*remote sensing*

MDPI

*Article*

# A Fully Unsupervised Machine Learning Framework for Algal Bloom Forecasting in Inland Waters Using MODIS Time Series and Climatic Products

**Pedro Henrique M. Ananias [1], Rogério G. Negri [1,2,*], Maurício A. Dias [3], Erivaldo A. Silva [3] and Wallace Casaca [4]**

[1] Graduate Program in Natural Disasters, São Paulo State University (UNESP), National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN), São José dos Campos 12245-000, Brazil
[2] Science and Technology Institute (ICT), São Paulo State University (UNESP), São José dos Campos 01049-010, Brazil
[3] Faculty of Science and Technology (FCT), São Paulo State University (UNESP), Presidente Prudente 19060-900, Brazil
[4] Institute of Biosciences, Letters and Exact Sciences (IBILCE), São Paulo State University (UNESP), São José do Rio Preto 15054-000, Brazil
* Correspondence: rogerio.negri@unesp.br

**Abstract:** Progressively monitoring water quality is crucial, as aquatic contaminants can pose risks to human health and other organisms. Machine learning can support the development of new effective tools for water monitoring, including the detection of algal blooms from remotely sensed image series. Therefore, in this paper, we introduce the Algal Bloom Forecast (ABF) framework, a fully automated framework for algal bloom prediction in inland water bodies. Our approach combines machine learning, time series of remotely sensed products (i.e., Moderate-Resolution Imaging Spectroradiometer (MODIS) images), environmental data and spectral indices to build anomaly detection models that can predict the occurrence of algal bloom events in the posterior period. Our assessments focused on the application of the ABF framework equipped with the support vector machine (SVM), random forest (RF), and long short-term memory (LSTM) methods, the outcomes of which were compared through different evaluation metrics such as global accuracy, the kappa coefficient, F1-Score and $R^2$-Score. Case studies covering the Erie (USA), Chilika (India) and Taihu (China) lakes are presented to demonstrate the effectiveness and flexibility of our learning approach. Based on comprehensive experimental tests, we found that the best algal bloom predictions were achieved by bringing together the ABF design with the RF model.

**Keywords:** algal bloom; remote sensing; MODIS; prediction; machine learning



**Citation:** Ananias, P.H.M.; Negri, R.G.; Dias, M. A.; Silva, E.A.; Casaca, W. A Fully Unsupervised Machine Learning Framework for Algal Bloom Forecasting in Inland Waters Using MODIS Time Series and Climatic Products . *Remote Sens.* **2022**, *14*, 4283. https://doi.org/10.3390/rs14174283

Academic Editor: Paraskevas Tsangaratos, Wei Chen, Ioanna Ilia and Haoyuan Hong

Received: 25 July 2022
Accepted: 25 August 2022
Published: 30 August 2022



## 1. Introduction

Water quality is of vital importance for life on Earth, mainly due to the recent increase in population and climate change [1,2]. As a result, there has been a growth in water demand for agricultural, industrial and domestic uses, thus increasing pressure on the global environment [3]. Another issue that has appeared in this context is the proliferation of cyanobacteria, which are responsible for severe damage to ecological structures and aquatic ecosystems [4]. This phenomenon is known as algal blooms, and it can lead to serious risks threatening the behavior and health of living beings [5].

Concerning the different approaches available for assessing the water quality, which includes the presence of algae, Chawla et al. [2] presented several bio-indicators, in particular, suspended sediment, turbidity, total phosphorus, dissolved organic content, temperature, and Secchi disks. Moreover, Gons [6] emphasized that the presence of chlorophyll-a (Chl-a) has been a very useful parameter in inspections of water purity. In fact, a high incidence

of Chl-a in lakes and rivers may be related to sudden changes in conditions such as surface temperature, wind speed, precipitation, water column stratification and water flow direction [4,5,7].

A straightforward way of monitoring water quality is to assess the water resources continuously [3]. However, several studies have reported a global decline in the use of hydrometric stations in isolation to perform such monitoring [2,3]. Their non-uniform spatial distribution and operational faults due to a lack of government investment are some potential factors that may demotivate the use of hydrometric stations for water monitoring. A feasible alternative to hydrometric stations is the application of remote sensing techniques [8,9]. For instance, Qi et al. [8] proposed an algorithm based on empirical orthogonal functions to estimate the concentration of the Chl-a in Lake Taihu, China. Similarly, Allen et al. [9] employed satellite images from the Moderate-Resolution Imaging Spectroradiometer (MODIS) sensor to predict algal blooms in a eutrophic coastal area. Spectral indices have also been used for algal bloom detection [10–14]. For example, Hu [10] introduced the floating algae index (FAI) to identify floating algae. Although they are flexible and very functional, spectral index-based methods are not capable of dealing with historic records of algal bloom events over time [5].

Recent technological advances have enabled the use of several imaging sensors, producing large spatio-temporal datasets. When combined with modern artificial intelligence methods, these datasets allow for the determination of the Chl-a concentration [15,16]. A representative example is the study conducted by Zhang et al. [17], in which a support vector machine (SVM)-based algorithm and Landsat-8/OLI images were used to map Chl-a in Chinese lakes. Similarly, Ananias and Negri [18] introduced a data-driven method to detect the occurrence of algal blooms in inland waters using a one-class SVM. By unifying Sentinel-2 multispectral data and random forest (RF) forecasts, Silveira et al. [19] addressed the issue of predicting the levels ofChl-a in two small water bodies. More recently, deep-learning-based methods have been demonstrated to be effective when approximating environmental parameters [15,20,21]. For instance, Barzegar et al. [22] introduced a hybrid model that combined a convolutional neural network and long short-term memory (LSTM) to estimate the Chl-a level in a lake in Greece. Cho et al. [20] applied LSTM nets for Chl-a detection in the Geum River (South Korea) over a horizon of one-to-four days, whereas Yu et al. [23] employed LSTM and wavelet concepts to predict long-term Chl-a occurrences in a lake in China. It is important to point out that Cho et al. [20] and Barzegar et al. [22] measured physicochemical variables collected using in situ sensors installed at specific positions of the Geum River (South Korea) and Lake Prespa (Balkan peninsula). As a result, the predicted Chl-a concentration was not mapped onto the spatial domain. Conversely, Yu et al. [23] also considered physicochemical variables to estimate the regional Chl-a concentration in Lake Dianchi (China), but their method was focused on generating annual predictions instead.

Despite their accuracy and high learning capacity, most machine/deep learning-based methods still rely on large datasets of labeled data to properly work. As pointed out by Yuan et al. [15], both remotely sensed data and machine learning methods can be used to deliver a forecast model built from ground-truth samples, which are usually collected from the study area. As a consequence, obtaining in situ information may be critical in a multitude of cases, thus motivating the development of new methodologies that do not depend on any reference data to be effective and functional.

In light of the above-presented discussion, in this paper, we propose a fully automatic machine/deep-learning-based methodology for the spatial prediction of algal bloom events in aquatic environments, integrating time-series of satellite images and environmental data acquired by means of aerial remote sensing. More specifically, we address the issue of the forecasting of algal blooms as an anomaly detection problem in image time series by applying threshold-based rules so as to identify and build a representative database of abnormal events and then generate a decision function for the prediction of algal-bloom-like anomalies. Three case studies concerning the Erie (USA), Chilika (India) and Taihu (China)

lakes were carried out to demonstrate the effectiveness and applicability of the proposed methodology. Images acquired using the MODIS sensor comprised these applications' primary data source.

In summary, the main contributions of this paper are:

(i) A fully unsupervised learning methodology, designed to characterize, detect and predict algal blooms as time-varying anomalous events in wetland areas.
(ii) The proposed approach is modular, i.e., it can be integrated with any classification model in addition to those presented in our formalization.
(iii) A conceptual formalization that can be extended to other environmental issues besides inland water anomaly detection for algal blooms.

This paper is organized as follows. Section 2 briefly presents some basic notations, machine learning models and spectral indices, and formalizes the development of the proposed approach. In Section 3 we discuss the experiment design, study areas and the obtained results. Finally, Section 4 concludes the current study.

## 2. Methods

### *2.1. Machine Learning Models*

Let $\mathcal{I}$ be an image remotely obtained by means of a sensor of which the pixels $s \in \mathcal{S} \subset \mathbb{N}^2$ are associated with an attribute vector $\mathbf{x}_s = [x_{s:1}, \ldots, x_{s:n}]$, defined on a feature space $\in \mathcal{X} \subset \mathbb{R}^n$. Based on a given set of images represented by a time series of $t$ instants, we assume the notation $\mathcal{I}^{(\ell)}$, with $\ell = 1, \ldots, t$, to describe the spatio-temporal representation of these data.

The image classification problem consists of assigning a class $\omega_k \in \Omega = \{\omega_1, \ldots, \omega_c\}$ to each $s \in \mathcal{S}$, by applying a function $G : \mathcal{X} \to \mathcal{Y}$ over $\mathbf{x}_s$, where $\mathcal{Y} = \{1, \ldots, c\}$. In order to properly apply $G$, a pre-defined training set $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, \ldots, m\}$ is required, where $(\mathbf{x}_i, y_i)$ indicates that $\mathbf{x}_i$ is assigned to a particular class $\omega_a$ if the number $y_i \in \mathbb{N}$ is equal to $a$. Conveniently, we denote here by $\mathcal{C}^{(\ell)}$ the output produced from $G$ for each position $s$ of $\mathcal{I}^{(\ell)}$.

Given the importance of image classification in several applications of remote sensing, creating more efficient and accurate methods has become a persistent challenge [24,25]. Among various examples, consolidated methods such as support vector machine (SVM) [26], random forest (RF) [27] and the recent neural network-based approaches have been of paramount importance in supporting a variety of remote sensing applications.

Introduced by Vladimir Vapnik [28], the SVM method comprises a supervised learning algorithm that aims to determine classes via a hyperplane $g(\mathbf{x}) = K(\mathbf{x}, \mathbf{w}) + b$, of which the separating margin is the maximum [26,29], and $\mathbf{w}$ and $b$ are inner parameters. Here, $K(\cdot, \cdot)$ is a kernel function [30] which is conveniently defined according to the complexity of the classification problem.

