 2009 [44], exhibited a relatively large deviation from the in situ data (R^2 = 0.591, RMSE = 2.226 μg/L, MAPE = 35.884%, and MAE = 1.843 μg/L) (Table 6; Figure 10f).

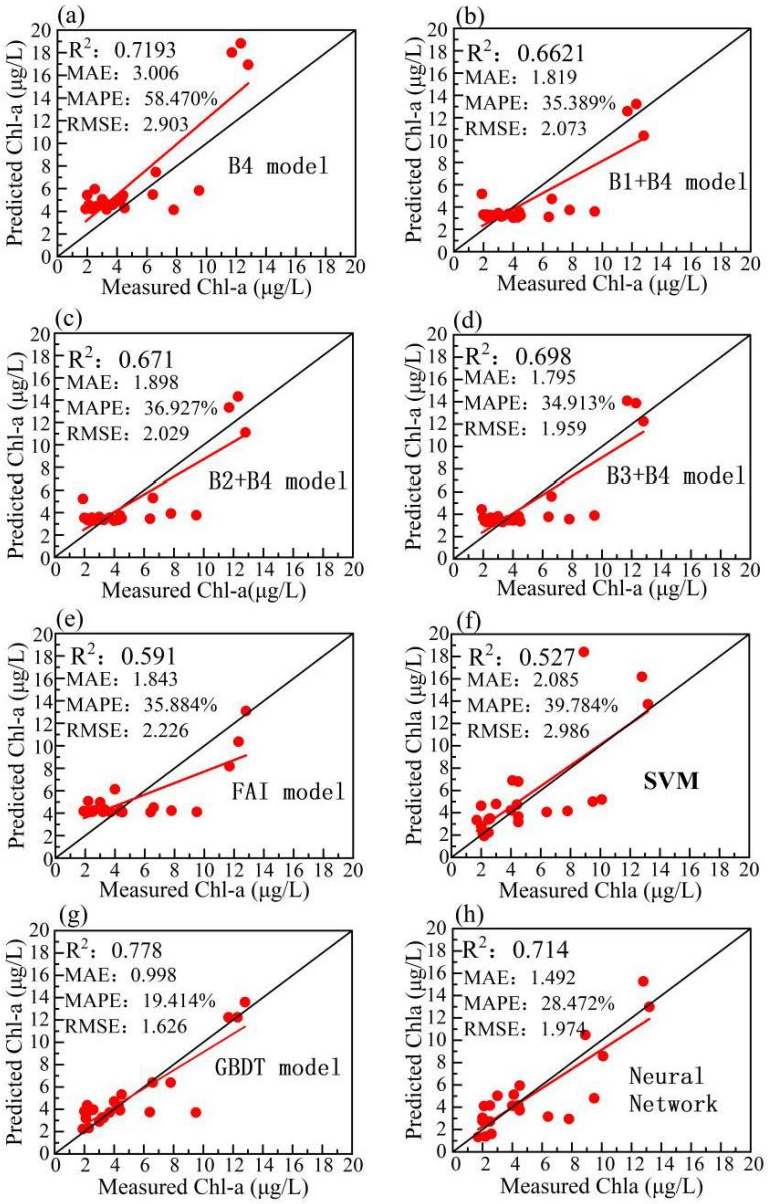

Figure 10. Performance evaluation of Chl-a retrievals using B4 model (**a**), B1 + B4 model (**b**), B2 + B4 model (**c**), B3 + B4 model (**d**), FAI model (**e**), SVM model (**f**), GBDT model (**g**) and neural network model (**h**).

