# Enhancing Algal Bloom Level Monitoring with CYGNSS and Sentinel-3 Data

Yan Jia [1], Zhiyu Xiao [1], Liwen Yang [1], Quan Liu [1], Shuanggen Jin [2,3], Yan Lv [4] and Qingyun Yan [5,*]

1 Department of Surveying and Geoinformatics, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; yan.jia@njupt.edu.cn (Y.J.); 1022173214@njupt.edu.cn (Z.X.); 1024072119@njupt.edu.cn (L.Y.); 1023172916@njupt.edu.cn (Q.L.)
2 School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454003, China; sgjin@shao.ac.cn
3 Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China
4 Faculty of Electronic Information Engineering, Huaiyin Institute of Technology, Huaian 223003, China
5 School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
* Correspondence: 003257@nuist.edu.cn

**Abstract:** Algal blooms, resulting from the overgrowth of algal plankton in water bodies, pose significant environmental problems and necessitate effective remote sensing methods for monitoring. In recent years, Global Navigation Satellite System–Reflectometry (GNSS-R) has rapidly advanced and made notable contributions to many surface observation fields, providing new means for identifying algal blooms. Additionally, meteorological parameters such as temperature and wind speed, key factors in the occurrence of algal blooms, can aid in their identification. This paper utilized Cyclone GNSS (CYGNSS) data, Sentinel-3 OLCI data, and ECMWF Re-Analysis-5 meteorological data to retrieve Chlorophyll-a values. Machine learning algorithms were then employed to classify algal blooms for early warning based on Chlorophyll-a concentration. Experiments and validations were conducted from May 2023 to September 2023 in the Hongze Lake region of China. The results indicate that classification and early warning of algal blooms based on CYGNSS data produced reliable results. The ability of CYGNSS data to accurately reflect the severity of algal blooms opens new avenues for environmental monitoring and management.

**Keywords:** GNSS-R; CYGNSS; reflectivity; machine learning; algal blooms

## 1. Introduction

The development of society and the rise of urban construction near inland lakes with abundant water resources have led to increased industrial, agricultural, and residential activities, contributing to significant water pollution and eutrophication [1]. The presence of phytoplankton indicates eutrophication [2,3]. These blooms deplete oxygen levels in the water and produce toxins, posing serious threats to aquatic life and the safety of drinking water for nearby cities [4,5]. Consequently, effective methods and tools for detecting algal blooms are crucial for the protection and management of water resources in lakes and surrounding areas. To minimize the negative impacts of these blooms, researchers have adopted various detection methods. The in situ measurement methods involve directly measuring the concentration of various algae in the water. However, these methods are limited in their monitoring scope and the number of samples they can obtain [6]. Additionally, field measurements are costly, time-consuming, and inadequate for quickly obtaining comprehensive information about blooms across lakes and monitoring their changes.

Over decades of development, satellite remote sensing technology has been successfully applied to the monitoring of algal blooms and related research. Remote sensing offers the advantages of low cost and wide observation coverage, allowing for the rapid

acquisition of large-area feature information [7]. However, the temporal resolution of optical remote sensing is generally low; for example, the revisit period of Landsat-8 is 16 days, while Sentinel-2 missions are at least 5 days [8]. Moreover, optical remote sensing is significantly hindered by cloud cover, making it difficult to obtain reliable images on cloudy or rainy days [9], which results in incomplete observation sequences and further lengthens the time intervals between image acquisitions [10]. Microwave remote sensing is also a major means of Earth observation, with synthetic aperture radar (SAR) being one of the primary methods. SAR detects the scattering surface by receiving the backscattering of transmitted electromagnetic waves. SAR is sensitive to algal blooms, as Wang et al. [11] discovered that SAR backscattering is suppressed at the lake surface where blooms occur, resulting in dark zones in SAR images. Bresciani et al. [12] demonstrated the feasibility of correlating Chlorophyll-a ($chl\_a$) concentration with SAR backscattering coefficients during different bloom phases. They retrieved $chl\_a$ concentration and SAR backscattering coefficients at various stages of the bloom, showing a viable correlation between the two. Although SAR images can overcome weather constraints that affect optical images, the revisit period of most onboard SAR missions remains not short enough (at least 7 days), and their temporal resolution is not significantly better than that of optical images [13].

Global Navigation Satellite System–Reflectometry (GNSS-R) is an emerging microwave remote sensing technology that uses specific receivers to capture GNSS signals reflected from the ground to observe Earth's surface [14]. Its main features include a wide range of signal sources, global coverage of major land and oceans, and low cost [15]. GNSS-R satellites offer flexible spatial resolution and very short revisit periods. Operating in the L-band, they can penetrate clouds, are unaffected by weather conditions, and can monitor reflective surfaces on a 24 h, all-weather basis [16,17]. In recent years, GNSS-R data have been applied in various fields, such as sea surface wind retrieval, sea ice measurement [18,19], ocean altimetry [20], flood and inland water mapping [21–23], and soil moisture retrieval [24–26].

Algal bloom detection has become a key GNSS-R research topic in recent years. Algal blooms smooth the water surface, enhancing the forward scattering strength of radar waves. GNSS signals reflecting off smooth water surfaces exhibit coherent reflections, and changes in water surface roughness can be effectively identified using the Doppler Delay Map of the onboard GNSS-R. These capabilities make Cyclone GNSS (CYGNSS) data feasible for hydrographic monitoring. CYGNSS is the latest constellation mission using GNSS-R technology and is also widely employed in the study. Rodriguez-Alvarez et al. [27] were the first to use CYGNSS satellite data to analyze changes in ocean surface roughness in the Gulf of Mexico to detect algal blooms. Ban et al. [28] proposed a model to estimate red tide density at the sea surface from GNSS-R observations, demonstrating the potential of GNSS-R technology for rapid preliminary monitoring of red tides. Zhang et al. [29] utilized the power ratio of GNSS-R data to identify blooms in Taihu Lake by analyzing coherent reflections on the water surface where blooms occur, also discussing the impact of wind speed on identification results. Zhen et al. [30] combined GNSS-R data with auxiliary meteorological data to detect blooms in Taihu Lake, showing that incorporating meteorological data can improve detection accuracy. These studies primarily rely on vegetation indices (e.g., Normalized Difference Vegetation Index, NDVI) as validated data. The method is not sensitive enough to recognize mild to slight "algal blooms" and "algal blooms" with low algal densities. When the NDVI is negative, it is difficult to choose a suitable threshold between algal bloom and algal bloom, which leads to inconsistency in the interpretation of algal bloom distribution and area. Consequently, using vegetation indices as validation data may lead to false positives in bloom modeling results, causing false alarms. Additionally, due to their characteristic thickness and stickiness, algae floating on the lake's surface reduce the tension on the water's surface. During bloom formation, algae multiply and aggregate, evolving from a hidden bloom to a dominant one, eventually smoothing the water reflection surface and increasing the coherent reflection component received by GNSS-R. Current research primarily focuses on identifying algal blooms after

major outbreaks, with insufficient monitoring of less severe blooms and their development processes, limiting the practical application and impact of these research findings.

To accurately monitor and predict algal blooms, this paper uses Sentinel-3 OLCI (Ocean and Land Color Instrument) data to retrieve *chl_a* concentration and proposes a machine learning (ML) model for bloom monitoring using CYGNSS reflectivity and meteorological data. The validation of the *chl_a* concentration retrieval results and the prediction model of algal bloom outbreak level are conducted in the Hongze Lake area of China. This paper is organized as follows: Section 2 describes the study area and the data used; Section 3 describes the methodology for retrieval of *chl_a* concentration, the strategy for grading algal bloom levels, and the adopted ML prediction model; Section 4 presents the classification results obtained; Section 5 gives the discussion of the article; and Section 6 provides the conclusions of this study.

## 2. Study Area and Materials

### 2.1. Study Area

Hongze Lake (Figure 1) is located in the lower reaches of the Huaihe River in northwestern Jiangsu Province, China. It is the largest lake in the Huaihe River Basin and one of the five largest freshwater lakes in China. Geographically, Hongze Lake is situated between 33°06′–33°40′N and 118°10′–118°52′E. The lake is influenced by a monsoon climate, resulting in relatively abundant annual precipitation. The water quality of Hongze Lake is classified as medium-eutrophic, with the main pollutants being organic matter [31]. The average annual water temperature is 16.3 °C, with summer water temperatures exceeding 28 °C, which is conducive to the proliferation of algae. Consequently, the lake has experienced frequent algal blooms in recent years, significantly impacting local fishery resources, the ecological environment, and the water supply for nearby residents and agriculture.
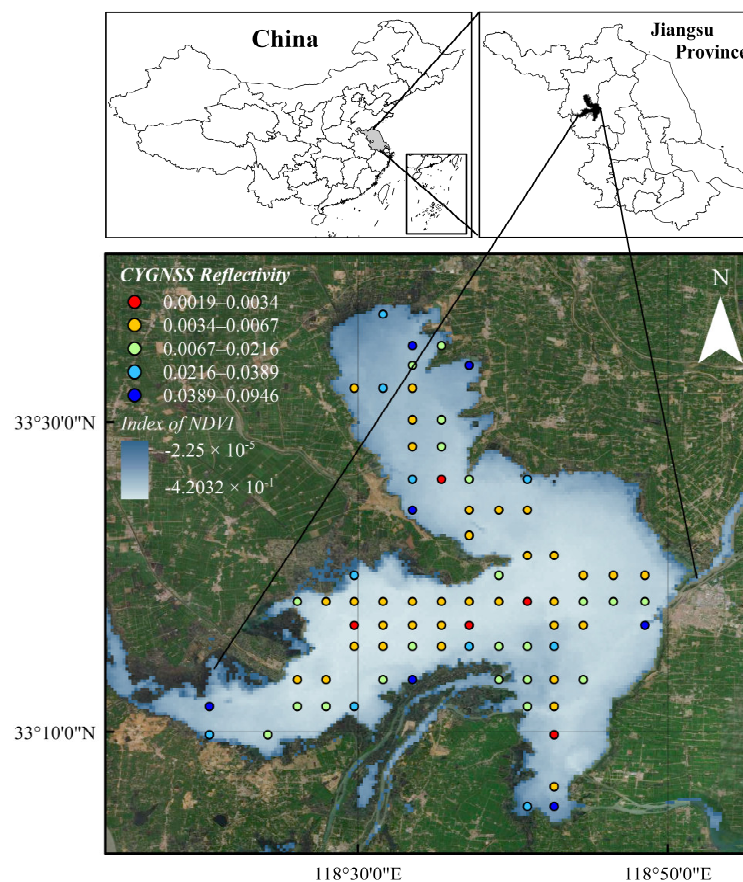