As described by Bruzzone et al. [31], starting from a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, +1\} : i = 1, \ldots, m\}$, where $y_i = \pm 1$ indicates membership between two classes, the SVM training stage comprises the calculation of parameters $\mathbf{w}$ and $b$ after solving the following optimization problem:

$$\max_{\alpha} \left( \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \right) \quad s.t. \begin{cases} \sum_{i=1}^{n} y_i \alpha_i = 0 \\ 0 \leq \alpha_i \leq C \\ i = 1, \ldots, n \end{cases} \tag{1}$$

where $\alpha_i \in \mathbb{R}$ are Lagrange multipliers and $C \in \mathbb{R}_0^+$ is a penalty factor applied to misclassifications. Regarding the kernel function, the radial basis function (RBF) $K(\mathbf{x}_i, \mathbf{w}) = e^{-\gamma \|\mathbf{x}_i - \mathbf{w}\|^2}$, with $\gamma \in \mathbb{R}_0^+$, is highlighted as a convenient option.

The Lagrange multipliers obtained when solving Equation (1) allow the determination of the classification rule $G(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b\right)$. For a comprehensive discussion of SVM and kernel functions, see [32].

The RF method was introduced by Breiman [27], and it consists of generating a classification rule $G : \mathcal{X} \to \mathcal{Y}$ from an ensemble of decision trees [33]. For a given training set $\mathcal{D}$, the bootstrap sampling process is applied to generate $n_{est}$ replicas of this set. For each replica, a subset with up to $n_{att}$ attributes is randomly taken and then used to train a single decision tree. Parameters such as the maximum depth ($p_{depth}$) and the minimum examples per leaf for splitting ($p_{split}$ and $p_{leaf}$, respectively) have to be properly tuned before training. A detailed discussion of these parameters can be found in [27].

After training the set of trees, the sub-classifiers $G_k : \mathcal{X} \to \mathcal{Y}$, $k = 1, \ldots, n_{est}$, are combined so that the attribute vectors $\mathbf{x} \in \mathcal{X}$ are labeled according to a class $\omega_a \in \Omega$ and Equation (2):

$$G(\mathbf{x}) = \underset{a \in \{1,\ldots,c\}}{\arg\max} \left\{ \sum_{k=1}^{n_{est}} \delta_a(G_k(\mathbf{x})) \right\}, \tag{2}$$

where $\delta_a(G_k(\mathbf{x})) = 1$, if $G_k(\mathbf{x}) = a$; otherwise, $\delta_a(G_k(\mathbf{x})) = 0$.

In recent years, neural network-inspired methods have emerged as an effective alternative for remote sensing image classification [15]. Such models are distinguished by a high generalization capability when coping with uncorrelated data.

Among the different neural network models proposed in the literature, the long short-term memory (LSTM) [34] architecture allows the analysis of temporal effects on the learning process. The diagram in Figure 1 summarizes an elementary part of this model and its interaction. First, let us assume that $\mathbf{x}_i$, $i = 1, \ldots, m$, is a set of patterns (i.e., attribute vectors) ordered by the indices $i$ and sequentially submitted to concatenation, element-wise multiplication and vector sum operations ($\odot$, $\otimes$ and $\oplus$), as to the sigmoid ($\sigma$) and hyperbolic tangent ($\varphi$) functions. Next, we also set the following elements that are used during the training process: $\mathbf{c}_{i-1}$ and $\mathbf{c}_i$ as a vector with the "previous state" and "current" of the network; $\mathbf{f}_i$ as a vector comprising "forgetting factors" for each component of the input vector; $\mathbf{i}_i$ as a "signal modulation" for input information; and $\mathbf{h}_{i-1}$ and $\mathbf{h}_i$ as the network input and output vectors in the "previous" and "current" states, respectively.

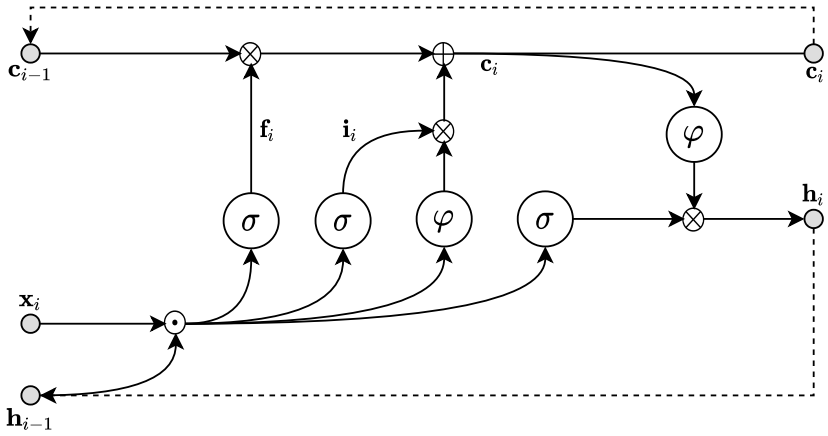


**Figure 1.** Overview of the LSTM architecture.

After successively processing $\mathbf{x}_i$ and considering the respective predictions $\mathbf{h}_i$, the training process stops when network convergence occurs. After that, the classification is performed via a softmax function and outputs $\mathbf{h}_i$, of which the class indicators $y_i$ are known from $(\mathbf{x}_i, y_i) \in \mathcal{D}$. According to this approach, the combination between the network and the softmax function comprises the classification model $G$. A more extensive and detailed discussion about LSTM-based models and their parameters can be found in [34].

### *2.2. Algal Bloom Detection via Spectral Index Thresholding*

Spectral indices comprise a particular type of descriptor generated from remote sensing images for discriminating specific sets of targets. Several spectral indices have been proposed in the literature, for example, to characterize water bodies [35–37] and vegetation health [38–41]. For the application of algal bloom detection, spectral indices are used by applying rigid thresholds. For instance, Zhao et al. [42] utilized the normalized difference vegetation index (NDVI) [38] to identify algae when this index achieved values greater than a certain threshold ($-0.15$). Similarly, Xu [37] employed the modified normalized difference water index (MNDWI) [43] to detect algae by taking MNDWI values less than zero. The surface algae bloom index (SABI) [44] and floating algae index (FAI) [45] are other spectral-type indices specifically designed to capture the level of algae incidence.

Table 1 summarizes the spectral indices and their respective characterizations for algae occurrences. Here, $x_{Red}$, $x_{Green}$, $x_{Blue}$, $x_{NIR}$, and $x_{SWIR}$ represent the spectral behavior measured in terms of red, green, near-infrared, and short-wave infrared wavelength bands, respectively, whereas $\lambda_{Red}$, $\lambda_{NIR}$ and $\lambda_{SWIR}$ are the midpoints of each mentioned band.

**Table 1.** Summary of spectral indices and thresholds used for algal bloom detection.

| Spectral Index | Expression | Algae Threshold | Reference |
|---|---|---|---|
| NDVI | $\frac{x_{NIR} - x_{Red}}{x_{NIR} + x_{Red}}$ | $>-0.15$ | [42] |
| MNDWI | $\frac{x_{Green} - x_{SWIR}}{x_{Green} + x_{SWIR}}$ | $<0$ | [37] |
| SABI | $\frac{x_{NIR} - x_{Red}}{x_{Blue} + x_{Green}}$ | $>-0.1$ | [44] |
| FAI | $x_{NIR} - \left[x_{Red} + (x_{SWIR} - x_{Red}) \times \frac{\lambda_{NIR} - \lambda_{Red}}{\lambda_{SWIR} - \lambda_{Red}}\right]$ | $>-0.004$ | [45] |

### *2.3. Characterizing and Forecasting an Algal Bloom as an Anomalous Event*

In this section, we introduce the proposed methodology for algal bloom characterization and prediction in inland water bodies. In Section 2.3.1 we presents a conceptual formalization of our proposal, whereas in Section 2.3.2 we provide relevant details regarding the implementation of the computational framework.

#### 2.3.1. Conceptual Formalization

The current methodology for predicting anomalous events in aquatic environments, referred to here as *anomalous behavior forecasting* (ABF), relies on the combination of machine learning concepts and time-varying remotely sensed data. Figure 2 illustrates the main components of the proposed ABF methodology.

The first component accounts for building a spatio-temporal database to train a given machine learning model that is applied for predicting anomalous events. In this stage, five basic parameters are defined: a *region of interest* containing the spatial domain, where the predictive mapping will be performed; a *period of analysis* comprising the range of instants to collect information; the number of *past instants* that compose the input objects (i.e., the attribute vector) in the predictive model; the number of instants concerning the *forecasting period*; and the *spatial context size* used to estimate the local algae concentration. Next, a database is created which gathers information from different sensors and products regarding the *region of interest* and the *period of analysis*. This information is then structured according to $t + 1$ instants, with the first $t$ moments employed to train the prediction model, and the last instant (*early instant*) dedicated to performing the predictions.

The data are automatically collected via the application programming interface (API) for the Google Earth Engine (GEE) platform. More specifically, the constructed database is composed of images acquired by means of the MODIS sensor (MOD09GA.006—500 m

of spatial resolution) and products provided by the National Aeronautics and Space Administration Global Land Data Assimilation projects (NASA GLDAS 2.1) [46] (wind speed, temperature, precipitation, pressure and evapotranspiration), the World Wide Fund for Nature Hydrological Data and Maps Based on Shuttle Elevation Derivatives at Multiple Scales (WWF HydroSHEDS) [47] (drainage direction), the Japan Aerospace eXploration Agency Advanced Land Observing Satellite Digital Surface Model World 3D (JAXA AW3D30) [48] (elevation) and the National Aeronautics and Space Administration MODIS Terra Land Surface Temperature and Emissivity Daily Global (NASA MOD11A1.006) [49] (emissivity). The above-described data are then re-sampled in terms of their spatial resolution to match MODIS data (i.e., 500 m).