Among the machine learning algorithms tested, the GBDT algorithm exhibited the best performance, while the SVM had the poorest performance (R^2 = 0.527, RMSE = 2.986 μg/L, MAPE = 39.784%, and MAE = 2.085 μg/L). Given the same input variables as the GBDT model, the neural network had a higher inversion accuracy than the traditional statistical models (R^2 = 0.706), showing a capacity to estimate the concentration of Chl-a in the coastal water. In general, the neural network tended to underestimate the Chl-a concentration, and the reason for this might be that the samples used to train and validate the neural network were inadequate. In a previous study (Song et al. [45]), the artificial neural network performed well for the estimation of the TSM and Chl-a concentration, but it required a large dataset for training. With a smaller dataset and faster computation, the GBDT model could generate a prediction algorithm, allowing better generalization and thus making it more appropriate for estimating the temporal variation of the Chl-a concentration than other models.

The concentration of Chl-a in the coastal sea surface of the Beibu Bay of Guangxi obtained by remote sensing inversion ranged from 0 to 35 μg/L, similar to the field survey results of Yang Bin and Zhong Qiuping et al. [46], indicating that the retrieval effect of the selected remote sensing inversion model for the Chl-a concentration was close to the field data, and the remote sensing inversion results had high reference significance.

4.2. Spatial and Temporal Distribution of Chl-a

4.2.1. Spatial Difference of Chl-a

The average distribution of the Chl-a concentration in the coastal waters of the Guangxi Beibu Gulf is shown in Figure 11. The concentration of coastal sea surface chlorophyll was high in the nearshore waters and low in the offshore waters, and it gradually decreased from north to south. Nutrients in the Beibu Gulf in Guangxi are mainly transported from coastal land sources [47], and the pollution inputs from the Nanliu River are the largest. The Nanliu River basin is large, bringing together industrial, agricultural, and urban sewage from adjacent land, which carries a large amount of nutrients. According to the Marine Environment Quality Bulletin of the Guangxi Zhuang Autonomous Region, in 2016, the inflow COD_{Cr} of the Nanliu River in 2016 was 174,049 t, and the total amount of ammonia nitrogen, nitrate nitrogen, and nitrite nitrogen was 9868 t. The estuarine and coastal zones are transitional zones between land and sea with relatively shallow water depths, so river inflows with high-input terrigenous nutrient content mix with offshore salt water. The annual average SST is about 20.3~29.9 °C. High light intensity and high SST promote the growth and reproduction of phytoplankton and algae, resulting in an increase in the Chl-a concentration [48].