**Figure 1.** Distribution of averaged CYGNSS reflection points in the Hongze Lake.

## 2.2. CYGNSS Data

The CYGNSS satellite is a constellation utilizing GNSS-R technology that has received great attention and has achieved many research results since its operation. The mission was launched by the National Aeronautics and Space Administration (NASA) in December 2016 into a low Earth orbit at an altitude of 580 km and an inclination of 35°. Its constellation contains eight microsatellites that can receive both direct and ground-reflected signals from Global Positioning System (GPS) satellites, and its sampling area covers all regions between 38° north and south latitudes. Each satellite is equipped with a dual-base radar featuring multiple reception channels, allowing the reception of up to four signals simultaneously, resulting in observations at 32 different points. The CYGNSS constellation offers high temporal resolution and short revisit periods, averaging 7.2 h for oceans and 1–2 days for land [32]. The spatial resolution of the data varies theoretically from 0.5 to 25 km, depending on whether the reflections are specular (Fresnel zone) or diffuse (Shining zone) [33]. Since 2019, the sampling time of CYGNSS has been reduced from 1 s to 0.5 s, increasing the minimum spatial resolution to $3.5 \times 0.5$ km [34]. Although the mission was originally designed to monitor tropical cyclones, it was subsequently found to be perceptive of surface changes. In this paper, CYGNSS L1 V3.1 data (downloaded from NASA https://search.earthdata.nasa.gov/search, accessed on 1 May 2023) were acquired from May to September 2023. The surface-reflected power calculated from CYGNSS observations was used to categorize algal bloom outbreaks in the Hongze Lake region.

## 2.3. Auxiliary Data

### 2.3.1. Sentinel-3 OLCI Data

This paper utilizes multispectral data from the Sentinel-3 OLCI satellite as auxiliary data. ESA started the Sentinel-3 constellation mission in February 2016, and the Ocean and Land Color Instrument (OLCI) on board the A and B satellites included in the mission provides a wealth of water ecological remote sensing data [35]. Sentinel-3 OLCI provides a full resolution of 1200 m and a reduced resolution of 300 m. Meanwhile, its Ocean product has a revisit period of 1~2 days. OLCI inherits the technical characteristics of the previous ENVISAT Medium Resolution Imaging Spectrometer (MERIS) sensor, designed for watercolor and ecological remote sensing. It features several bands in the transition zone from red to near-infrared. As a result, Sentinel-3 OLCI data are widely used to monitor and identify the distribution area, intensity, and biomass of phytoplankton in eutrophic, optically complex inland waters and near-shore seas worldwide. In this paper, Sentinel-3 OLCI multispectral image data are used to retrieve *chl_a* concentration by leveraging the bands of red to near-infrared wavelength. These data assist in the classification of bloom levels.

### 2.3.2. ERA5-Land Data

ECMWF Re-Analysis-5 (ERA5)-Land is a meteorological dataset released by the European Center for Medium-Range Weather Forecasts (ECMWF), containing atmospheric stratification data on a global scale from 1970 to the present. It is available through the European Space Agency (ESA) Climate Database (https://cds.climate.copernicus.eu/, accessed on 1 May 2023). ERA5-Land is the latest ECMWF reanalysis dataset and represents a globally important surface element data integrated from multiple sources worldwide. This dataset provides hourly meteorological data with a spatial resolution of $0.1° \times 0.1°$. The main meteorological factor that affects the identification of blooms is wind speed. Wind speed influences the degree of water surface smoothing and the distribution of algae [36]. In this paper, wind speed data from the ERA-5 Land dataset is used to assist in monitoring the algal bloom outbreak process.

## 3. Algal Bloom Level Monitoring Method

In this study, *chl_a* concentration was first retrieved from Sentinel-3 OLCI imagery using the Maximum Peak Height (MPH) algorithm and validated with in situ chl_a concentration measurement. The extensive coverage of OLCI data not only compensates for the limited in situ sampling but also provides more data for modeling. After data matching and quality control, the *chl_a* concentration was classified into 2 to 5 categories using the K-means clustering algorithm. These categorized levels were labeled and used as references and output levels to establish and train the monitoring model. Additionally, surface reflectivity was derived using the bistatic radar equation, and wind speed data were extracted from the ERA5-Land dataset to serve as inputs for the XGBoost monitoring model. This design ensures a well-structured approach, integrating multisource data and clustering methods to build an algal bloom monitoring model. The flowchart of this study is illustrated in Figure 2.
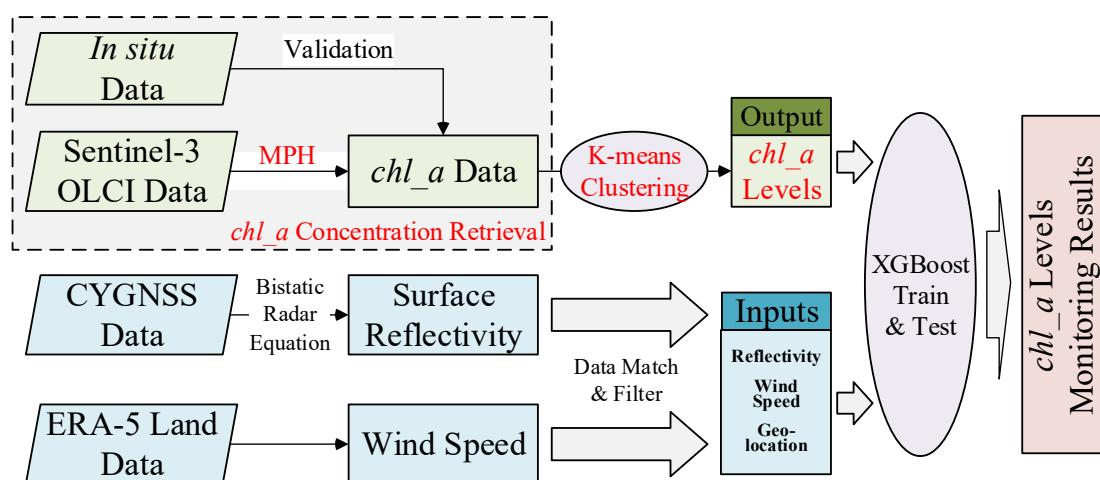


**Figure 2.** Flowchart of the study.

### 3.1. MPH Retrieval of chl_a Concentration

The MPH algorithm [37] is designed to provide quantitative estimates of *chl_a* in optically complex inland and nearshore waters and detect cyanobacterial blooms, surface-floating algae, and aquatic plants. It has undergone rigorous validation across a diverse array of inland lakes globally, covering a wide range of trophic levels and water types. The algorithm has consistently demonstrated strong overall performance and is widely applicable to most lakes, with the exception of certain oligotrophic systems. The algorithm utilizes the narrow red-edge bands at 681, 709, and 753 nm, with unique spectral features in the 665 and 885 nm bands for cyanobacteria detection. It identifies the location of the maximum peak in the 753 nm band to determine the presence of floating material and cyanobacterial blooms. Specifically, the algorithm focuses on the 753 nm maximum reflectivity peak to ascertain whether the water surface is heavily colonized with phytoplankton or macrophytes. Matthews [38] improved the algorithm by increasing the number of bands to 620, 664, 681, 709, 753, and 885 nm. During the retrieval process, the peak reflectivity ($R_{max,0}$, $R_{max,1}$) and the corresponding wavelengths ($\lambda_{R_{max,0}}$, $\lambda_{R_{max,1}}$) were determined for two different spectral ranges used for the MPH calculations. Here, $R_{max,0}$ represents the maximum reflectivity peaks at the band center wavelengths of 681 and 709, and $R_{max,1}$ represents the maximum reflectivity peaks at the band center wavelengths of 681, 709, and 753. The MPH values, $MPH_0$ and $MPH_1$, are calculated from these maximum reflectivity peaks, where $brr_i$ denotes the peak reflectivity of band with the center wavelength being *i* and the integer (e.g., $I_{619}$) denotes the band center wavelength:

$$MPH_0 = R_{max,0} - brr_{664} - \left[ (brr_{885} - brr_{664}) \cdot \left( \lambda_{R_{max,0}} - I_{664} \right) / (I_{885} - I_{664}) \right] \quad (1)$$

$$MPH_1 = R_{max,1} - brr_{664} - \left[ (brr_{885} - brr_{664}) \cdot \left( \lambda_{R_{max,1}} - I_{664} \right) / (I_{885} - I_{664}) \right] \tag{2}$$