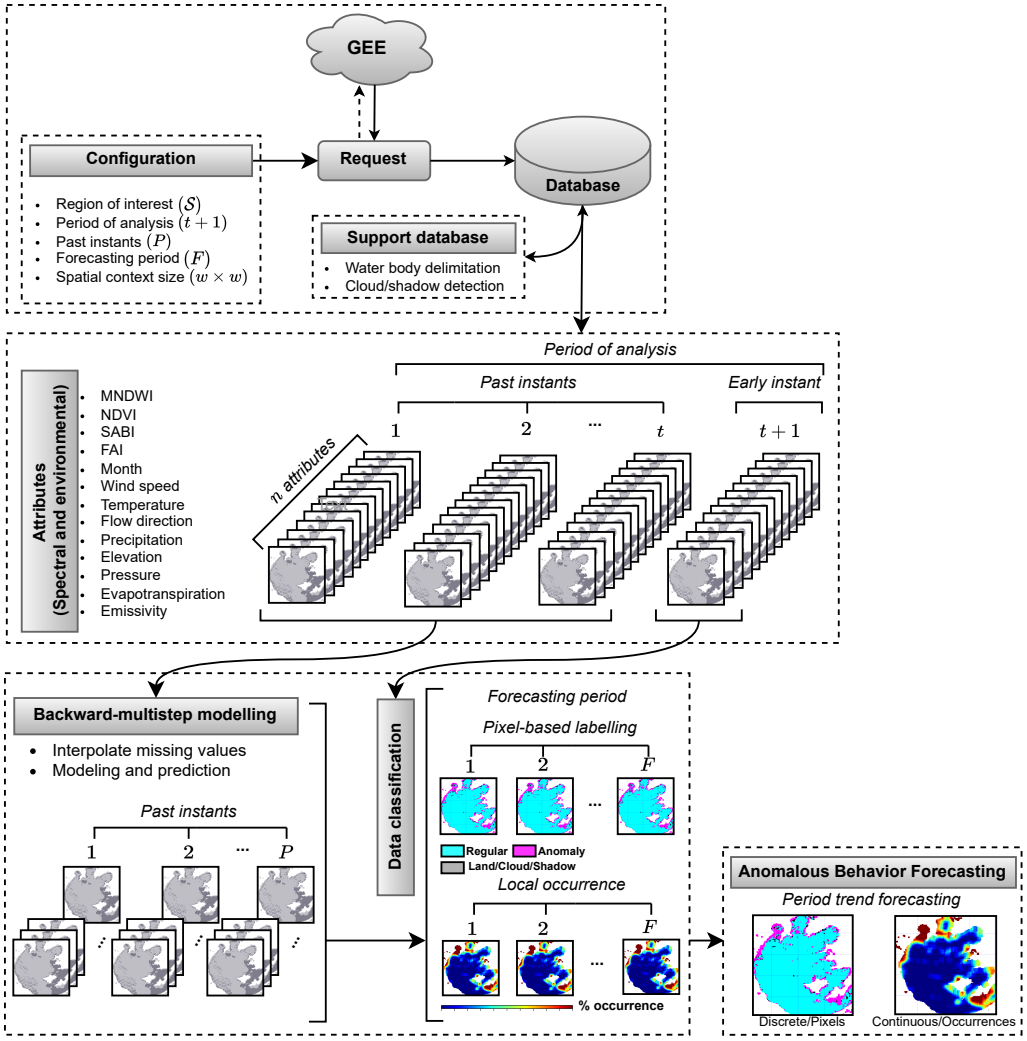


**Figure 2.** Overview of the proposed framework.

By simultaneously taking images from the MODIS sensor, a *support database* is also built, containing information about the water body region (see Section 2.3.2—Support Data) present in the *region of interest* and about cloud and shadow occurrences over the MODIS images, which impose information gaps on the time series and the demand for posterior treatment. After identifying the water body region, the entire database is reduced to its boundaries. In our approach, we assume that $\mathcal{S}$ is limited to the water body area.

After defining both primary and support databases, several spectral indices are extracted from the remote sensing images and then integrated with the environmental information to compose the attribute vectors that express an algal-bloom-like anomalous event. It is worth mentioning that such a process is locally-executed without using the

cloud-computing facilities offered by the GEE platform. Formally, assuming that the *period of analysis* is composed of $t+1$ instants, with one representing the actual (current) and the remaining ones ($t$) representing past instants, a time series $\mathcal{I}^{(\ell)}$ is built for $\ell = 1, \ldots, t+1$, such that $\mathcal{I}^{(\ell)}(s) = \mathbf{x}_s^{(\ell)} = \left[x_{s:1}^{(\ell)}, \ldots, x_{s:n}^{(\ell)}\right]$ is an attribute vector in the $\ell$-th instant. The vector components consist of MNDWI, NDVI, SABI and FAI spectral indices; the instant month number; and the environmental variables such as wind speed ($\mathrm{m\,s^{-1}}$), temperature (K), drainage direction, precipitation ($\mathrm{kg/m^2/s}$), elevation (m), pressure (Pa), evapotranspiration ($\mathrm{kg/m^2/s}$) and emissivity.

Next, the data are re-structured according to a "backward-multistep" scheme, where the $P-1$ previous instants of each observation are concatenated in the form of a single attribute vector. More precisely, for every $s \in \mathcal{S}$ taken from $\mathcal{I}^{(\ell)}$, $\ell = P+1, \ldots, t-F+1$, the attribute vector $\mathbf{x}_s^{(\ell|P)} = \left[x_{s:1}^{(\ell-P)}, \ldots, x_{s:n}^{(\ell-P)}, \ldots, x_{s:1}^{(\ell)}, \ldots, x_{s:n}^{(\ell)}\right] \in \widetilde{\mathcal{X}}$ is defined, where $\widetilde{\mathcal{X}}$ represents a time-extended attribute space. For the cases wherein a certain component $x_{s:i}^{(k)}$ is influenced by the presence of cloud/shadow, we fill out the respective values by using interpolation over time according to the McKinney approach [50].

For each generic position $s$ and instant $\ell$, we assign labels related to the occurrence of anomalies in the next $F$ instants, herein denoted by $\mathbf{y}_s^{(\ell|F)} = \left[y_s^{(\ell+1)}, \ldots, y_s^{(\ell+F)}\right]$, where $y_s^{(k)} = +1$ or $-1$ indicate the occurrence of an algal bloom (i.e., an anomaly) at the $k$-th instant. These labels are established by an ensemble scheme of majority voting involving thresholding rules on the NDVI, MNDWI, SABI and FAI spectral indices (discussed in Section 2.2). Under these conditions, $y_s^{(k)} = +1$ if all the four spectral indices endorse the occurrence of an anomaly with respect to position $s$ and instant $k$; otherwise, $y_s^{(k)} = 0$. The rationale behind using only the instants from $P+1$ to $t-F+1$ when defining the vectors $\mathbf{x}_s^{(\ell|P)}$ and $\mathbf{y}_s^{(\ell|F)}$ is that if we assume $\ell = P-1$, the attempt of retrieve the $P-1$ previous instants to determine $\mathbf{x}_s^{(\ell|P)}$ will cause a disruption in the lower bound of the time series. Analogously, $\ell = t-F+2$ will cause an upper-bound break when trying to define $\mathbf{y}_s^{(\ell|F)}$. Figure 3 elucidates the above-presented explanation.

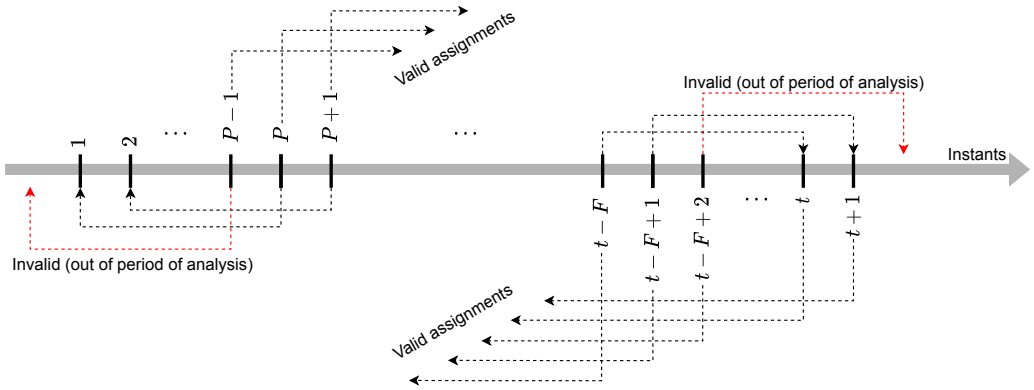


**Figure 3.** Representation of instants and past and future relations.

Note that the vectors $\mathbf{x}_s^{(\ell|P)}$ and $\mathbf{y}_s^{(\ell|F)}$, $\ell = P+1, \ldots, t-F+1$ and $s \in \mathcal{S}$ allow one to define different training sets $\mathcal{D}^{(q)} = \left\{(\mathbf{x}_s^{(\ell|P)}, y_s^{(\ell+q)}) : s \in \mathcal{S}; \ell = P+1, \ldots, t-F+1\right\}$, $q = 1, \ldots, F$. As a result, each $\mathcal{D}^{(q)}$ is used to train a classifier $G_q : \widetilde{\mathcal{X}} \to \{+1, -1\}$ that predicts the occurrence of algae at the $q$-th instant after the most recent instant (i.e., $t+1$), characterized by the attribute vectors associated with each position/pixel of $\mathcal{I}^{(t+1|P)}$, thus resulting in a classification $\mathcal{C}^{(q)}$, $q = 1, \ldots, F$. Note that a useful trait of our approach is that any machine learning method can be used to define $G_q$, $q = 1, \ldots, F$, as those have already been discussed in Section 2.1.

From classifications $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(F)}$, the local anomaly concentrations are estimated by applying a convolution filter that checks the percentage of anomaly occurrences into a neighborhood of $w \times w$ pixels wide. Positions outside $\mathcal{S}$ or affected by the occurrence of cloud/shadow are discarded. The respective outcomes are denoted here by $\mathcal{O}^{(1)}, \ldots, \mathcal{O}^{(F)}$.

Finally, based on the predictions of $\mathcal{I}^{(t+1|P)}$, a *period trend forecast* is computed, in which the median of the estimates obtained at each position is considered. Under these conditions, the representations $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(F)}$ give place to a "discrete occurrence" trend representation. Similarly, a characterization is also obtained for the occurrence concentration of the forecasting period through $\mathcal{O}^{(1)}, \ldots, \mathcal{O}^{(F)}$, thus generating a "continuous occurrence" representation.

#### 2.3.2. Implementation Details and Parametrization

We now provide some details on the implementation of our methodology. The full source code is publicly available for use (see the *Data Availability Statement* below).

**Programming Language and Libraries:** Python 3.8 [51] was adopted as the programming language. We also utilized the modules Numpy [52] and Pandas [53] for data organization procedures; Scikit-Learn [54] to run the SVM and RF implementations; and Keras [55] to implement the LSTM model.

**Google Earth Engine API:** The Google Earth Engine application programming interface (API) [56] was used to obtain MODIS sensor data (MOD09GA.006 and MOD11A1.006), and the products GLDAS 2.1, HydroSHEDS and AW3D30. In our methodology, the product MOD09GA.006 was atmospherically corrected [57].

**Spectral Indices and Thresholds:** In order to appropriately select the best set of thresholds for the spectral indices, we followed the methods of [37,42,44,45] to take the values established in these studies.Although different thresholds may result in more accurate outputs for certain study areas, the authors in [37,42,44,45] verified that the variations were minimal for a multitude of areas analyzed. Therefore, we opted to take the pre-established thresholds given in [37,42,44,45] to drive our method, but with the advantage of keeping them as free parameters in our implementation, making the method flexible enough to meet the requirements of other study areas and applications of interest.