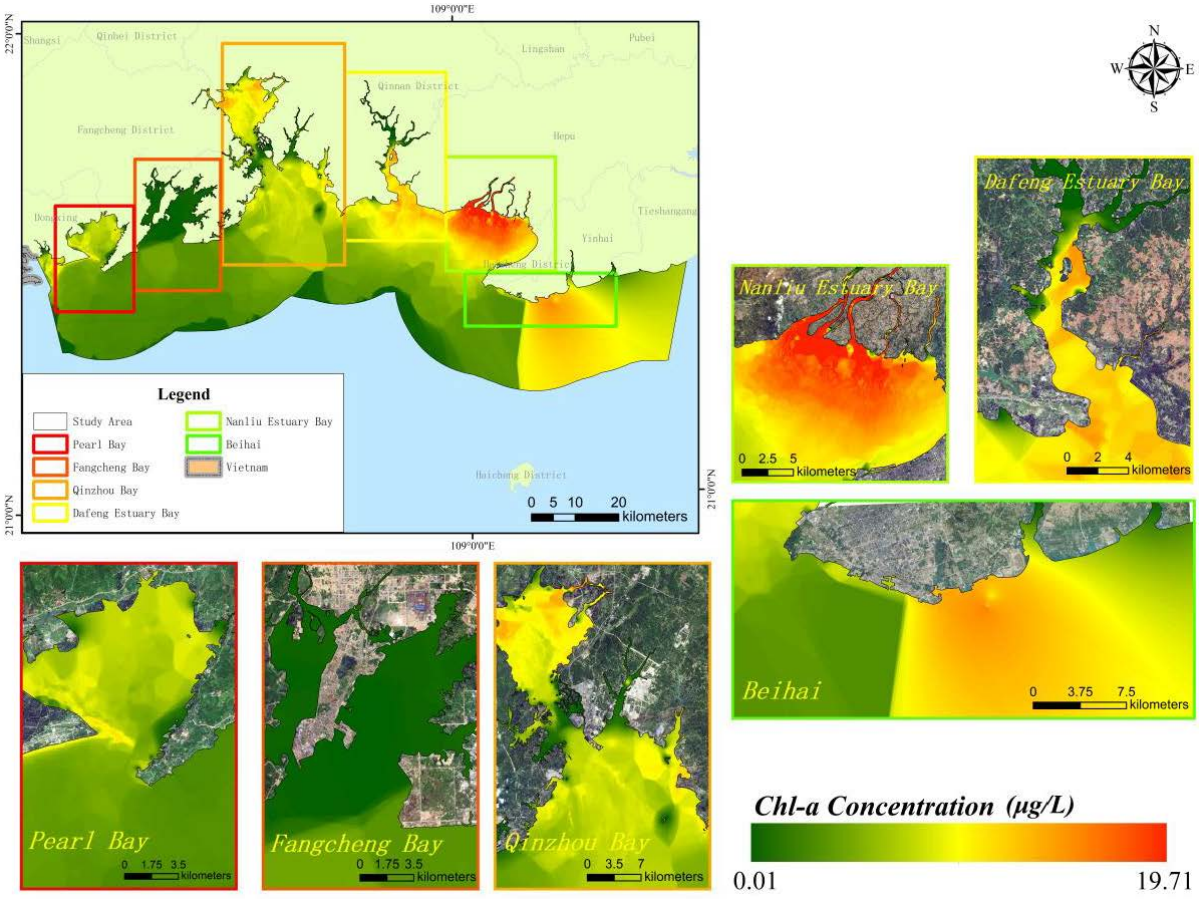

Figure 11. Distribution of the Chl-a concentration in the coastal waters of the Beibu Gulf in Guangxi.

The runoff diluted water lifts the level in the nearshore sea, and the offshore sea level is relatively low; nutrients then plume out to sea from the estuary. The Beibu Gulf in Guangxi has a tropical and subtropical monsoon climate. The monsoon strengthens the emergence of a current alongshore area, which is also affected by the Coriolis force, and the seaward runoff moves westward along the coast. Therefore, from the perspective of the source, the concentration of Chl-a in the coastal waters of the Beibu Gulf in Guangxi was higher in the nearshore water and lower in the offshore water due to the input of terrestrial nutrients. From the perspective of transport, the nutrients in the Beibu Gulf in Guangxi diffuse from the estuary to the offshore sea. Under the influence of the monsoon, the nutrients flow westward along the coast, and the concentration of Chl-a gradually decreases when moving westward.

4.2.2. Temporal Variation of Chl-a

The concentration of Chl-a in the coastal sea surface of the Beibu Gulf in Guangxi exhibited strong seasonal changes, with the following ranking being apparent: summer > autumn > spring and winter (Figure 12).