Sun-Induced Chlorophyll Fluorescence peaks ($SICF_{peak}$), Sun-Induced Phycocyanin Absorption and Fluorescence peaks ($SIPAF_{peak}$), Normalized Difference Vegetation Index ($NDVI$), and Backscattering and Absorption-Induced Reflectivity peaks ($BAIR_{peak}$) were also computed for water body delineation, where

$$NDVI = (brr_{885} - brr_{664}) / (brr_{885} + brr_{664}) \tag{3}$$

$$SICF_{peak} = brr_{664} - brr_{619} - \left[ (brr_{681} - brr_{619}) \cdot (I_{664} - I_{619}) / (I_{681} - I_{619}) \right] \tag{4}$$

$$SIPAF_{peak} = brr_{681} - brr_{664} - \left[ (brr_{709} - brr_{664}) \cdot (I_{681} - I_{664}) / (I_{709} - I_{664}) \right] \tag{5}$$

$$BAIR_{peak} = brr_{709} - brr_{664} - \left[ (brr_{885} - brr_{664}) \cdot (I_{709} - I_{664}) / (I_{885} - I_{664}) \right] \tag{6}$$

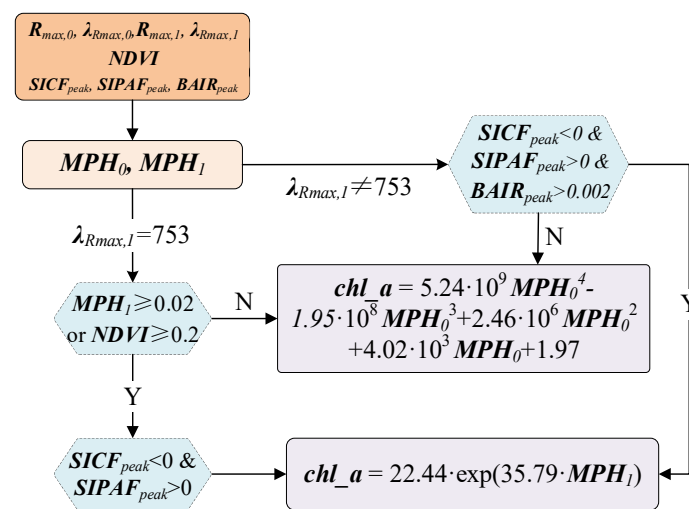The specific implementation flow of the MPH algorithm is shown in Figure 3:



**Figure 3.** Flowchart of MPH to obtain *chl_a* concentration.

### 3.2. Classification Method for Algal Blooms

Algal blooms are defined by the proliferation and aggregation of various algal species, ultimately leading to large-scale lake disasters. The *chl_a* concentration was classified accordingly for the different stages of the algal bloom outbreak. The "Technical Specification for Classification and Monitoring of Algal Blooms" (DB44/T 2261-2020) [39] is a local standard of Guangdong Province, China, which was implemented on 28 March 2021 and issued by the Guangdong Provincial Department of Ecology and Environment. The document details the classification methods for common algal blooms, such as diatoms, greens, and methanogens, through the density of algae, as well as the concentration of *chl_a*, as shown in Table 1.

**Table 1.** Guangdong Province Algal Bloom Classification Standard.

| Algal Bloom Classification | Cyanobacterial Density (Cells/L) | *chl_a* Concentration (μg/L) |
|---|---|---|
| no algal bloom | $0 < D < 2 \times 10^6$ | $C < 10$ |
| no visible algal bloom | $2 \times 10^6 < D < 1 \times 10^7$ | $10 < C < 15$ |
| mild algal bloom | $1 \times 10^7 < D < 5 \times 10^7$ | $15 < C < 50$ |
| moderate algal bloom | $5 \times 10^7 < D < 1 \times 10^8$ | $50 < C < 100$ |
| heavy algal bloom | $D > 1 \times 10^8$ | $C > 100$ |

Each area has unique factors such as water temperature, nutrient levels, and ecosystem characteristics. Therefore, algal bloom classification standards applicable to one area may not be suitable for another. Based on this, this paper adopts a data-driven classification criterion for *chl_a* concentration retrieved by the MPH algorithm. To standardize the data in the study area, a clustering method is employed. K-means, a commonly used unsupervised learning algorithm, classifies a dataset into *k* clusters [40].

### 3.3. Calculation of CYGNSS Surface Reflectivity

GNSS-R technology relies on bistatic radar to obtain surface reflection signals, and its received signals are usually described by a bistatic radar model [41]:

$$P^r = P_{RL}^{coh} + P_{RL}^{inc} \tag{7}$$

In this equation, the received signal ($P^r$) from CYGNSS consists of coherent signal ($P_{RL}^{coh}$) and incoherent signal ($P_{RL}^{inc}$). The coherent signal dominates the received signal when the surface roughness is lower and smoother, and the main source of the coherent signal is the first Fresnel reflection region near the specular reflection point. As the distance from the specular reflection point increases, the coherent component proportion of the signal decreases rapidly. The coherent component of the reflected signal power can be expressed as

$$P_{RL}^{coh} = \left(\frac{\lambda}{4\pi}\right)^2 \frac{P_t G_t G_r}{(R_r + R_t)^2} \Gamma_{RL}(\theta) \tag{8}$$

where $\lambda$ is the wavelength, $P_t$ is the peak power of the transmitted GNSS signal, $G_t$ is the gain of the transmitting antenna, and $G_r$ is the gain of the receiving antenna. $R_r$ is the distance between the specular reflection point and the GNSS-R receiver, $R_t$ is the distance between the specular reflection point and the GNSS transmitter, $\Gamma_{RL}(\theta)$ is the specular reflectivity at the specular reflection point.

Meanwhile, the calculation of the incoherent component of the reflected signal power can be expressed as

$$P_{RL}^{inc} = \frac{\lambda^2 P_t G_t G_r R_{PL}}{(4\pi)^3} \sigma_{RL} \tag{9}$$

where $\sigma_{RL}$ is the bistatic radar cross-section in m$^2$, and $R_{PL}$ is the Fresnel coefficient. When the surface is relatively flat and smooth, the signal can be considered to be mainly a coherent component, i.e., $P_{RL}^{coh} = P_{RL}^{inc}$, and then the surface reflectivity $\Gamma_{RL}(\theta)$ can be expressed as [42–46]

$$\Gamma_{RL}(\theta) = \frac{\sigma_{RL}(R_r + R_t)^2}{4\pi R_t^2 R_r^2} \tag{10}$$

### 3.4. XGBoost Algorithm and Accuracy Evaluation

Extreme Gradient Boosting (XGBoost) is an efficient and powerful gradient boosting framework widely used for various ML tasks, including multiclass classification problems [47]. The core idea of XGBoost is to iteratively train a series of weak learners, typically decision trees, and progressively add new learners to enhance the accuracy of the model. Residuals from the previous model are fitted by training a new decision tree, with the predictions of this new tree serving as improvements to the current model. The impact of each tree on the overall model is managed through the learning rate. XGBoost optimizes model performance by refining the objective function and employs regularization to control model complexity, thereby mitigating overfitting. The final XGBoost classification model is obtained by iterating until the training ends when the error on the validation set no longer decreases.

## 4. Results and Analysis

### 4.1. chl_a Retrieval Results and Validation

In this study, Sentinel-3 OLCI imagery from May to September 2023 was used to retrieve *chl_a* concentration within the Hongze Lake region using the MPH algorithm. Due

to limitations imposed by weather conditions and revisit cycles, reliable OLCI images were available for 22 days during this period. To validate the reliability of the *chl_a* concentration retrieval, in situ data were collected in the Hongze Lake region on 11 May 2023, and were used for comparison with the retrieval results. The *chl_a* concentration results for one day each month are selected and presented here (see Figure 4).
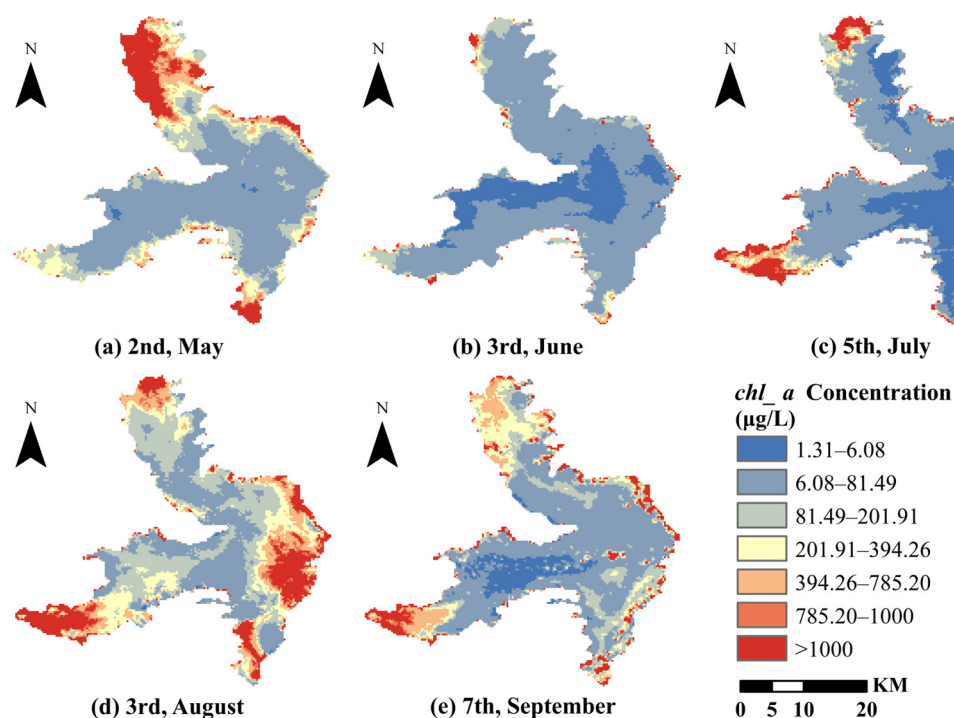