**Model Parameter Tuning:** The selection of suitable hyperparameters for each classification method (Section 2.1) was conducted by applying the randomized grid search procedure [58–60] with a five-fold cross-validation. The hyperparameter space-search taken for the LSTM model comprised $n_{epochs} \in \{5, 10, 20, \ldots, 100\}$, $p_{dropout} \in \{0, 0.1, \ldots, 0.5\}$, $n_{layer} \in \{16, 32, \ldots, 512\}$. Hyperparameters related to the RF method ranged as part of the following sets: $n_{est} \in \{1, 5, \ldots, 250\}$, $p_{depth} \in \{1, 2, \ldots, 30\}$, $p_{split} \in \{2, 4, \ldots, 20\}$, $p_{leaf} \in \{2, 4, \ldots, 20\}$ and $n_{att} \in \{\sqrt{dim(\widetilde{\mathcal{X}})}, 100\%, 75\%, 50\%\}$. Regarding the SVM method, the examined hyperparameters varied as follows: $\gamma \in \{10^{-1}, 10^{-2}, \ldots, 10^{-7}\}$ and $C \in \{1, 2, 3, \ldots, 1000\}$.

**Support Database:** This database comprised the water body spatial boundaries and the occurrences of cloud/shadow during the period of analysis. Concerning water body identification, we employed the "water_mask" sub-product from MODIS (MOD44W.006/"Terra Land Water Mask") to deliver water surface mapping with 250 m spatial resolution. Furthermore, bitwise operations were applied on the "state_1km" sub-product to detect the presence of clouds and shadows.

**Period of Analysis:** This consisted of an image time series used as input data to train a machine learning model in our ABF framework. Depending on the availability of data for each study area, we were able to take different "periods of analysis" when applying our methodology. Since the period of analysis used to train the predictive models may influence the outputs, we ran an extensive battery of tests to select the most suitable training window: 180 days. More specifically, we used the

following periods when performing our experimental tests: 30, 60, 90, 180, 365, 730 and 1825 days.

**Past Instants and Forecasting Period:** The parameters of *past step instants* (the number of past instants required to generate the attribute vectors) and the *forecasting period* (the number of future instants that comprise the prediction interval) are related to the performance and computational cost of the proposed approach. Suitable values for these parameters were analyzed after considering a battery of tests from distinct study areas (Sections 3.1 and 3.2). Figure 4 depicts a median profile, in terms of the $R^2$-Score, regarding the estimated algae occurrence concentration. The results indicated that four *past step instants* and a *forecasting period* of five instants produced the best outcomes.

**Spatial Context Size:** This parameter defines the dimension $w \times w$ of the convolution filter applied on the classifications $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(F)}$, to measure the occurrence of algae. After performing a series of tests where $w$ varied in 3, 5 and 7, the adoption of $w = 7$ provided the best consistency.

**Data Projection:** With the aim of alleviating the computational cost, the principal component analysis (PCA) [61] technique was applied to the full set of attribute vectors $\mathbf{x}_s^{(\ell|P)}$ (i.e., considering all instants and positions), thus allowing the projection of the data onto a feature space of reduced dimensions without losing significant accuracy in the results. More specifically, after a preliminary battery of tests regarding all study areas, the proposed methodology achieved the average values of 0.97, 0.96 and 0.96 for $R^2$-Scores when data projections were applied with 99%, 95% and 90% of explained variance, respectively. However, the computational time drastically decreased by about 70% and 50% when comparing the multidimensional projection with 90% against 99% and 95%, respectively. Consequently, we found that the choice of 90% provided a good trade-off between satisfactory performance and computational cost.

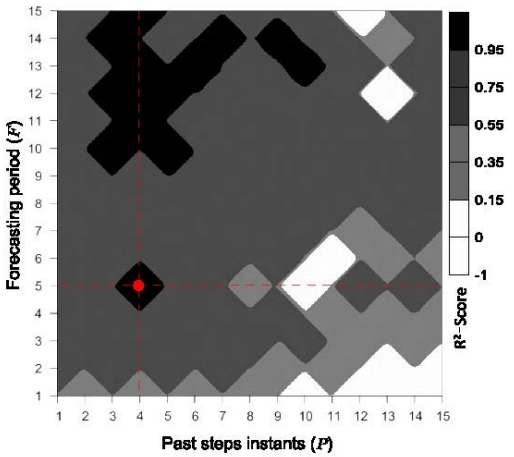


**Figure 4.** Median performance in terms of $R^2$-Scores for different configurations of *past steps instants* and *forecasting period* parameters. The red dot indicates the best performance values.

## 3. Results

Three case studies covering distinct study areas were carried out to assess the results after training the predictive models. The study areas comprised portions of Lake Erie (USA), Chilika (India) and Taihu (China), as shown in Figure 5.

The SVM, RF and LSTM methods were used as classification models in our approach to generate three algal bloom forecasters, which we refer to as ABF-SVM, ABF-RF and ABF-LSTM. The performance of each trained model was measured in terms of several validation metrics, in particular, overall accuracy, kappa coefficient [62], percentages of true/false-positives/negatives (denoted by TP, TN, FP and FN), F1-Score [63], $R^2$-Score [64]

and root mean square error (RMSE) [65]. Hypothesis tests with 5% significance for the kappa coefficient [62] were also computed in an effort to statistically inspect the results.

The experiments were run on a desktop computer with an AMD Ryzen 9 3900X 12-core processor, 32 GB of RAM and an Ubuntu Linux version 20.04 operating system.

### *3.1. Study Areas*

Located between the states of Ohio, New York, Pennsylvania and Michigan, Lake Erie (Figure 5a) is one of the Great Lakes of Saint Lawrence. According to Bolsenga and Herdendorf [66], this lake is approximately 388 km in length and 92 km wide and has an average depth of 19 m. It also presents a total area and volume of 25,657 km$^2$ and 484 km$^3$, respectively. Affected by regional urbanization and agricultural growth, this lake has presented algal blooms since the 1960s [67]. As reported in Stumpf et al. [68], the peak months for algal blooms are typically August and September.

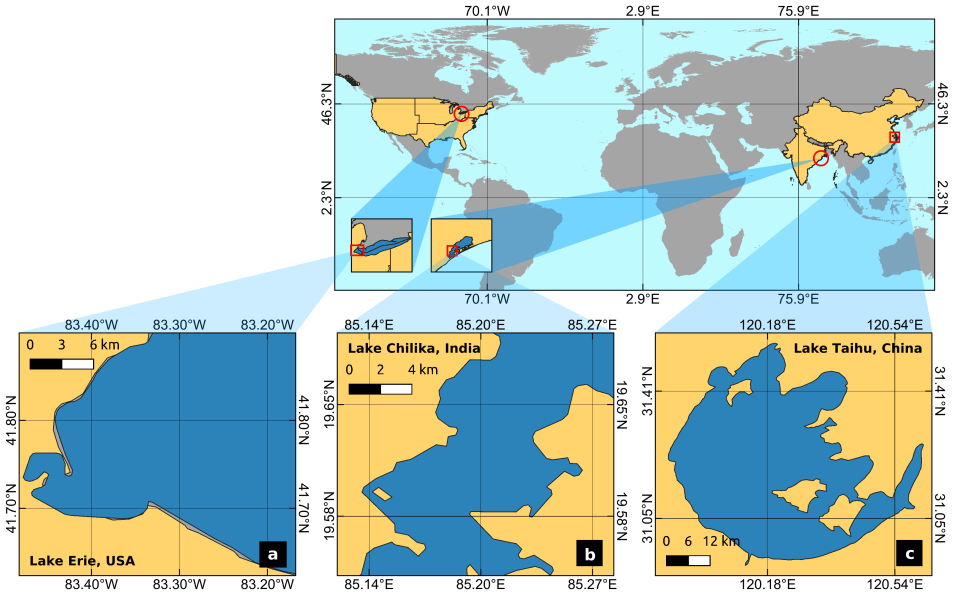


**Figure 5.** Spatial locations of the study areas.

Lake Chilika (Figure 5b) is located on the east coast of Orissa, India. In addition to providing an income source for about 150,000 fishermen [69], this lake comprises the largest brackish water body in Asia, and it strongly contributes to the regional economy [70]. Furthermore, according to Panigrahi [70], such a lake presents an area that varies from 906 km$^2$ (summer) to 1165 km$^2$ (monsoon) with a depth ranging from 0.40 to 4.90 m. In addition to issues such as constant changes in salt concentration, freshwater weeds and declines in fish productivity, Chilika lake has been suffering from a constant process of eutrophication, which leads to high concentrations of algae [71].

The third study area is Lake Taihu (Figure 5c), China. Although it is considered a source of water for more than 40 million people, this lake has frequently been affected by the presence of algal blooms [72]. According to Gao et al. [73], Lake Taihu has an area of 2338 km$^2$ and an average depth of 1.9 m. The wet and dry periods usually occur between April to July and February to March, respectively.

### *3.2. Reference Data*

As previously mentioned, the proposed methodology was assessed through the assessment of different study areas and periods. Regarding the first area (Lake Erie—Figure 5a), 10 October 2013, and 26 July 2015 were considered as *early instants*. Consequently, the respective *forecasting periods* were 11–15 October 2013 (Figure 6a–e) and 27–31 July 2015 (Figure 7a–e). Regarding the second area (Lake Chilika—Figure 5b), 24 November 2019 was considered as an *early instant*; therefore, the *forecasting period* encompassed 25–29 November

2014 (Figure 8a–e). Finally, for the third area (Lake Taihu—Figure 5c) we selected the periods of 24 August 2016 and 24 July 2017 as *early instants*, thus providing the *forecasting periods* shown in Figures 9a–e and 10a–e.

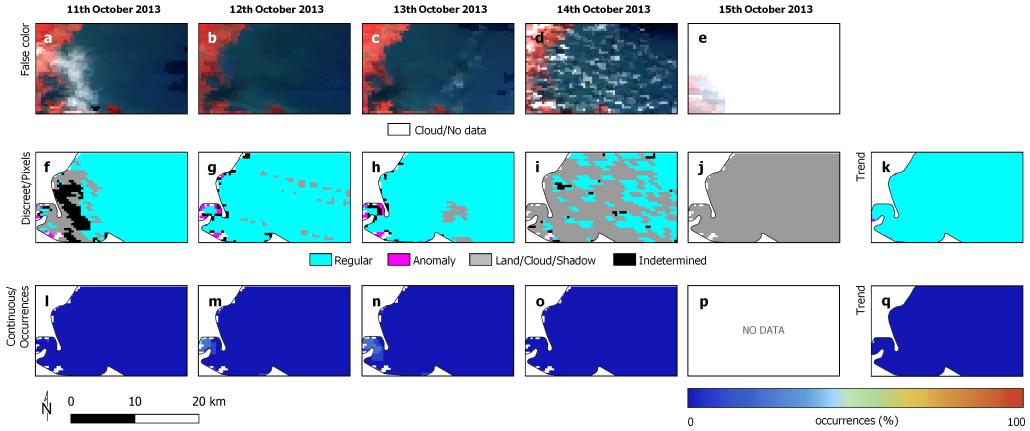


**Figure 6.** Images (**a**–**e**) in false-color composition (NIR, red and green bands) for the Lake Erie study area from 11–15 October 2013. Regular, anomaly, land/cloud/shadow and undetermined samples (**f**–**j**) are identified by cyan, magenta, gray and black polygons, respectively; (**l**–**p**) show the local anomaly concentration at each instant; (**k**,**q**) represent the discrete and local anomaly concentration references for the whole forecasting period.