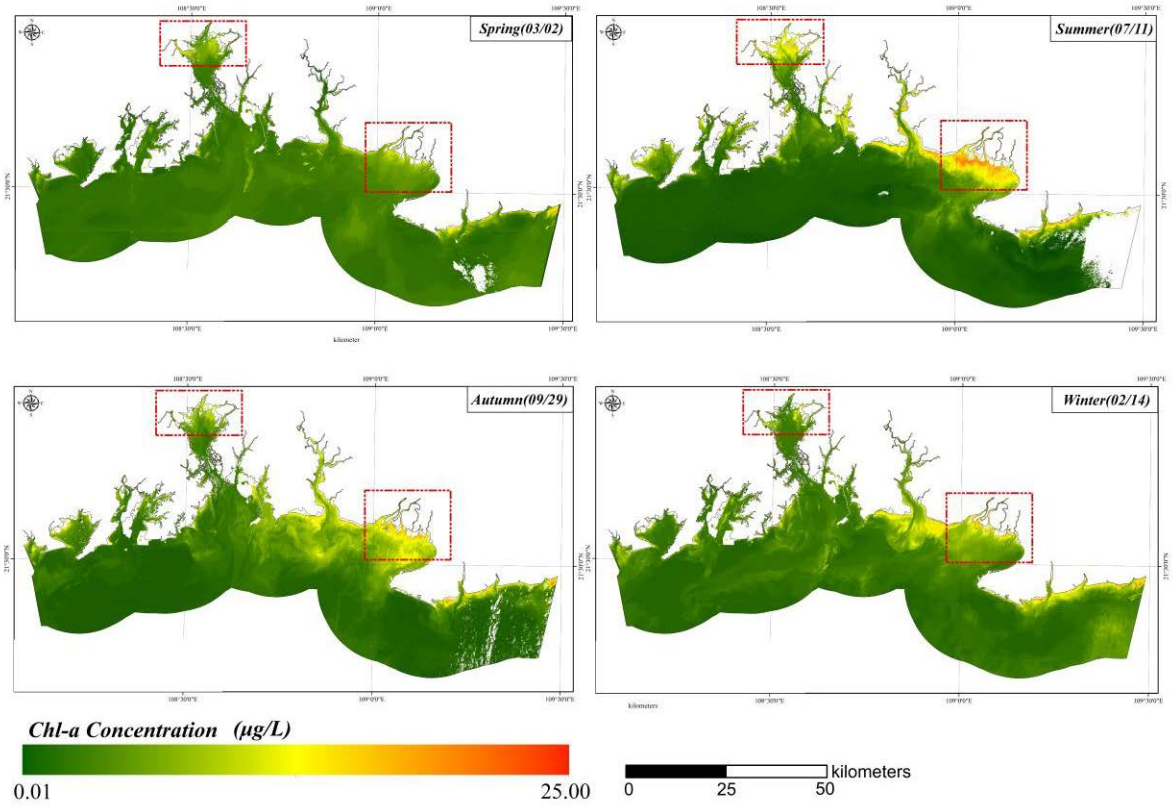

Figure 12. Distribution of Chl-a concentrations in the summer and autumn of 2018.

The Chl-a concentration in summer and autumn was mainly affected by the climate [49,50]. In summer, the Beibu Gulf receives abundant rainfall. The wet season ranges from July to September, and the average monthly precipitation exceeds 100 mm, accounting for approximately 55–70% of the total annual rainfall. River runoff is the largest in summer, and many rivers along the coast flow into the Beibu Gulf, taking agricultural and industrial wastewater into the gulf. The inflows have strong diffusion force, carrying large amounts and high concentrations of nutrients. As the sea level is raised by the runoff, the resultant concentration gradient strengthens the water exchange and increases the thickness of the mixing layer. There are high temperatures in summer, with the highest SST being up to 34.1 °C. Under such environmental conditions, phytoplankton proliferate in large numbers. Therefore, the concentration of Chl-a was the highest in the summer, and the area with a

high value was concentrated in the estuarine area of the region. The southwest monsoon prevails in summer, generating a coastal component of the wind force, and it flows along the coast, driven by the monsoon [51]. The runoff raised the sea level and further promoted southwest flow along the coast by the gradient in the sea level. Therefore, in summer, nutrients along the Beibu Gulf are transported from east to west. The Chl-a concentration declined gradually from east to west. In autumn, rainfall decreases, and runoff into the sea decreases. The river-diluted water contracts to the shore, and the concentration of Chl-a decreases compared with that in the summer.