**Figure 4.** Results of *chl_a* concentration retrieval based on MPH algorithm.

In most parts of Hongze Lake, *chl_a* concentration was low, with levels near the lake's center dropping below 100 µg/L, indicating no algal blooms. In contrast, higher *chl_a* concentration was observed near the shores, particularly along the western and northern edges of the lake, where levels exceeded 100 µg/L. These areas, characterized by relatively shallow water and intensive planting and aquaculture activities, experience significant eutrophication due to fertilizer and feed pollution. Consequently, they are more prone to algal bloom outbreaks.

Since the Sentinel-3 constellation did not capture valid OLCI images on 11 May 2023, this study validated the *chl_a* concentration retrieval results using data from 9 and 14 May, which were close to that date. Figure 5 shows the distribution between retrieval results and in situ data for *chl_a* concentration. The base map shows the retrieval results of concentration, and the dots show the in situ data. The visualization offers a clear comparison between the retrieval model's *chl_a* distribution and real-world data, providing valuable insights into its performance. By highlighting both the alignment and discrepancies between the modeled retrieval values and in situ measurements, it sheds light on the model's accuracy and reliability. The results demonstrate that the retrieval model captures the spatial distribution of *chl_a* in aquatic environments with notable precision, indicating strong predictive capabilities and suitability for real-world applications.
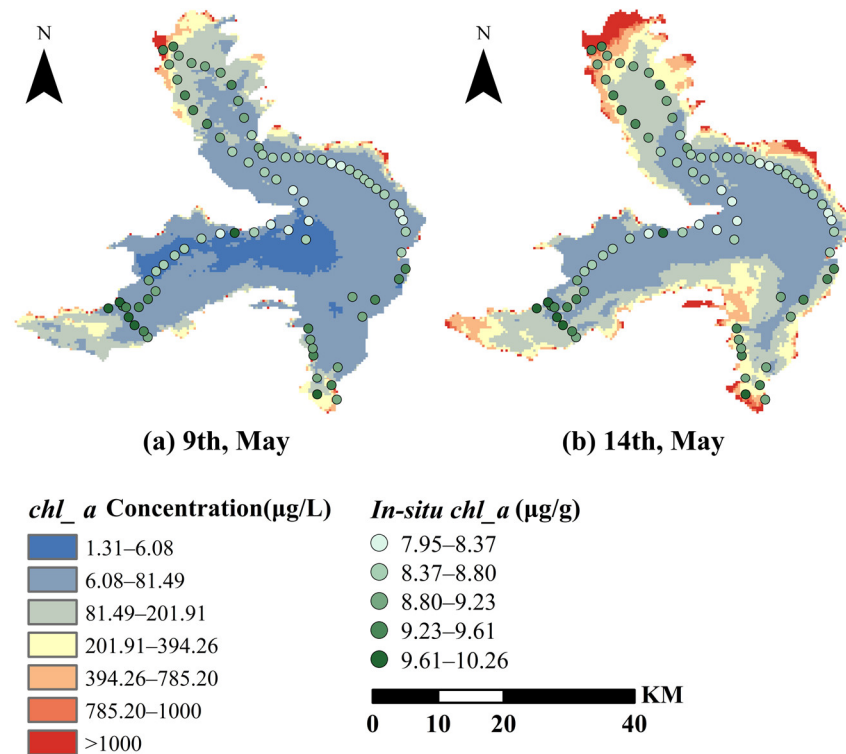
**(a) 9th, May**  **(b) 14th, May**

**Figure 5.** The map of retrieved *chl_a* concentration results and in situ measurements.

Figure 6 displays the correlation between the *chl_a* concentration retrievals for 9 and 14 May and the in situ measurements from 11 May. Although the units are inconsistent, there is a strong correlation between the two datasets, which significantly reflects and verifies the reliability of the retrieval results. The retrieval results of these two days were compared with the in situ data, respectively, and both of them reflected a relatively apparent correlation with coefficients of $R = 0.7593$ for 9 May and $R = 0.7933$ for 14 May. Meanwhile, the RMSEs on the 9th and 14th were 56.081 ug/L and 86.865 ug/L, respectively.
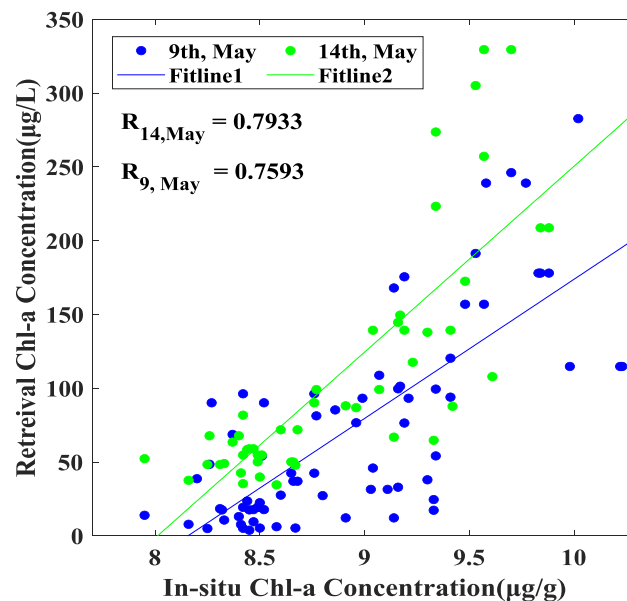


**Figure 6.** Relationship between retrieval results of *chl_a* concentration on 9 May and 14 May and measured *chl_a* concentration on May 11.

The comparison results indicate a high degree of alignment between the modeled retrieval values and in situ measurements, confirming the accuracy and reliability of the retrieval process. These results lay a solid foundation for the subsequent classification of the retrieval values and their integration into monitoring models.

In this study, the retrieved *chl_a* concentration is classified into different categories using the K-means algorithm to establish an algal bloom monitoring model. This classification method helps to accurately define different stages of algal bloom development, enhancing the predictive accuracy of the model. Furthermore, by comparing the *chl_a* concentrations across different categories, it becomes more effective to identify and monitor potential algal bloom outbreak areas in the water, providing crucial support for environmental monitoring and water quality management.

### 4.2. Algal Bloom Level Monitoring Based on CYGNSS Data

In this study, an input dataset was constructed using CYGNSS data, ERA5-Land data, and Sentinel-3 OLCI image data from the experimental area. The input variables include the geographic locations of sampling points, reflectivity, and wind speed. The output is based on the classification results of previously retrieved *chl_a* concentration. The retrieved *chl_a* concentration was clustered and classified using the K-means method, and the XGBoost model was employed to predict *chl_a* concentration levels through geolocation, CYGNSS reflectivity, and wind speeds.

The CYGNSS sampling process resulted in a random discrete distribution of ground reflect points across the study area, and the spatial resolution of the *chl_a* concentration retrieval results was not consistent with the CYGNSS reflectivity data. Therefore, the CYGNSS data outside the water body were removed, and the *chl_a* concentration retrieval results were resampled to 1 km with the CYGNSS reflectivity data in this study. When *chl_a* concentration is at a very high level, algae significantly accumulate on the water surface, directly affecting CYGNSS reflectivity. Thus, quality control procedures were applied to exclude anomalous data. Specifically, data with CYGNSS reflectivity of less than 0.02 were removed when *chl_a* concentration exceeded 190 μg/L. The number of samples used in the learning model after data quality control is shown in Table 2.