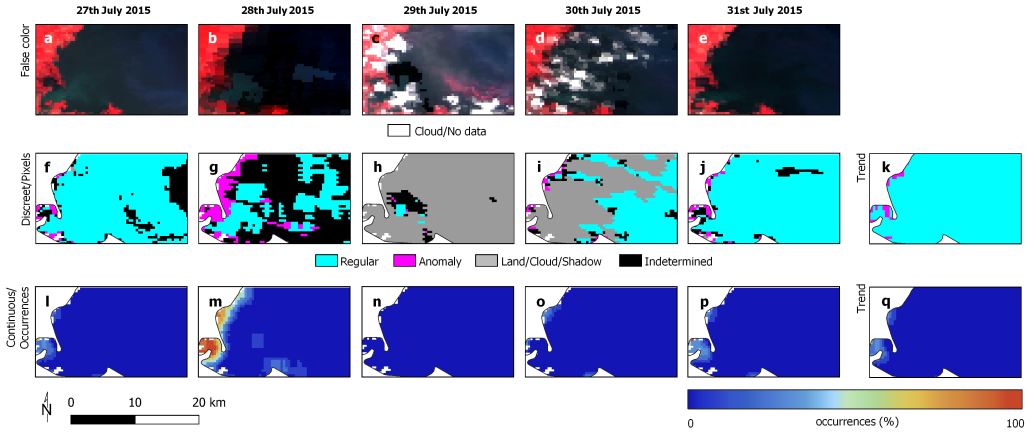


**Figure 7.** Images (**a**–**e**) in false-color composition (NIR, red and green bands) for the Lake Erie study area from 27–31 July 2015. Regular, anomaly, land/cloud/shadow and undetermined samples (**f**–**j**) are identified by cyan, magenta, gray and black polygons, respectively; (**l**–**p**) show the local anomaly concentration at each instant; (**k**,**q**) represent the discrete and local anomaly concentration references for the whole forecasting period.

It is worth stressing that in the following experiments, we simulated a real-world application, i.e., the *forecasting period* was unknown for the ABF method. Thus, based on the current behavior at the *early instants*, predictions were made regarding the occurrence of algal blooms in the *forecasting period*.

The predictive performance of the ABF method was assessed using reference datasets (i.e., ground-truth data) defined through the same process as that discussed in Section 2.3, producing the training data labels (i.e., $\mathbf{y}_s^{(\ell|F)}$ as "regular" or "anomaly"). More precisely: (i) each instant in the *forecasting period* was analyzed according to the thresholding of the spectral indices MNDWI, NDVI, FAI and SABI, calculated based on MODIS images (Section 2.2); (ii) locations/pixels simultaneously identified as anomalies by all four thresholding processes were admitted as "anomalies"; (iii) locations/pixels simultaneously identified as non-anomalies by all four thresholding processes were admitted as "regular"; (iv) lo-

cations/pixels with a partial agreement regarding the occurrence of anomalies or under the influence of cloud/shadow were admitted as "indeterminate" or "land/cloud/shadow" and then disregarded as a reference; (v) considering the anomalies and non-anomalies pixels detected during the *forecasting period*, the median value was used to determine the local trend and generate a discrete reference; and (vi), to create the reference for the local anomaly concentration, a convolution filter of $7 \times 7$ pixels was applied to obtain the local percentage of anomalies at each instant of the *forecasting period*, followed by the local trend computation according to the median value.

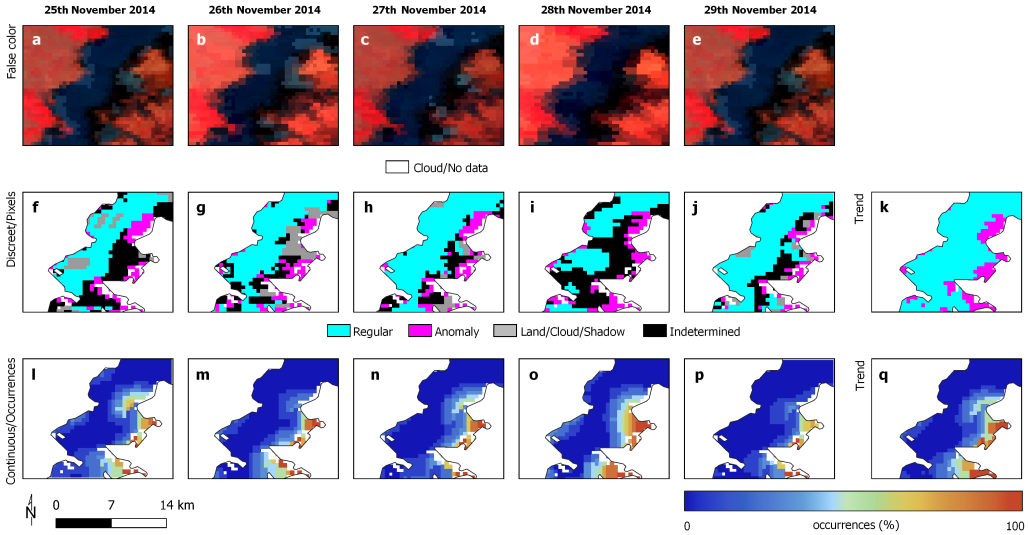


**Figure 8.** Images (**a**–**e**) in false-color composition (NIR, red and green bands) for the Lake Chilika study area from 25–29 November 2014. Regular, anomaly, land/cloud/shadow and indeterminate samples (**f**–**j**) are identified by cyan, magenta, gray and black polygons, respectively; (**l**–**p**) show the local anomaly concentration at each instant; (**k**,**q**) represent the discrete and local anomaly concentration references for the whole forecasting period.

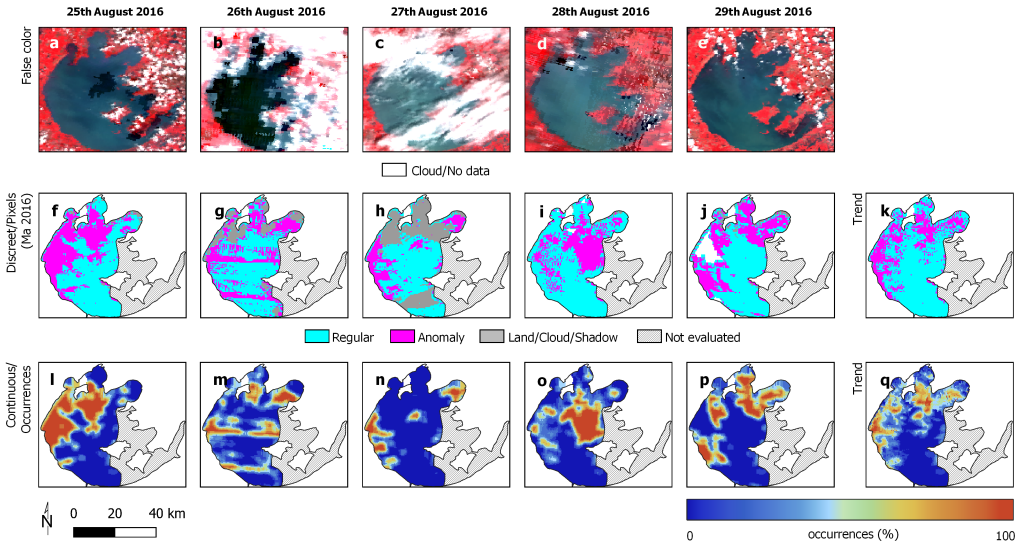


**Figure 9.** Images (**a**–**e**) in false-color composition (NIR, red and green bands) for the Lake Taihu study area from 25–29 August 2016. Regular, anomaly, land/cloud/shadow and undetermined samples (**f**–**j**) are identified by cyan, magenta, gray and black polygons, respectively; (**l**–**p**) show the local anomaly concentration at each instant; (**k**,**q**) represent the discrete and local anomaly concentration references for the whole forecasting period.

Concerning the reference dataset for Lake Taihu, we used the available ground-truth data from the National Earth System Science Data Center (NESSDC) [74,75]. In this case, the occurrence of anomalies was determined based on the locations/pixels wherein the Chl-a concentration was above 20 µg/L [76,77]; otherwise, the location was assumed to be regular/non-anomaly. Finally, the discrete and local anomaly concentration references were produced following the previously discussed steps (v) and (vi).

The reference datasets are illustrated in Figures 6–10k–q. For the sake of presentation, the estimates at each instant are depicted in Figures 6–10f–j and Figures 6–10l–p. Table 2 summarizes the number of anomalies and regular positions/pixels regarding the discrete reference of each study area and *early instant*.

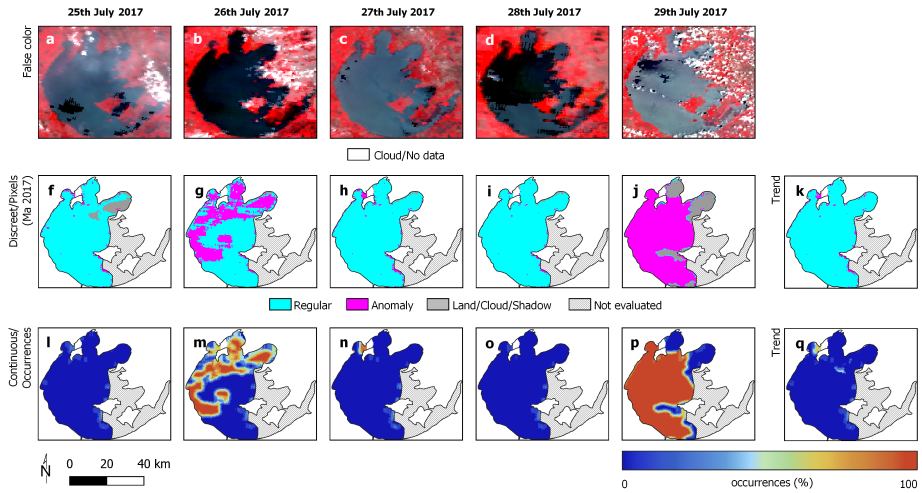


**Figure 10.** Images (**a**–**e**) in false-color composition (NIR, red and green bands) for the Lake Taihu study area from 25–29 July 2017. Regular, anomaly, land/cloud/shadow and undetermined samples (**f**–**j**) are identified by cyan, magenta, gray and black polygons, respectively; (**l**–**p**) show the local anomaly concentration at each instant; (**k**,**q**) represent the discrete and local anomaly concentration references for the whole forecasting period.