In spring and winter, the river runoff was the smallest, the terrestrial nutrient input was lower, the sea surface temperature was the lowest (19.0–24.0 °C), the growth of phytoplankton was inhibited, and the Chl-a concentration in the sea surface was the lowest. The wind is light over the Beibu Gulf in spring, and the effect of the Beibu Gulf monsoon on aquatic mixing is greatly reduced, causing a smaller mixing layer. Additionally, controlled by the reduction in flow, the Chl-a concentration is generally low in the spring and winter in the Beibu Gulf. The climate data of the sampling point (GX10 referenced as an example) on the sensing date of the satellite is shown in Table 8.

Table 8. Climate data of the sampling point (GX10) on the sensing date of the satellite.

Date	Temperature	Wind Direction	Wind Strength
14 April 2015	14.3	SE	<3
23 October 2015	27.2	S	<3
3 June 2016	31.2	S	<3
28 December 2016	18	N	1
14 February 2017	17.44	N	1
2 March 2017	17.22	SW	1
28 October 2017	25.72	N	3–4

5. Conclusions

Using Landsat 8 OLI remote sensing images combined with measured Chl-a concentrations, the GBDT model was used to study the coastal waters of the Beibu Gulf in Guangxi and analyze the spatial–temporal distribution of the Chl-a concentration. The main research results are as follows:

1. Compared with the performance of different models, the GBDT model can significantly improve the accuracy of Chl-a concentration inversion, proving that it can be a new method for remote sensing inversion of the water quality parameters. When B4, B3 + B4, B3, B1 − B4, B2 + B4, B1 + B4, and B2 − B4 were considered the characteristic variables of the GBDT model, the inversion accuracy of the model was the highest (MAE = 0.998 μg/L, MAPE = 19.413%, RMSE = 1.626 μg/L, and R^2 = 0.778).
2. The spatial distribution of the Chl-a concentration was highest in the nearshore and lowest in the offshore waters in the Beibu Gulf in Guangxi. The Chl-a concentration was highest in the summer, and the concentration in autumn was lower, while concentrations in spring and winter were the lowest. The ranking of Chl-a concentrations, from high to low, across multiple bays was as follows: Nanliu River Estuary Bay, Dafeng River Estuary Bay, Qinzhou Bay, Beihai Pearl Harbor, and Fangcheng Bay.

Limited by the revisiting period of the satellite and the quality of the satellite images, the data used to train and validate the GBDT model may be relatively small in size, which may have influenced the inversion accuracy of the model. To estimate and learn the spatial and temporal distribution of Chl-a concentrations more precisely, an increased amount of data may be included in future work in combination with multi-source satellite remote sensing data.

Author Contributions: Resources, H.Y.; writing—original draft preparation, Y.H.; formal analysis, Y.W.; data curation, W.Z.; visualization, K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Landsat 8 OLI satellite data (https://glovis.usgs.gov (accessed on 1 January 2021)).