**Table 2.** The number of samples used in the learning model.

| Date | Amount of Data Pre-Filter | Amount of Data After Filter |
|:---:|:---:|:---:|
| 2 May 2023 | 18 | 17 |
| 3 May 2023 | 19 | 14 |
| 9 May 2023 | 21 | 11 |
| 13 May 2023 | 23 | 6 |
| 14 May 2023 | 35 | 12 |
| 16 May 2023 | 24 | 10 |
| 19 May 2023 | 29 | 15 |
| 20 May 2023 | 32 | 14 |
| 3 Jun 2023 | 31 | 21 |
| 7 Jun 2023 | 22 | 8 |
| 8 Jun 2023 | 21 | 5 |
| 9 Jun 2023 | 11 | 8 |
| 10 Jun 2023 | 35 | 22 |
| 14 Jun 2023 | 16 | 7 |
| 3 Aug 2023 | 9 | 8 |
| 12 Aug 2023 | 25 | 13 |
| 31 Aug 2023 | 13 | 4 |
| 1 Sep 2023 | 36 | 21 |
| 7 Sep 2023 | 41 | 30 |
| 8 Sep 2023 | 32 | 17 |
| 9 Sep 2023 | 25 | 10 |
| 10 Sep 2023 | 19 | 6 |

Figure 7 illustrates the *chl_a* concentration values corresponding to CYGNSS reflection points, which represent the total number of samples used in our monitoring model. Most of the reflection points have *chl_a* concentration under 100 μg/L, with a small number of data points showing *chl_a* concentration over 1000 μg/L. Table 3 shows the K-means clustering results of different categories, in which the algal bloom levels are categorized from "No Bloom" to "Heavy", and the numbers 0–5 represent the corresponding *chl_a* concentration ranges under different clustering results.
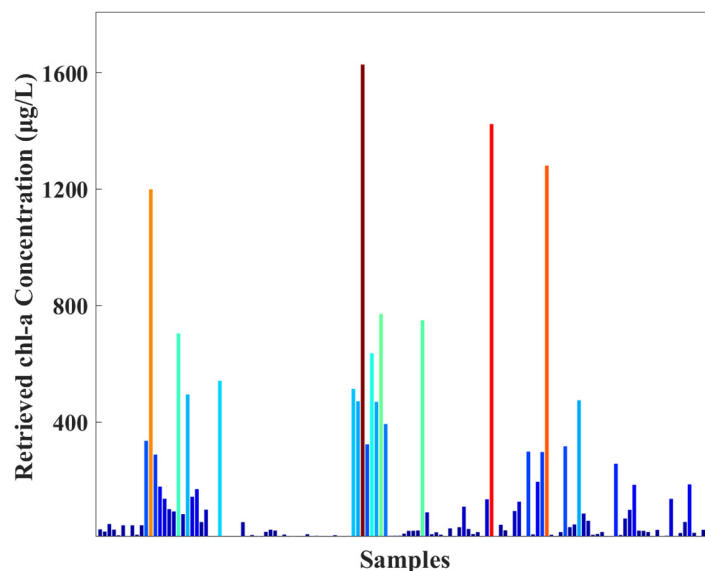


**Figure 7.** *chl_a* concentration values corresponding to CYGNSS reflection points, the colors (blue to red) represent increasing concentration.

**Table 3.** Classification Results of K-means Algorithm.

| Algal Bloom Classification | *chl_a* Concentration (μg/L) | 2Class | 3Class | 4Class | 5Class |
|:---:|:---:|:---:|:---:|:---:|:---:|
| No Bloom | <200 | 0 | 0 | 0 | 0 |
| Light | 200–600 | 0 | 1 | 1 | 1 |
| Mild | 600–1000 | | 1 | 2 | 2 |
| Moderate | 1000–1500 | 1 | 2 | 3 | 3 |
| Heavy | >1500 | 1 | 2 | 3 | 4 |

To evaluate the accuracy of the monitoring model's classification results for algal blooms, a five-fold cross-validation (CV) method was used. It effectively utilizes every part of the dataset, minimizing the randomness associated with data splitting. This is particularly beneficial when the dataset is small, as it maximizes the use of training samples, thereby enhancing the model's robustness and stability.

This study uses the confusion matrix and the corresponding *Accuracy* to evaluate the performance of the model. *Accuracy* is the ratio of the number of correctly classified samples to the total number of samples, and $C_{ij}$ represents that the $i$th class of data is categorized into the $j$th class.

$$Accuracy = \frac{\sum_{i=1}^{n} C_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij}} \tag{11}$$

The accuracy results of predicting *chl_a* concentration classes from CYGNSS reflectivity data and wind speed data at a spatial resolution of 1 km are shown in Figure 8. "2 Classes" to "5 Classes" represent the categorization results of the clustering, while "GD Classes" represents the categorization according to the Guangdong Province standard. The overall accuracy of the model was highest when the data were clustered into "3 Classes",

with an overall accuracy of 0.955. The classification accuracy gradually decreased as the number of classification classes increased. Notably, the classification result obtained after clustering is lower than that obtained using the Guangdong Province standard, which had an accuracy of 0.698. This discrepancy may be attributed to the fact that the clustering mainly subdivided data points with *chl_a* concentration less than 100 µg/L, a range in which the *chl_a* concentration characteristics were not distinctly categorized.
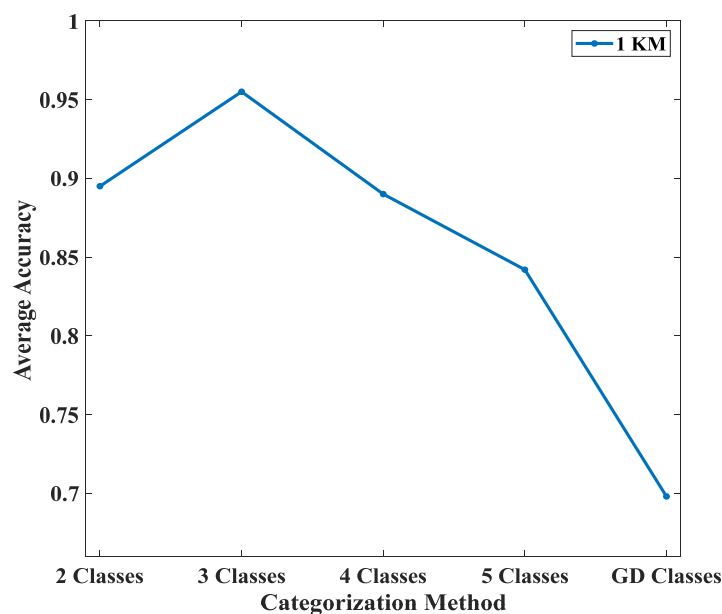


**Figure 8.** Accuracy of predicted *chl_a* concentration category by XGBoost at 1 KM resolution.

Additionally, the confusion matrix demonstrates the classification results by showing how correctly and incorrectly samples from each class *i* are predicted to be in class *j*. Diagonal elements represent correct classifications, while off-diagonal elements indicate misclassifications, helping to identify common errors and confusion between categories. Figures 9–11 illustrate the confusion matrices of model predictions obtained by different classification methods at a 1 km spatial resolution. Categories 0 to 4 refer to different levels of algal blooms, corresponding to Table 3. Figure 9 demonstrates the results of categorizing the retrieval values of *chl_a* concentration into Classes 2 and 3. The accuracy, average precision, average recall, and average F1 index obtained by the model under this categorization method are maintained at a high level, with fewer misclassifications occurring. However, in the binary categorization, the data representing category 0, which has a lower *chl_a* concentration, actually shows a relatively high concentration. The sample imbalance in the dataset makes the model less sensitive to the distinguishing features between the two categories. This is one of the reasons why there is more misclassification in category 1, as shown in Figure 9a. But, Figure 10 shows the results of categorizing the retrieval values of *chl_a* concentration into categories 4 and 5. As the number of classification categories increases, the misclassification phenomenon begins to increase, and the classification accuracy of the model appears to be reduced to a certain extent. However, the overall classification accuracy remains around 0.85, which is within an acceptable range. Figure 11 demonstrates the results of classification according to the local standards of Guangdong Province. When classified according to the local standards, the differences in features between different categories are not distinct enough, resulting in more cases of misclassification.
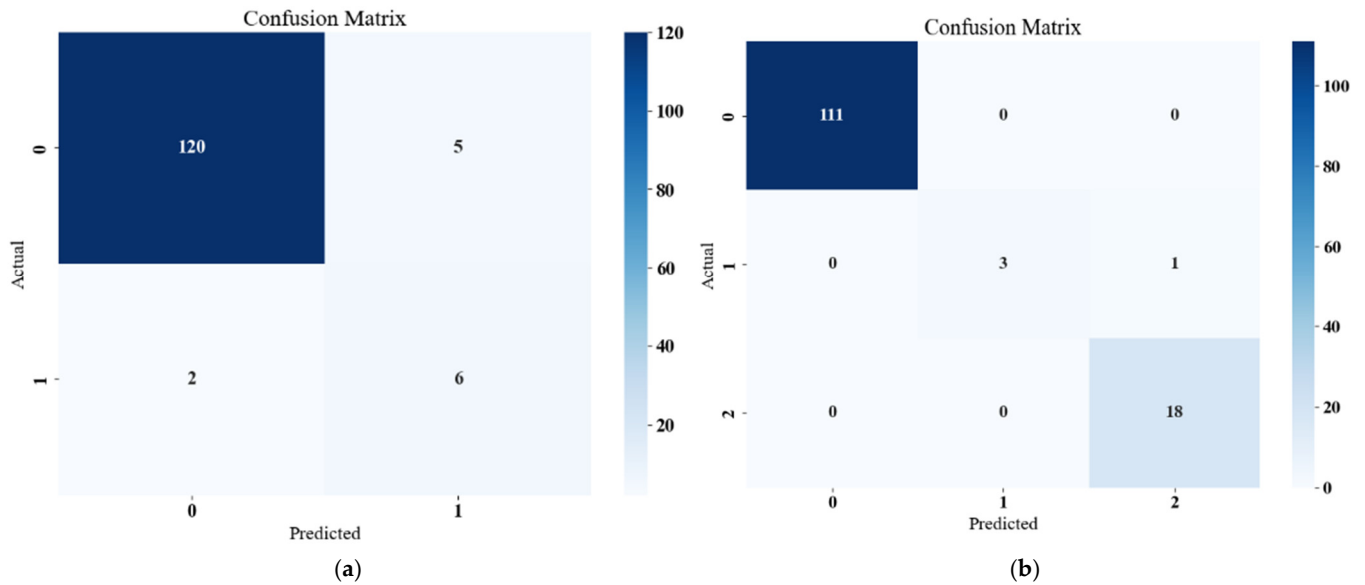
**Figure 9.** Model classification confusion matrix for 2 Classes (**a**) and 3 Classes (**b**) classification criterion.
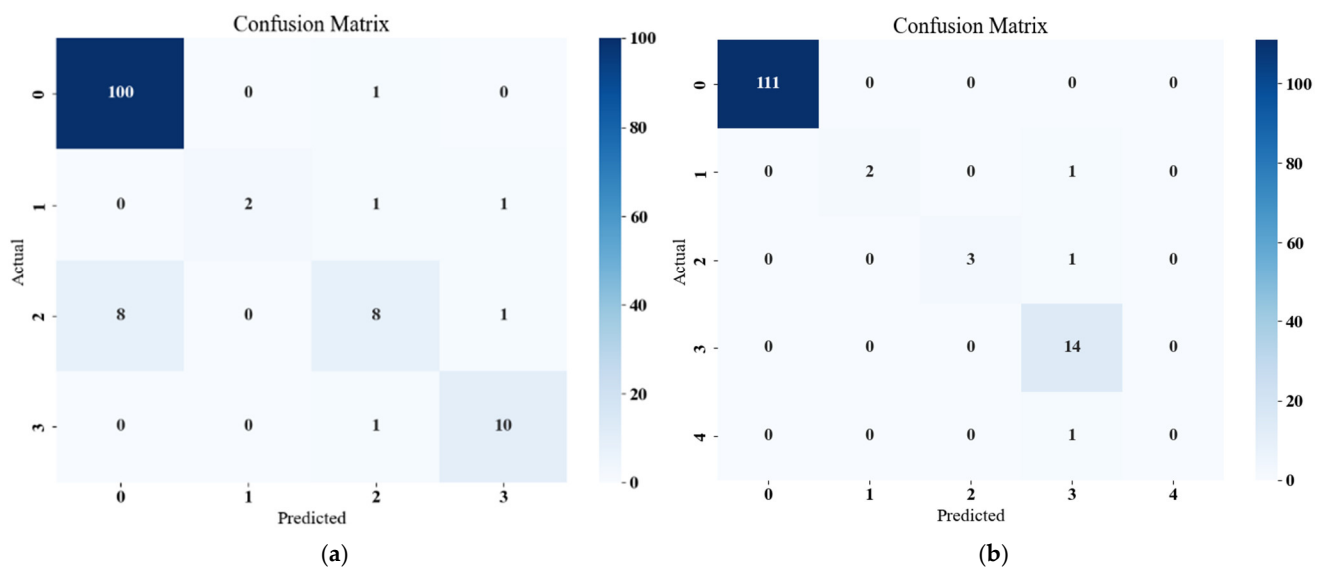