**Table 2.** Summary of regular and anomaly pixels regarding each study area and early instants.

| | **Lake Erie** | |
|---|---|---|
| | **11–15 October 2013** | **27–31 July 2015** |
| Regular | 277,328 | 233,137 |
| Anomaly | 19,041 | 13,210 |
| Average occurrence | 0.8% | 3.2% |
| Stand. deviation occurrences | 4.5% | 11.3% |
| | **Lake Chilika** | |
| | 25–29 November 2014 | |
| Regular | 41,272 | |
| Anomaly | 21,692 | |
| Average occurrence | 19.4% | |
| Stand. deviation occurrences | 26.5% | |
| | **Lake Taihu** | |
| | 25–29 August 2016 | 25–29 July 2017 |
| Regular | 631,398 | 1,259,410 |
| Anomaly | 72,399 | 142,685 |
| Average occurrence | 22.6% | 19.1% |
| Stand. deviation occ. | 27.7% | 33.8% |

### *3.3. Experimental Analysis*

Figure 11 shows the evaluated datasets and validation metrics for the trained models ABF-SVM, ABF-RF and ABF-LSTM when predicting algal blooms in all study areas.

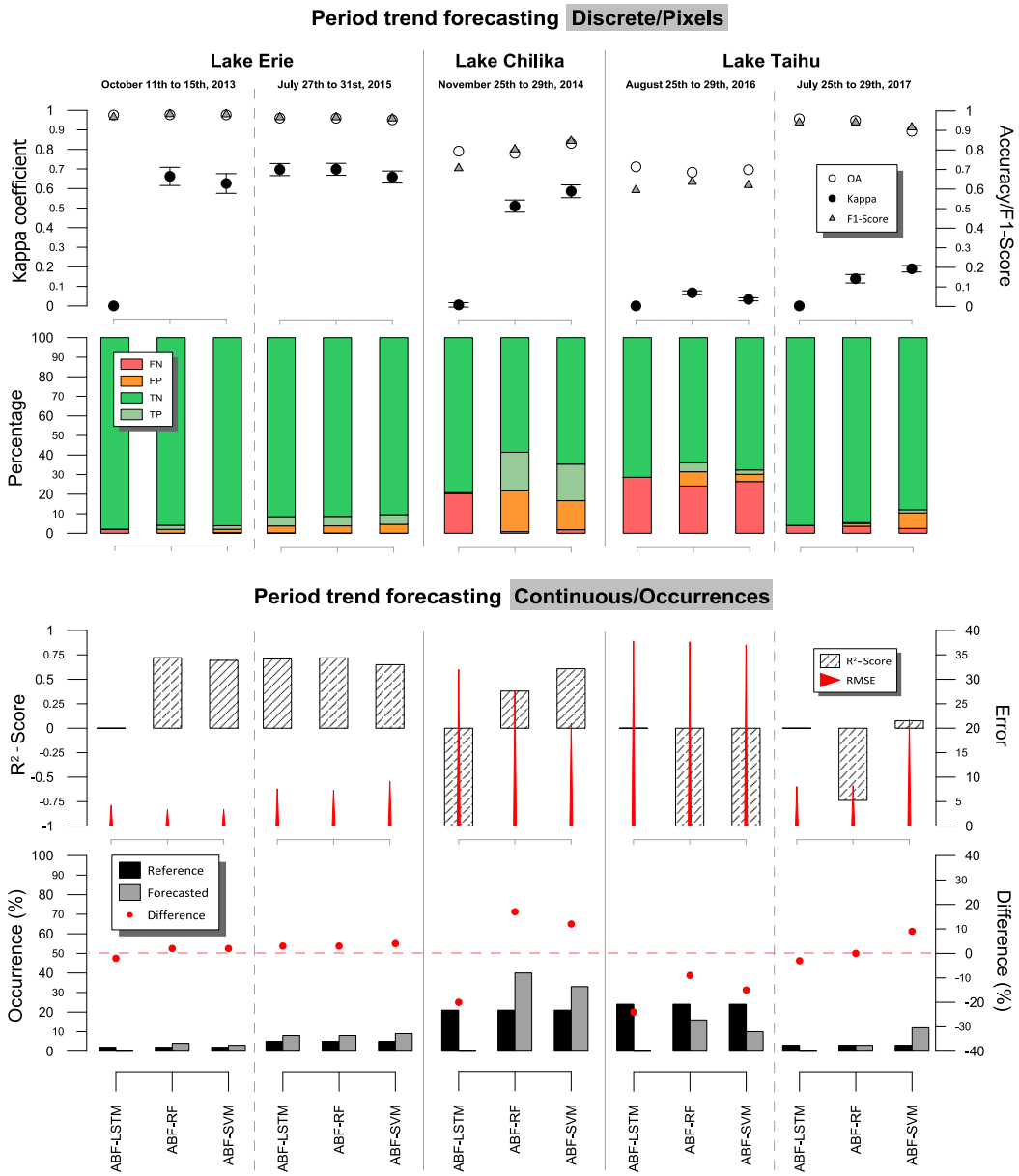


**Figure 11.** Quantitative assessment of the forecasting models. The first plot shows the global accuracy, F1-Scores and kappa coefficient values, whereas the second one shows the performance of each model based on their True(T)-False(F)/Positive(P)-Negative(N) scores. The third plot compares the models by means of $R^2$-Scores and RMSE measures. Finally, the last plot lists the predicted/referenced occurrence percentages. Dashed red line represents the value where the difference was null.

Regarding the first study area (Lake Erie), although the ABF-RF model demonstrated a global accuracy and an F1-Score with a slight difference compared to the ABF-LSTM and ABF-SVM ones, it achieved the best kappa coefficient among all competitors. Note that all forecasting models presented a similar performance regarding the true-positive/false-negative percentages. However, ABF-LSTM did not detect TP values in the first period (11–15 October 2019). As "positive" indicates the occurrence of anomalies (i.e., algal blooms), the opposite was verified for the term "negative", i.e., anomalies were not identified (i.e., regular). Additionally, according to the $R^2$-Scores and RMSE values computed

through the occurrence concentration reference, one can observe that the ABF-RF model produced a high $R^2$-Score, a low RMSE and a good trade-off between the expected (reference) and predicted (forecasted) data. Finally, it is worth mentioning that all trained models were capable of predicting the anomaly occurrence with a low difference w.r.t. the reference data (dashed red line at zero).

Concerning the second study area (Lake Chilika), the ABF-SVM model provided better outcomes than the ABF-RF one, as expressed by the global accuracy, F1-Score and kappa coefficient. Moreover, the ABF-SVM delivered lower FP scores than other models. Analyzing the forecasting results in terms of percentages of occurrences, the ABF-SVM provided more suitable $R^2$-Scores and RMSE values than ABF-RF and ABF-LSTM. On the other hand, one can observe that ABF-LSTM generated predictions with higher negative divergence values w.r.t. omission errors when compared to ABF-RF and ABF-SVM, in which the divergences were positive (i.e., inclusion errors).

For the Lake Taihu study area—specifically, for the first period, 25–29 August 2016—although the ABF-LSTM forecaster produced a high global accuracy and F1-Score, the FN percentage was also high, thus producing a lower kappa. A similar conclusion was observed when analyzing the scores of the ABF-SVM model, except for the FP values. In general, one can see that the ABF-RF model performed better than the others. Regarding the performance analysis in regard to the concentration of occurrences, the negative $R^2$-Score and the high RMSE values corroborated the omission behavior demonstrated by all predictive models. Nonetheless, it is worth pointing out that the ABF-RF model provided the best outcome in comparison to the LSTM- and SVM-based ones.

Now, focusing on the second period (25–29 July 2017) for Lake Taihu, both high global accuracy values and F1-Scores were achieved by all prediction models. However, for the kappa coefficient, the ABF-SVM forecaster performed better than others. Furthermore, one can verify that a higher inclusion error frequency (i.e., FP values) was generated by the SVM-based approach. Considering the occurrence concentration analysis, although the ABF-SVM achieved a slight positive $R^2$-Score, the corresponding RMSE value was higher when compared to those of the other evaluated models.

In Figure 12b—period: 11–15 October 2013, it can be seen that the ABF-LSTM model did not perform satisfactorily w.r.t. the discrete reference, as shown in Figure 12a. On the contrary, the ABF-RF and ABF-SVM forecasters were able to successfully predict the presence of algal blooms in the studied area. Regarding the second period (27–31 July 2015), Figure 13b–d depicts the similar performance observed for all predictive models and reference data (Figure 13a). The occurrence mapping illustrated in Figure 13e–h illustrated the gathered FP errors (Figure 13e). However, note that the ABF tended to be substantially accurate considering a regional context.

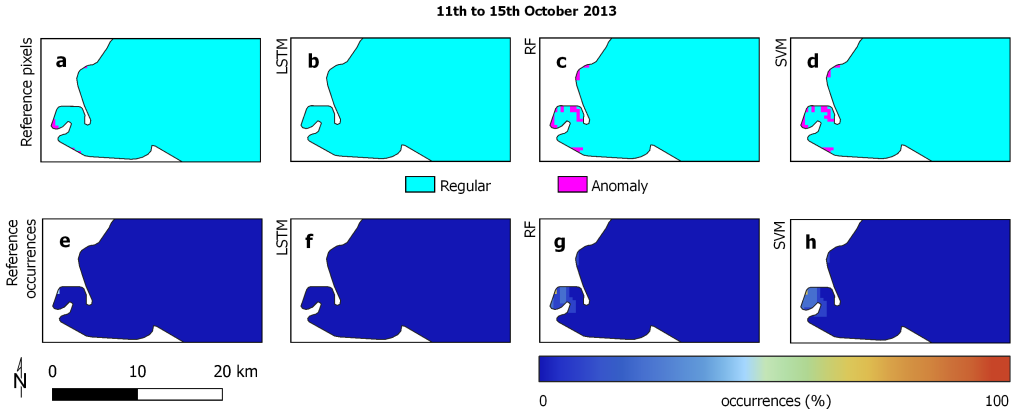


**Figure 12.** Period trend forecasting based on pixels (**a**–**d**) and occurrences (**e**–**h**) from 11–15 October 2013, related to Lake Erie. Plots (**a**,**b**) represent the mapping of the reference data.