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

References

1. Brooks, B.W.; Lazorchak, J.M.; Howard, M.D.A.; Johnson, M.V.; Morton, S.L.; Perkins, D.A.K.; 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]
2. Carmichael, W.W. Health effects of toxin-producing cyanobacteria: "The CyanoHABs". *Hum. Ecol. Risk Assess.* **2001**, *7*, 1393–1407. [CrossRef]
3. Carvalho, L.; McDonald, C.; de Hoyos, C.; Mischke, U.; Phillips, G.; Borics, G.; Poikane, S.; Skjelbred, B.; Solheim, A.L.; Van Wichelen, J.; et al. Sustaining recreational quality of European lakes: Minimizing the health risks from algal blooms through phosphorus control. *J. Appl. Ecol.* **2013**, *50*, 315–323. [CrossRef]
4. Duan, H.; Ma, R.; Xu, X.; Kong, F.; Zhang, S.; Kong, W.; Hao, J.; Shang, L. Two-Decade Reconstruction of Algal Blooms in China's Lake Taihu. *Environ. Sci. Technol.* **2009**, *43*, 3522–3528. [CrossRef]
5. Gao, D.; Li, C.; Liu, G.; Zhang, H. The species composition and distribution of phytoplankton in the Beibu Bay. *J. Zhanjiang Ocean Univ.* **2001**, *21*, 13–18.
6. Dörnhöfer, 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]
7. Li, X.; Wei, A.; Jiang, S.; Wang, T.; Ji, X.; Zhang, Y.; Jiao, X. Retrieval of chlorophyll-a and total suspended matter concentrations from sentinel-3OLCI imagery by C2RCC algorithm in south yellow sea. *Environ. Monit.* **2020**, *12*, 6–12.
8. Li, Y.; Huang, J.; Wei, Y.; Lu, W. Inversing Chlorophyll Concentration of Taihu Lake by Analytic Model. *Natl. Remote Sens. Bull.* **2006**, *10*, 169–175.
9. Yang, W.; Chen, J.; Mausushita, B. Algorithm for Estimating Chlorophyll-a Concentration in Case II Water Body Based on Bio-Optical Model. *Spectrosc. Spectr. Anal.* **2009**, *29*, 38–42.
10. Chang, N.; Imen, S.; Vannah, B. Remote Sensing for Monitoring Surface Water Quality Status and Ecosystem State in Relation to the Nutrient Cycle: A 40-Year Perspective. *Crit. Rev. Environ. Sci. Technol.* **2015**, *45*, 101–166. [CrossRef]
11. Sagan, V.; Peterson, K.T.; Maimaitijiang, M.; Sidike, P.; Sloan, J.; Greeling, B.A.; Maalouf, S.; Adams, C. Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Sci. Rev.* **2020**, *205*, 103187. [CrossRef]
12. 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]
13. Xue, K.; Zhang, Y.; Duan, H.; Ma, R.; Loiselle, S.; Zhang, M. A Remote Sensing Approach to Estimate Vertical Profile Classes of Phytoplankton in a Eutrophic Lake. *Remote Sens.* **2015**, *7*, 14403–14427. [CrossRef]
14. Pyo, J.; Duan, H.; Baek, S.; Kim, M.S.; Jeon, T.; Kwon, Y.S.; Lee, H.; Cho, K.H. A convolutional neural network regression for quantifying cyanobacteria using hyperspectral imagery. *Remote Sens. Environ.* **2019**, *233*, 111350. [CrossRef]
15. Liu, H.; Yan, L. Back-Propagation Network Model for Predicting the Change of Eutrophication of Qiandao Lake. *Bull. Sci. Technol.* **2008**, *24*, 411–416.
16. 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] [PubMed]
17. Deng, L.; Zhou, W.; Cao, W.; Zheng, W.; Wang, G.; Xu, Z.; Li, C.; Yang, Y.; Hu, S.; Zhao, W. Retrieving Phytoplankton Size Class from the Absorption Coefficient and Chlorophyll A Concentration Based on Support Vector Machine. *Remote Sens.* **2019**, *11*, 1054. [CrossRef]
18. Peterson, K.T.; Sagan, V.; Sidike, P.; Cox, A.L.; Martinez, M. Suspended Sediment Concentration Estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers Using an Extreme Learning Machine. *Remote Sens.* **2018**, *10*, 1503. [CrossRef]
19. Gonzalez Vilas, L.; Spyrakos, E.; Torres Palenzuela, J.M. Neural network estimation of chlorophyll a from MERIS full res-olution data for the coastal waters of Galician rias (NW Spain). *Remote Sens. Environ.* **2011**, *115*, 524–535. [CrossRef]
20. Pahlevan, N.; Smith, B.; Schalles, J.; Binding, C.; Cao, Z.; Ma, R.; Alikas, K.; Kangro, K.; Gurlin, D.; Nguyen, H.; et al. Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A ma-chine-learning approach. *Remote Sens. Environ.* **2020**, *240*, 111604. [CrossRef]