**Figure 10.** Model classification confusion matrix for 4 Classes (**a**) and 5 Classes (**b**) classification criterion.

### 4.3. Impact of Different Scales on the Effectiveness of Algal Bloom Monitoring Models

The spatial resolution of the CYGNSS data was greater than the resolution of the *chl_a* concentration retrieval results, so the *chl_a* concentration was projected onto a grid of the same resolution as CYGNSS and averaged. During the averaging process, the *chl_a* concentration may not be represented accurately due to the large differences between pixels within the grid. Therefore, in this study, the retrieved *chl_a* concentrations were also projected onto 2 km and 3 km grids to explore the impact of classification prediction accuracy of *chl_a* concentration at different resampling resolutions.
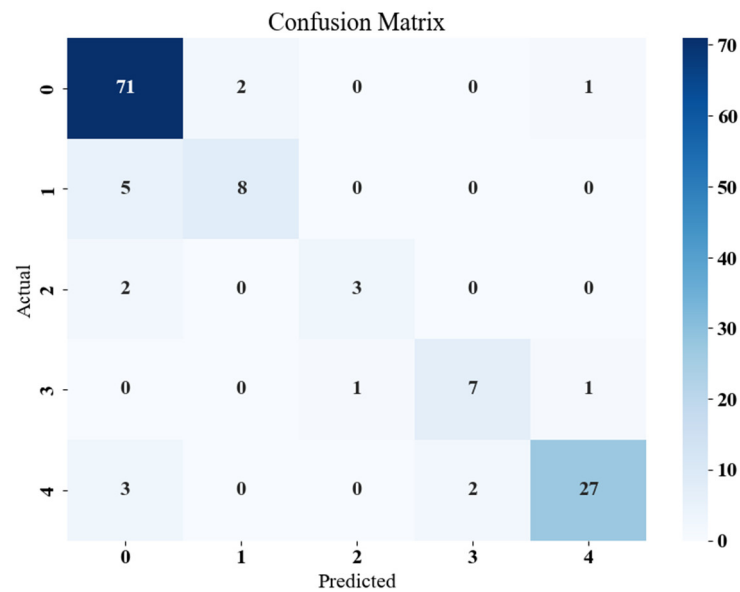
**Figure 11.** Model classification confusion matrix of Guangdong local classification criterion.

The five-fold validation accuracies of the classification results at different spatial resolutions are shown in Table 4, as well as Figure 12. When the spatial resolution was reduced from 1 km to 2 km or 3 km, the classification accuracy of the model subsequently decreased. The accuracy of the classification results at a 1 km spatial resolution under different classification methods was generally better than at other resolutions. This indicates that higher spatial resolution can obtain higher classification accuracy to some extent.

**Table 4.** Five-fold CV accuracy of various classification methods across different spatial resolutions.

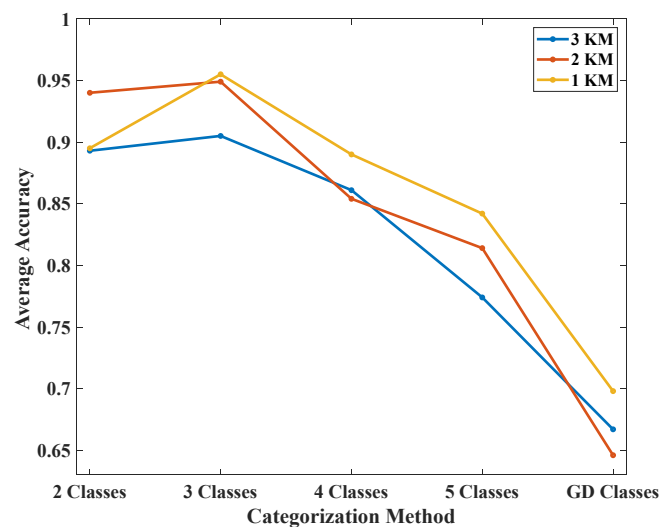| Classification | Average Accuracy (AA) | | |
| --- | --- | --- | --- |
| | 3 KM | 2 KM | 1 KM |
| 2 Classes | 0.893 | 0.940 | 0.895 |
| 3 Classes | 0.905 | 0.949 | 0.955 |
| 4 Classes | 0.861 | 0.854 | 0.890 |
| 5 Classes | 0.774 | 0.814 | 0.842 |
| GD Classes | 0.667 | 0.646 | 0.698 |



**Figure 12.** Accuracy of 5-fold CV of different classification methods at different spatial resolutions.

## 5. Discussion

In this study, *chl_a* concentration is used to determine the level of algal blooms. *chl_a* is a general indicator for phytoplankton and a specific marker for cyanobacteria. Most current remote sensing methods, including the one used in this study, cannot accurately distinguish between "green" algae and "blue–green" algae (cyanobacteria). The term "green algae" refers to a broad group encompassing many different types of phytoplankton, whereas cyanobacterial blooms, especially toxic ones, are relatively rare compared with green algal blooms [9,10]. Although cyanobacteria and phytoplankton blooms differ in nature, they share similar reflectance characteristics in the visible and near-infrared (VIS-NIR) spectral regions, making them hard to distinguish [9,10]. This study does not attempt to differentiate phytoplankton species but instead detects algal blooms by utilizing the characteristic of biomass correlation as *chl_a* smooths the water surface. As a result, it focuses on detecting "algal blooms" without differentiating between green algae and cyanobacteria, or distinguishing between plant and bacterial life.

CYGNSS (Cyclone Global Navigation Satellite System), by measuring the intensity and phase of reflected GPS signals, can capture water surface roughness [18–20]. Since changes in algal blooms are often related to wind speed, ERA5-Land wind speed data are integrated into the model for better results [26,36]. CYGNSS, consisting of a constellation of eight small satellites, provides a high temporal resolution by frequently scanning Earth's surface [14]. This is a significant advantage over traditional passive radar satellites, which have lower temporal resolution. By leveraging the high temporal resolution and CYGNSS's sensitivity to surface roughness, combined with wind speed data, a machine learning model is constructed to explore CYGNSS's potential for early warning of harmful algal blooms (HABs).

However, this study is limited by the availability of in situ *chl_a* concentration data, which are insufficient to validate the model at a higher temporal and spatial resolution. Therefore, this study first utilizes Sentinel-3 OLCI imagery and the MPH algorithm to estimate *chl_a* concentration in lakes, validating the retrieved values. These verified *chl_a* concentrations are then used as the ground truth to train and test the algal bloom monitoring model, ultimately yielding a graded early-warning system for HABs. In addition, given the limitations of the CYGNSS sampling process, it was impossible to sample all data ranges within the lakes in a short period of time. This is also a primary reason for utilizing *chl_a* concentration values obtained from OCLI satellite data for modeling purposes, rather than relying solely on field measurements. Future studies can consider combining data from multiple lakes and extending the time series. More in situ measurements collected in different seasons should be used for robust validation of the model.