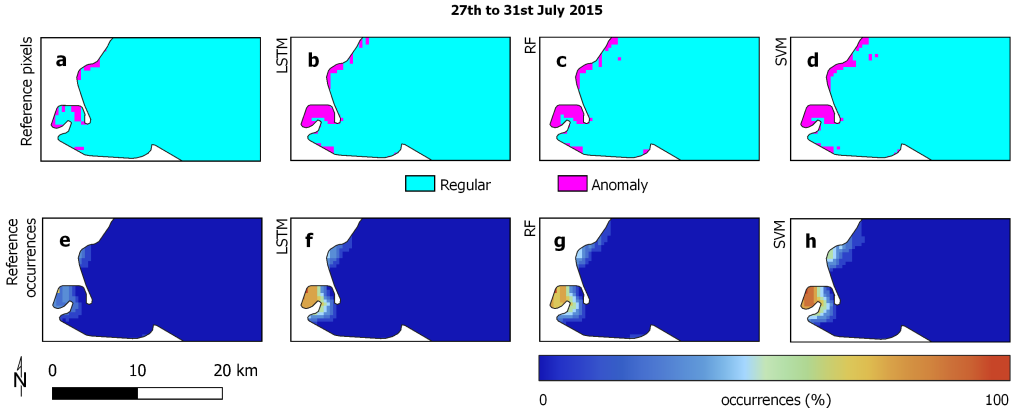


**Figure 13.** Period trend forecasting based on pixels (**a**–**d**) and occurrences (**e**–**h**) from 27–31 July 2015, related to Lake Erie. Plots (**a**,**b**) represent the mapping of the reference data.

Figure 14 presents the forecasting map for Lake Chilika (period: 25–29 November 2014). Although the ABF-LSTM failed in predicting the anomalies (Figure 14b), the RF- (Figure 14c) and SVM-based models (Figure 14d) exhibited more assertive results in both discrete and occurrence concentration representations (Figure 14a,e).

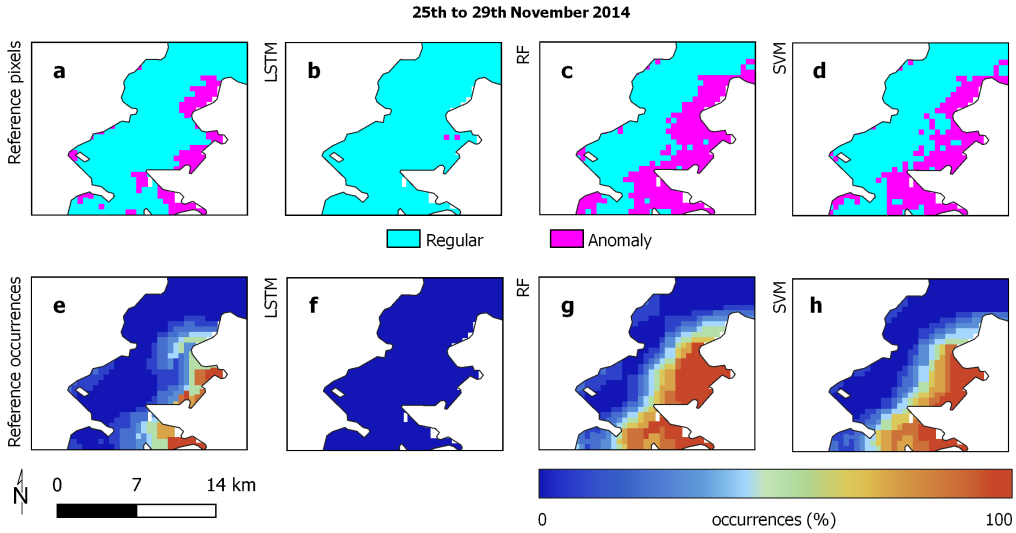


**Figure 14.** Period trend forecasting based on pixels (**a**–**d**) and occurrences (**e**–**h**) from 25–29 November 2014, related to Lake Chilika. Plots (**a**,**b**) represent the mapping of the reference data.

Regarding Lake Taihu—specifically, the first period, 25–29 August 2016—several divergences in performance could be observed between the ABF-derived models (Figure 15b–d) and the reference data (Figure 15a). Although the first image indicated anomalous behavior in the upper left region extending to the center, the ABF-LSTM and ABF-SVM models did not display these pixels. The same was true for the ABF-RF forecaster, although it mapped a larger portion of anomalies in the upper-central region. The same behavior was observed for the concentration occurrence mappings (Figure 15f–h).

As shown in Figure 16a–h, which refer to the second period of the Chinese lake, the performance previously observed through Figure 11 was also confirmed, with the three predictive models giving superior performance in comparison to the first period. A high correlation was achieved for the reference data (Figure 16a) and the ABF-RF model (Figure 16c,g). The ABF-LSTM model did not detect anomalies for this particular study area, whereas the ABF-SVM was capable of successfully accomplishing this task.

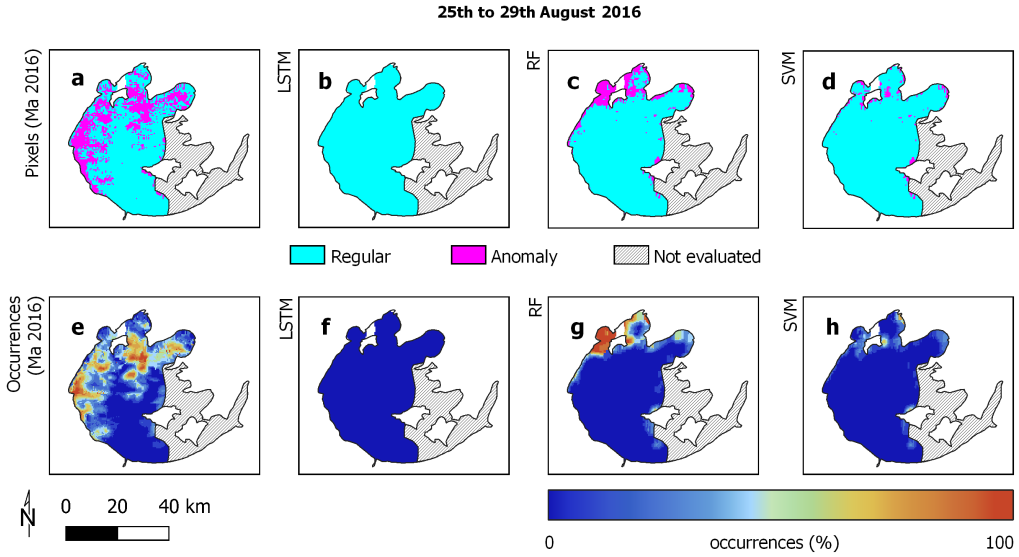


**Figure 15.** Period trend forecasting based on pixels (**a**–**d**) and occurrences (**e**–**h**) from 25–29 August 2016, related to Lake Taihu. Plots (**a**,**b**) represent the mapping of the reference data.

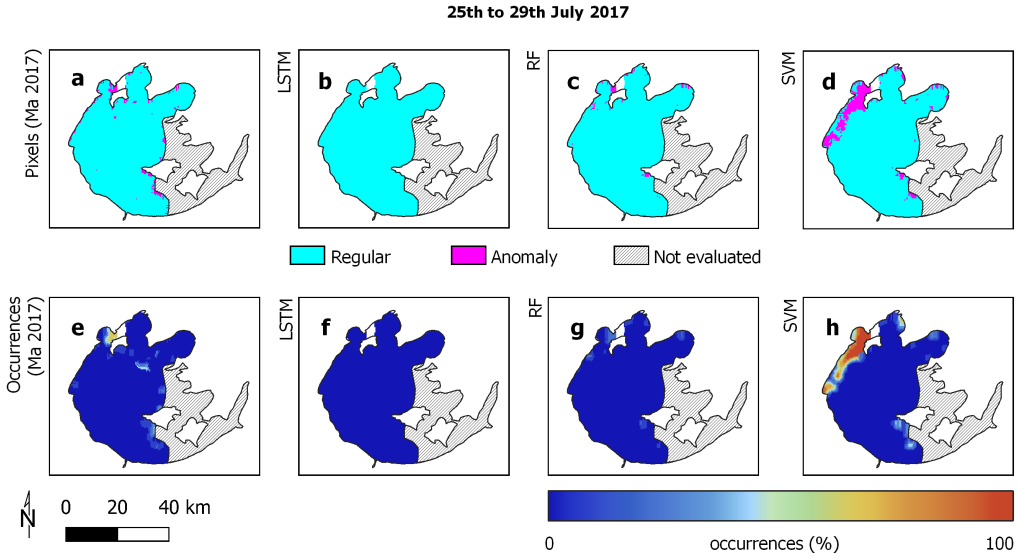


**Figure 16.** Period trend forecasting based on pixels (**a**–**d**) and occurrences (**e**–**h**) from 25–29 July 2017, related to Lake Taihu. Plots (**a**,**b**) represent the mapping of the reference data.

In order to assess the significance of the results, hypothesis tests for the kappa coefficient were performed. For the first period related to the Lake Erie study area, ABF-RF and ABF-SVM performed equivalently, whereas for the second period, all forecasters were statistically equal. Regarding the second study area, the ABF-RF and ABF-SVM models were again statistically equivalent and significant in relation to ABF-LSTM. The results from Lake Taihu demonstrated the superiority of the ABF-RF model; however, the performance of this forecaster was equivalent to that of ABF-SVM in the second period.

Finally, Figure 17 shows the average run-time for each prediction model for all study areas and periods. The ABF-LSTM model had a higher run-time due to the low-convergence behavior commonly obtained by neural network models. In contrast, the ABF-RF model delivered the lowest computation burden for all datasets. Given that the increase in the number of pixels to be processed was proportional to the study area dimensions, the cost assigned to Lake Taihu tended to be higher than in other areas.