21. Wang, Q.; Chen, D.; Gao, X.; Wang, F.; Li, J.; Liao, W.; Wang, Z.; Xie, G. Microscopic pore structures of tight sandstone reservoirs and their diagenetic controls: A case study of the Upper Triassic Xujiahe Formation of the Western Sichuan Depression, China. *Mar. Petrol. Geol.* **2020**, *113*, 104119. [CrossRef]
22. Sagi, O.; Rokach, L. Approximating XGBoost with an interpretable decision tree. *Inform. Sci.* **2021**, *572*, 522–542. [CrossRef]
23. Zhang, J.; Liang, Q.; Jiang, R.; Li, X. A Feature Analysis Based Identifying Scheme Using GBDT for DDoS with Multiple Attack Vectors. *Appl. Sci.* **2019**, *9*, 4633. [CrossRef]
24. Wang, C.; Zhang, J.; Yu, G. Cluster Analysis of Pedestrian Mobile Channels in Measurements and Simulations. *Appl. Sci.* **2019**, *9*, 886. [CrossRef]
25. Kawatani, T.; Yamaguchi, T.; Sato, Y.; Maita, R.; Mine, T. Prediction of Bus Travel Time over Intervals between Pairs of Adjacent Bus Stops Using City Bus Probe Data. *Int. J. Intell. Transp. Syst. Res.* **2021**, *19*, 456–467.
26. Hou, C.; Cao, B.; Fan, J. A data-driven method to predict service level for call centers. *IET Commun.* **2021**, *2*, 1–12. [CrossRef]
27. Sun, R.; Wang, G.; Cheng, Q.; Fu, L.; Chiang, K.; Hsu, L.; Ochieng, W.Y. Improving GPS Code Phase Positioning Accuracy in Urban Environments Using Machine Learning. *IEEE Internet Things J.* **2021**, *8*, 7065–7708. [CrossRef]
28. Huang, P.; Wang, L.; Hou, D.; Lin, W.; Yu, J.; Zhang, G.; Zhang, H. A feature extraction method based on the entropy-minimal description length principle and GBDT for common surface water pollution identification. *J. Hydroinform.* **2021**, jh2021060. [CrossRef]
29. Zhao, D.; Zhu, L.; Sun, H.; Li, J.; Wang, W. Fengyun-3D/MERSI-II Cloud Thermodynamic Phase Determination Using a Machine-Learning Approach. *Remote Sens.* **2021**, *13*, 2251. [CrossRef]
30. Zou, Y.; Chen, Y.; Deng, H. Gradient Boosting Decision Tree for Lithology Identification with Well Logs: A Case Study of Zhaoxian Gold Deposit, Shandong Peninsula, China. *Nat. Resour. Res.* **2021**, 1–21. [CrossRef]
31. Li, R.; Cui, L.; Zhao, Y.; Zhou, W.; Fu, H. Long-term trends of ambient nitrate (NO_3^-) concentrations across China based on ensemble machine-learning models. *Earth Syst. Sci. Data* **2021**, *13*, 2147–2163. [CrossRef]
32. Chen, J.; Huang, G.; Chen, W. Towards better flood risk management: Assessing flood risk and investigating the potential mechanism based on machine learning models. *J. Environ. Manag.* **2021**, *293*, 112810. [CrossRef]
33. Wang, J.; Li, P.; Ran, R.; Che, Y.; Zhou, Y. A Short-Term Photovoltaic Power Prediction Model Based on the Gradient Boost Decision Tree. *Appl. Sci.* **2018**, *8*, 689. [CrossRef]
34. Zhang, T.; He, W.; Zheng, H.; Cui, Y.; Song, H.; Fu, S. Satellite-based ground PM2.5 estimation using a gradient boosting decision tree. *Chemosphere* **2021**, *268*, 128801. [CrossRef] [PubMed]
35. Meng, R.; Shen, W.; Ji, Q.; Rao, Y.; Hao, L. The application of GBDT model in remote sensing water depth introverse. *Environ. Ecol.* **2021**, *3*, 1–5.