Two separate legends are used to better distinguish between the retrieval and field results. These legends have different classification criteria and ranges, which primarily resulted in the retrieved *chl_a* concentration values not agreeing with the field data. Despite using different symbols, we can observe that the overall trends between the retrieval and field results are largely consistent, showing higher values near the lake bank and lower values in the center. The scatter plot also demonstrates a strong correlation between the two datasets. Additionally, the classification and early-warning results in waters near the lake bank are still reliable and accurate. For example, as shown in Figure 11, in regions where algal blooms are more severe (higher bloom values), the classification for level "4" has 27 correct predictions and only two misclassifications, demonstrating high accuracy.

The performance of the algal blooms monitoring model, based on CYGNSS data, is evaluated using a confusion matrix. The confusion matrix is a tool used to assess the performance of classification algorithms, particularly for multiclass problems [13]. It compares the model's predictions with actual labels, providing a breakdown of correct and incorrect classifications. Each element $C_{ij}$ in the matrix represents the number of samples predicted to be class *j* when the actual label is class *i*. A sufficient sample size helps the model learn the true distribution of the data, thereby improving prediction accuracy and robustness. In this multiclass scenario, most data are categorized as "No Bloom" or "Mild",

resulting in relatively stable and accurate prediction accuracy for the "No Bloom" class compared with other categories (over 90%). This is due to data imbalance, which could be addressed in future research by adjusting the distribution of bloom and no-bloom samples to further explore the model's ability to detect algal blooms [48].

Increasing the number of classification categories does not always lead to better prediction performance. The effectiveness of classification depends on the model's complexity, data quality, feature distinctiveness, and generalization ability [18–23]. More categories introduce greater complexity, potentially resulting in higher computational costs and issues like overfitting or underfitting [29]. If the differences between categories are subtle, increasing categories may reduce the model's ability to differentiate between similar classes, decreasing accuracy. In cases of insufficient training data or imbalanced class distributions, adding categories could worsen performance for certain classes due to poor generalization. Therefore, adding more categories does not guarantee improved results; effectiveness must be evaluated based on the specific dataset and model. Additionally, false positives can occur across all classes, not just adjacent ones. In imbalanced datasets, increasing categories might lead to poorer performance in predicting some categories, and the occurrence of false positives could become more random [48].

This study uses OLCI satellite data and the MPH algorithm to derive *chl_a* concentration. Although the MPH algorithm was initially developed based on MERIS satellite data, the OLCI satellite inherits the technology of MERIS, and the central wavelengths of its bands are consistent with those of MERIS (see Table 5), making the OLCI data well suited for the algorithm [49]. Furthermore, although the MPH algorithm was originally developed for specific regions, the results of this study demonstrate that the algorithm performs reliably in new areas as well, proving its adaptability and showing potential for broader application in other regions.

**Table 5.** Corresponding bands between MERIS and OLCI Center.

| Wavelength of the Band | MERIS Bands | OLCI Bands |
| --- | --- | --- |
| 619 | 6 | Oa7 |
| 664 | 7 | Oa8 |
| 681 | 8 | Oa10 |
| 709 | 9 | Oa11 |
| 753 | 10 | Oa12 |
| 885 | 14 | Oa18 |

CYGNSS's high temporal resolution provides a significant advantage to this study. Despite its lower spatial resolution, its frequent revisit times compensate for this limitation, particularly in conditions with cloud cover or adverse weather, where CYGNSS can still provide consistent observations. This high temporal resolution is crucial for capturing rapidly evolving water surface phenomena, such as the early stages of algal blooms. Compared with other satellites that offer high spatial resolution but lower revisit frequencies, CYGNSS provides a more continuous time series, enabling real-time monitoring and response to dynamic changes. While there is a trade-off in detection precision, CYGNSS's high-frequency observations exhibit significant potential for application in rapidly changing environments.

## 6. Conclusions

In this study, an integrated approach is proposed to monitor algal blooms in Hongze Lake using CYGNSS data, Sentinel-3 OLCI data, and ERA5-Land meteorological data. By combining multiple data sources with learning techniques, reliable results were obtained. Sentinel-3 OLCI data offers high-resolution watercolor information, essential for accurately estimating *chl_a* concentration when processed with the MPH algorithm. CYGNSS data demonstrates great potential in identifying and classifying algal blooms, with GNSS-R technology providing stable and reliable data for monitoring water surface changes due to its high temporal resolution and the ability to penetrate clouds.

The Extreme Gradient Boost (XGBoost) model was successful in predicting *chl_a* concentration and classifying algal bloom levels using CYGNSS reflectivity and ancillary meteorological data. During this study, K-means clustering was employed to classify *chl_a* concentration, with a five-fold CV performed to assess model performance. The results show significant improvement in classification accuracy as the spatial resolution increased from 3 km to 2 km and 1 km. Specifically, the three-class clustering model at 1 km resolution achieved the highest overall accuracy of 0.955, while the standard classification results in Guangdong Province have a relatively low accuracy of 0.698. This indicates that the spatial resolution of the data and the classification standard significantly affect the results.

The research findings in this paper provide new insights and methods for algal bloom monitoring, which are of great significance for the protection of lake ecosystems. By further optimizing the model and improving data resolution, the accuracy of algal bloom classification can be enhanced, enabling more effective responses to algal bloom problems and better protection of the ecological environment.

## References

1.  Xie, R.; Pang, Y.; Bao, K. Spatiotemporal Distribution of Water Environmental Capacity—A Case Study on the Western Areas of Taihu Lake in Jiangsu Province, China. *Environ. Sci. Pollut. Res.* **2014**, *21*, 5465–5473. [CrossRef] [PubMed]
2.  Cheng, X.; Li, S. An analysis on the Evolvement Processes of Lake Eutrophication and Their Characteristics of the Typical Lakes in the Middle and Lower Reaches of Yangtze River. *Chin. Sci. Bull.* **2006**, *51*, 1603–1613. [CrossRef]
3.  Zou, H.; Pan, G.; Chen, H.; Yuan, X. Removal of cyanobacterial blooms in Taihu Lake using local soils. II. Effective removal of Microcystis aeruginosa using local soils and sediments modified by chitosan. *Environ. Pollut.* **2006**, *141*, 201–205. [CrossRef]
4.  Wu, J.; Xu, Q.; Gao, G.; Shen, J. Evaluating Genotoxicity Associated with Microcystin-LR and Its Risk to Source Water Safety in Meiliang Bay, Taihu Lake. *Environ. Toxicol.* **2006**, *21*, 250–255. [CrossRef]
5.  Hu, C.; Lee, Z.; Ma, R.; Yu, L.; Li, D.; Shang, S. Moderate Resolution Imaging Spectroradiometer (MODIS) Observations of Cyanobacteria Blooms in Taihu Lake, China. *J. Geophys. Res. Ocean.* **2010**, *C4*, 115. [CrossRef]
6.  Zhou, B.; Cai, X.; Wang, S.; Yang, X. Analysis of the Causes of Cyanobacteria Bloom: A Review. *J. Resour. Ecol.* **2020**, *11*, 405–413.
7.  Zhang, T.; Hu, H.; Ma, X.; Zhang, Y. Long-term Spatiotemporal Variation and Environmental Driving Forces Analyses of Algal Blooms in Taihu Lake based on Multi-source Satellite and Land Observations. *Water* **2020**, *12*, 1035. [CrossRef]
8.  Dhillon, M.S.; Dahms, T.; Kübert-Flock, C.; Steffan-Dewenter, I.; Zhang, J.; Ullmann, T. Spatiotemporal Fusion Modelling Using STARFM: Examples of Landsat 8 and Sentinel-2 NDVI in Bavaria. *Remote Sens.* **2022**, *14*, 677. [CrossRef]
9.  Wang, L.; Xu, X.; Yu, Y.; Gui, R.; Xu, Z.; Pu, F. SAR-to-optical Image Translation Using Supervised Cycle-consistent Adversarial Networks. *IEEE Access* **2019**, *7*, 129136–129149. [CrossRef]
10. Papale, D.; Belli, C.; Gioli, B.; Miglietta, E.; Ronchi, C.; Vaccari, F.P.; Valentini, R. ASPIS, A Flexible Multispectral System for Airborne Remote Sensing Environmental Applications. *Sensors* **2008**, *8*, 3240–3256. [CrossRef]
11. Wang, G.; Li, J.; Zhang, B.; Shen, Q.; Zhang, F. Monitoring Cyanobacteria-dominant Algal Blooms in Eutrophicated Taihu Lake in China with Synthetic Aperture Radar Images. *Chin. J. Oceanol. Limnol.* **2015**, *33*, 139–148. [CrossRef]
12. Bresciani, M.; Adamo, M.; Carolis, G.D.; Matta, E.; Pasquariello, G.; Vaičiūtė, D.; Giardino, C. Monitoring Blooms and Surface Accumulation of Cyanobacteria in the Curonian Lagoon by Combining MERIS and ASAR Data. *Remote Sens. Environ.* **2014**, *146*, 124–135. [CrossRef]
13. Landuyt, L.; Van Wesemael, A.; Schumann, G.J.P.; Hostache, R.; Verhoest, N.E.; Van Coillie, F.M. Flood Mapping based on Synthetic Aperture Radar: An Assessment of Established Approaches. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 722–739. [CrossRef]

14. Chew, C.C.; Small, E.E. Soil Moisture Sensing Using Spaceborne GNSS Reflections: Comparison of CYGNSS Reflectivity to SMAP Soil Moisture. *Geophys. Res. Lett.* **2018**, *45*, 4049–4057. [CrossRef]

15. Njoku, E.G.; Entekhabi, D. Passive Microwave Remote Sensing of Soil Moisture. *J. Hydrol.* **1996**, *184*, 101–129. [CrossRef]

16. Darrozes, J.; Roussel, N.; Zribi, M. The Reflected Global Navigation Satellite System (GNSS-R): From Theory to Practice. In *Microwave Remote Sensing of Land Surface*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 303–355.