36. Zhang, W.; Wei, Q.; Wu, T.; Lin, J.; Shao, G.; Ding, M. Prediction models of reference crop evapotranspiration based on gradient boosting decision tree(GBDT) algorithm in Jiangsu province. *Jiangsu J. Agric. Sci.* **2020**, *36*, 1169–1180.
37. Li, S.; Huang, H.; Dai, Z. Climate Change and Its Adaptation in Beibu Gulf of Guangxi in Recent 60 Years. *Ocean Dev. Manag.* **2017**, *34*, 50–55.
38. Xu, J. Preliminary study on Marine water quality monitoring system in Guangxi Beibu Gulf and its application in emergency monitoring. *Sci. Technol. Assoc. Forum* **2012**, *11*, 136–137.
39. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
40. Huo, S.; He, Z.; Su, J.; Xi, B.; Zhu, C. Using artificial neural network models for eutrophication prediction. *Procedia Environ. Sci.* **2013**, *18*, 310–316. [CrossRef]
41. Li, Y. Remote Sensing Retrieval Model for Chlorophyll-A Concentration of Water in Backwater Area, Three Gorges Reservioir. Master's Thesis, China University of Geosciences, Beijing, China, 2017.
42. Ye, H.; Yang, C.; Tang, S.; Chen, C. The phytoplankton variability in the Pearl River estuary based on VIIRS imagery. *Cont. Shelf Res.* **2020**, *207*, 104228. [CrossRef]
43. Hu, C. A novel ocean color index to detect floating algae in the global oceans. *Remote Sens. Environ.* **2009**, *113*, 2118–2129. [CrossRef]
44. Song, K.; Li, L.; Wang, Z.; Liu, D.; Zhang, B.; Xu, J.; Du, J.; Li, L.; Li, S.; Wang, Y. Retrieval of total suspended matter (TSM) and chlorophyll-a (Chl-a) concentration from remote-sensing data for drinking water resources. *Environ. Monit. Assess.* **2012**, *184*, 1449–1470. [CrossRef] [PubMed]
45. Yang, B.; Zhong, Q.; Zhang, C.; Lu, D.; Liang, Y.; Li, S. Spatio-temporal variations of chlorophyll a and primary productivity and its influence factors in Qinzhou Bay. *Acta Sci. Circumstantiae* **2015**, *35*, 1333–1340.
46. Li, P.; Guo, Z.; Mo, H.; Wang, D.; Lin, M. Temporal and spatial distribution of Guangxi inshore nutrients and evaluation of its potential eutrophication. *Trans. Oceanol. Limnol.* **2018**, *3*, 148–156.
47. Yu, Y.; Xing, X.; Liu, H.; Yuan, Y.; Wang, Y.; Chai, F. The variability of chlorophyll-a and its relationship with dynamic factors in the basin of the South China Sea. *J. Mar. Syst.* **2019**, *200*, 103230. [CrossRef]
48. Huynh, H.T.; Alvera-Azcarate, A.; Beckers, J. Analysis of surface chlorophyll a associated with sea surface temperature and surface wind in the South China Sea. *Ocean Dynam.* **2020**, *70*, 139–161. [CrossRef]

49. Wang, Y. Composite of Typhoon-Induced Sea Surface Temperature and Chlorophyll-a Responses in the South China Sea. *J. Geophys. Res.-Ocean.* **2020**, *125*, e2020JC016243. [CrossRef]
50. Chen, B.; Xu, G.; Ya, H.; Chen, X.; Xu, Z.; Shi, M. Transactions of oceanology and limnology. *Trans. Oceanol. Limnol.* **2020**, *2*, 43–54.
51. Liu, D.; Zhao, Q. Study on the spatial and temporal distribution of chlorophyll a concentration in Beibu gulf. *J. Mar. Sci.* **2019**, *37*, 95–102.