17. Li, L.; Elhajj, M.; Feng, Y.; Ochieng, W.Y. Machine learning based GNSS signal classification and weighting scheme design in the built environment: A comparative experiment. *Satell. Navig.* **2023**, *4*, 12. [CrossRef]

18. Li, W.; Cardellach, E.; Fabra, F.; Rius, A.; Ribó, S.; Martín-Neira, M. First spaceborne phase altimetry over sea ice using TechDemoSat-1 GNSS-R signals. *Geophys. Res. Lett.* **2017**, *44*, 8369–8376. [CrossRef]

19. Xie, Y.; Yan, Q. Stand-alone retrieval of sea ice thickness from FY-3E GNOS-R data. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2000305. [CrossRef]

20. Kucwaj, J.C.; Reboul, S.; Stienne, G.; Choquel, J.B.; Benjelloun, M. Circular regression applied to GNSS-R phase altimetry. *Remote Sens.* **2017**, *9*, 651. [CrossRef]

21. Ghasemigoudarzi, P.; Huang, W.; Silva, O.D.; Yan, Q.; Power, D.T. Flash flood detection from CYGNSS data using the RUSBoost-algorithm. *IEEE Access* **2020**, *8*, 171864–171881. [CrossRef]

22. Wei, H.; Yu, T.; Tu, J.; Ke, F. Detection and Evaluation of Flood Inundation Using CYGNSS Data during Extreme Precipitation in 2022 in Guangdong Province, China. *Remote Sens.* **2023**, *15*, 297. [CrossRef]

23. Yan, Q.; Liu, S.; Chen, T.; Jin, S.; Xie, T.; Huang, W. Mapping Surface Water Fraction Over the Pan-tropical Region Using CyGNSS Data. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5800914. [CrossRef]

24. Chen, Y.; Yan, Q. Unlocking the potential of CYGNSS for pan-tropical inland water mapping through multi-source data and transformer. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *133*, 104122. [CrossRef]

25. Jin, S.; Camps, A.; Jia, Y.; Wang, F.; Martin-Neira, M.; Huang, F.; Yan, Q.; Zhang, S.; Li, Z.; Edokossi, K.; et al. Remote sensing and its applications using GNSS reflected signals: Advances and prospects. *Satell. Navig.* **2024**, *5*, 19. [CrossRef]

26. Yan, Q.; Huang, W.; Jin, S.; Jia, Y. Pan-tropical soil moisture mapping based on a three-layer model from CYGNSS GNSS-R data. *Remote Sens. Environ.* **2020**, *247*, 111944. [CrossRef]

27. Rodriguez-Alvarez, N.; Oudrhiri, K. The Bistatic Radar as An Effective Tool for Detecting Andnonitoring the Presence Phytoplankton on the Ocean Surface. *Remote Sens.* **2021**, *13*, 2248. [CrossRef]

28. Ban, W.; Zhang, K.; Yu, K.; Zheng, N.; Chen, S. Detection of Red Tide over Sea Surface Using GNSS-R Spaceborne Observations. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5802911. [CrossRef]

29. Zhang, Y.; Wang, Y.; Zhou, S.; Meng, W.; Han, Y.; Yang, S. Analysis on Feasibility of Detecting Water Blooms in Taihu Lake with Spaceborne GNSS-R. *J. Beijing Univ. Aeronaut. Astronaut.* **2024**, *50*, 695–705.

30. Zhen, Y.; Yan, Q. Improving Spaceborne GNSS-R Algal Bloom Detection with Meteorological Data. *Remote Sens.* **2023**, *15*, 3122. [CrossRef]

31. Wu, Y.; Dai, R.; Xu, Y.; Han, J.; Li, P. Statistical assessment of water quality issues in Hongze Lake, China, related to the operation of a water diversion project. *Sustainability* **2018**, *10*, 1885. [CrossRef]

32. Ruf, C.S.; Atlas, R.; Chang, P.; Clarizia, M.P.; Garrison, J.L.; Gleason, S.; Zavorotny, V.U. New Ocean Winds Satellite Mission to Probe Hurricanes and Tropical Convection. *Bull. Am. Meteorol. Soc.* **2016**, *97*, 385–395. [CrossRef]

33. Eroglu, O.; Kurum, M.; Boyd, D.; Gurbuz, A.C. High Spatio-Temporal Resolution CyGNSS Soil Moisture Estimates Using Artificial Neural Networks. *Remote Sens.* **2019**, *11*, 2272. [CrossRef]

34. Egido, A.; Paloscia, S.; Motte, E.; Guerriero, L.; Pierdicca, N.; Caparrini, M.; Floury, N. Airborne GNSS-R Polarimetric Measurements for Soil Moisture and Above-Ground Biomass Estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 1522–1532. [CrossRef]

35. Mograne, M.A.; Jamet, C.; Loisel, H.; Vantrepotte, V.; Mériaux, X.; Cauvin, A. Evaluation of Five Atmospheric Correction Algorithms over French Optically-complex Waters for the Sentinel-3A OLCI Ocean Color Sensor. *Remote Sens.* **2019**, *11*, 668. [CrossRef]

36. Bai, X.; Hu, W.; Hu, Z.; Li, X. Importation of Wind-driven Drift of Mat-like Algae Bloom into Meiliang Bay of Taihu Lake in 2004 Summer. *Environ. Sci.* **2005**, *26*, 57–60.

37. Matthews, M.W.; Bernard, S.; Robertson, L. An Algorithm for Detecting Trophic Status (chlorophyll-a), Cyanobacterial-Dominance, Surface Scums and Floating Vegetation in Inland and Coastal Waters. *Remote Sens. Environ.* **2012**, *124*, 637–652. [CrossRef]

38. Matthews, M.W.; Odermatt, D. Improved Algorithm for Routine Monitoring of Cyanobacteria and Eutrophication in Inland and Near-coastal Waters. *Remote Sens. Environ.* **2015**, *156*, 374–382. [CrossRef]

39. *DB44/T 2261-2020*; Technical Specification for Classification and Monitoring of Algal Blooms. Market Supervision Administration of Guangdong Province: Guangzhou, China, 2020.

40. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Oakland, CA, USA, 1967; Volume 1, pp. 281–297.

41. Carreno-Luengo, H.; Luzi, G.; Crosetto, M. Sensitivity of CyGNSS Bistatic Reflectivity and SMAP Microwave Radiometry Brightness Temperature to Geophysical Parameters over Land Surfaces. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *12*, 107–122. [CrossRef]

42. Senyurek, V.; Lei, F.; Boyd, D.; Kurum, M.; Gurbuz, A.C.; Moorhead, R. Machine Learning-based CYGNSS Soil Moisture Estimates over ISMN Sites in CONUS. *Remote Sens.* **2020**, *12*, 1168. [CrossRef]

43. Senyurek, V.; Lei, F.; Boyd, D.; Gurbuz, A.C.; Kurum, M.; Moorhead, R. Evaluations of Machine Learning-based CYGNSS Soil Moisture Estimates Against SMAP Observations. *Remote Sens.* **2020**, *12*, 3503. [CrossRef]

44. Yang, T.; Wan, W.; Sun, Z.; Liu, B.; Li, S.; Chen, X. Comprehensive Evaluation of Using TechDemoSat-1 and CYGNSS Data to Estimate Soil Moisture over Mainland China. *Remote Sens.* **2020**, *12*, 1699. [CrossRef]

45. Ruf, C.S.; Gleason, S.; McKague, D.S. Assessment of CYGNSS Wind Speed Retrieval Uncertainty. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *12*, 87–97. [CrossRef]

46. Ruf, C.S.; Posselt, J.D.; Majumdar, S.; Gleason, S.; Clarizia, M.P.; Starkenburg, D.; Provost, D.; Zavorotny, V.U.; Murray, J.; Musko, S.; et al. *CYGNSS Handbook, Cyclone Global Navigation Satellite System: Deriving Surface Wind Speeds in Tropical Cyclones*; University of Michigan: Ann Arbor, MI, USA, 2016.

47. Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794.

48. Douzas, G.; Bacao, F.; Fonseca, J.; Khudinyan, M. Imbalanced learning in land cover classification: Improving minority classes' prediction accuracy using the geometric SMOTE algorithm. *Remote Sens.* **2019**, *11*, 3040. [CrossRef]

49. Kravitz, J.; Matthews, M.; Bernard, S.; Griffith, D. Application of Sentinel 3 OLCI for chl-a retrieval over small inland water targets: Successes and challenges. *Remote Sens. Environ.* **2020**, *237*, 111562. [CrossRef]