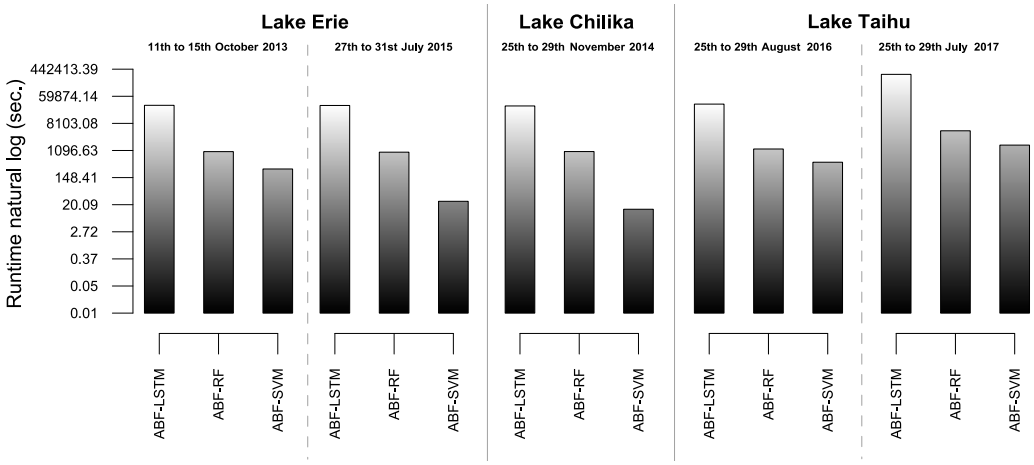


**Figure 17.** Computational run-times (average) for the study areas and evaluated periods.

### *3.4. Discussion*

Algal blooms are a very challenging issue, discussed in a plethora of studies, mainly due to their harmful effect on human beings and society. In the absence of ground-truth terrestrial data, the use of remote sensing data has arisen as a convenient tool in detecting and tracking algae blooms. Therefore, in this paper we have presented the ABF method, which has been designed to predict such events by applying machine learning methods to features extracted from remotely sensed image series and environmental data. Additionally, the proposed approach is fully automatic and can act as a core component in environmental monitoring systems, generating daily estimates based on the environmental behavior observed in past data and with a self-adapting ability to reflect local and temporal dynamics.

Based on the formalization of our methodology as given in Section 2.3.1, the ABF framework was validated in three different study areas and periods. The SVM, RF and LSTM methods were coupled as machine learning components to the ABF design. By performing an extensive battery of experiments, we verified that the combination ABF-RF was capable of generating predictions that accurately matched the expected results. The consistency of the predictions was demonstrated by computing several well-established metrics that used official data (NESSDC—Lake Taihu) or estimates derived from an ensemble of methodologies that are well-documented in the literature (Lakes Erie and Chilka).

Considering the obtained prediction maps (Figures 12–16), it is worth mentioning that the ABF approach did not produce biased results in terms of false positive errors. Moreover, the resulting maps were spatially consistent, as predictions/detections were not observed in portions of the study area not affected by algal blooms.

From an environmental point of view, the obtained results corroborate previous studies that discussed the insurgency of algae in several study areas. Concordantly, these studies highlight the presence of typical components, mainly phosphorus and nitrogen, as substantial factors in regard to algal insurgence events. Particularly, Kane et al. [78] and Scavia et al. [79] reported that Lake Erie's northwest-west-southwest contour has been a site of soluble reactive phosphorus discharge through the primary effluent from the Maumee River basin, and this was highlighted as one of the main factors responsible for algal blooms in this region. Similarly, Barik et al. [80] found that algal blooms in Lake Chilika occurred due to the local phosphorous concentration. Furthermore, the authors showed that there was nutrient migration from the northeast to southwest lake regions in the pre- and post-monsoon periods, respectively. Since the area and period considered in our analysis (Figures 5, 8 and 14) included the southwest and post-monsoon portion, we verified that the predictions generated by the ABF method were compatible with the reported local dynamics. Finally, Liu et al. [72] measured the usual concentrations of Chl-a in the northern and western portions of Lake Taihu during summer (June–August) and autumn (i.e., September–November), caused by the entry of nutrients through adjacent rivers, meaning that the main concentrationsof Chl-a were present in the northern-western

portion. This finding was also observed in the results depicted in Figures 15 and 16, thus confirming the accuracy of our method's predictions from an environmental perspective.

## 4. Conclusions

Forecasting the proliferation of potentially harmful algal blooms in water bodies is an essential task for fauna and flora preservation. Considering this environmental issue, in this study we introduced a fully unsupervised methodology for the prediction of algal blooms that relies on remotely sensed image series, climatic data and machine-learning-based designs. Three study cases were used to validate and discuss the proposed methodology. By making use of different image classification methods, in particular, SVM, RF and LSTM, our approach was capable of learning information about the transient occurrence of algae in a fully unsupervised fashion.

The prediction results and evaluation metrics demonstrated that the proposed methodology was capable of accurately forecasting the algal bloom phenomena for several lake regions, covering different levels of complexity and specificity. Regarding the evaluated forecasters, ABF-RF delivered more regular outputs for the three analyzed study areas, obtaining average scores of 95%, 94%, 51%, 0.38 and 8.22 for the validation metrics of overall accuracy, F1-Score, kappa coefficient, $R^2$-Score and RMSE, respectively, compared to 90%, 92%, 51%, 0.61 and 20.98 for ABF-SVM and 95%, 93%, 0%, 0.0 and 8.05 for ABF-LSTM.

Despite its robustness, accuracy and adaptability, there are a few aspects that must be observed when using the ABF methodology. First, ABF performs best for images with a low occurrence of cloud, as the temporal profile of algal proliferation tends to be uninformative for images with substantial cloud incidence. The dimensions of the study area could lead to an increased computational cost when training the machine learning models. Finally, Internet availability is another issue, as the method requires updated meteorological/climatic products.

In future works, we plan to (i) take other classification models as part of our ABF methodology; (ii) evaluate the use of the proposed methodology in other study areas; (iii) apply our learning apparatus to new application domains such as forest burning, deforestation and flooding; (iv) perform further investigations aiming at the early identification of the most relevant spectral and environmental variables, skipping some costly steps of the method so as to reduce its computational burden; and (v) develop new strategies to store image time series on local databases, thus enabling cases in which the internet is not available.

**Author Contributions:** Conceptualization—P.H.M.A., R.G.N. and W.C.; funding acquisition—R.G.N. and E.A.S.; investigation—P.H.M.A. and R.G.N.; methodology—P.H.M.A., R.G.N., M.A.D. and W.C.; validation—P.H.M.A., R.G.N., M.A.D., E.A.S. and W.C.; writing—original draft—P.H.M.A., R.G.N., M.A.D., E.A.S. and W.C.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the São Paulo Research Foundation (FAPESP), grants 2021/01305-6, 2021/03328-3 and 2016/24185-8, and National Council for Scientific and Technological Development (CNPq), grants 427915/2018-0, 304402/2019-2 and 316228/2021-4. The APC was partially funded by São Paulo State University (UNESP).

**Data Availability Statement:** The code of the proposed framework is freely available at https://github.com/pedroananias/abf.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Yang, K.; Luo, Y.; Chen, K.; Yang, Y.; Shang, C.; Yu, Z.; Xu, J.; Zhao, Y. Spatial–temporal variations in urbanization in Kunming and their impact on urban lake water quality. *Land Degrad. Dev.* **2020**, *31*, 1392–1407. [CrossRef]
2. Chawla, I.; Karthikeyan, L.; Mishra, A.K. A review of remote sensing applications for water security: Quantity, quality, and extremes. *J. Hydrol.* **2020**, *585*, 124826.

3. Mishra, A.K.; Coulibaly, P. Developments in hydrometric network design: A review. *Rev. Geophys.* **2009**, *47*, 1–24.
4. Wells, M.L.; Trainer, V.L.; Smayda, T.J.; Karlson, B.S.; Trick, C.G.; Kudela, R.M.; Ishikawa, A.; Bernard, S.; Wulff, A.; Anderson, D.M.; et al. Harmful algal blooms and climate change: Learning from the past and present to forecast the future. *Harmful Algae* **2015**, *49*, 68–93.
5. Shi, K.; Zhang, Y.; Qin, B.; Zhou, B. Remote sensing of cyanobacterial blooms in inland waters: Present knowledge and future challenges. *Sci. Bull.* **2019**, *64*, 1540–1556.
6. Gons, H.J. Optical Teledetection of Chlorophyll a in Turbid Inland Waters. *Environ. Sci. Technol.* **1999**, *33*, 1127–1132.
7. Rousso, B.Z.; Bertone, E.; Stewart, R.; Hamilton, D.P. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. *Water Res.* **2020**, *182*, 115959.
8. Qi, L.; Hu, C.; Duan, H.; Barnes, B.B.; Ma, R. An EOF-Based Algorithm to Estimate Chlorophyll a Concentrations in Taihu Lake from MODIS Land-Band Measurements: Implications for Near Real-Time Applications and Forecasting Models. *Remote Sens.* **2014**, *6*, 10694–10715.
9. Allen, J.I.; Smyth, T.J.; Siddorn, J.R.; Holt, M. How well can we forecast high biomass algal bloom events in a eutrophic coastal sea? *Harmful Algae* **2008**, *8*, 70–76.
10. Hu, C. A novel ocean color index to detect floating algae in the global oceans. *Remote Sens. Environ.* **2009**, *113*, 2118–2129.
11. Mishra, S.; Mishra, D.R. Normalized Difference Chlorophyll Index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. *Remote Sens. Environ.* **2012**, *117*, 394–406.
12. Zhang, Y.; Ma, R.; Duan, H.; Loiselle, S.A.; Xu, J.; Ma, M. A novel algorithm to estimate algal bloom coverage to subpixel resolution in Lake Taihu. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 3060–3068.
13. Houborg, R.; McCabe, M.F.; Angel, Y.; Middleton, E.M. Detection of chlorophyll and leaf area index dynamics from sub-weekly hyperspectral imagery. In Proceedings of the Remote Sensing for Agriculture, Ecosystems, and Hydrology XVIII, International Society for Optics and Photonics, Edinburgh, UK, 26–29 September 2016; Volume 9998, p. 999812.
14. Watanabe, F.; Alcantara, E.; Rodrigues, T.; Rotta, L.; Bernardo, N.; Imai, N. Remote sensing of the chlorophyll-a based on OLI/Landsat-8 and MSI/Sentinel-2A (Barra Bonita reservoir, Brazil). *An. Acad. Bras. Ciências* **2018**, *90*, 1987–2000.
15. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens. Environ.* **2020**, *241*, 111716.
16. Martínez-Álvarez, F.; Tien Bui, D. Advanced Machine Learning and Big Data Analytics in Remote Sensing for Natural Hazards Management. *Remote Sens.* **2020**, *12*, 301.
17. Zhang, T.; Huang, M.; Wang, Z. Estimation of chlorophyll-a Concentration of lakes based on SVM algorithm and Landsat 8 OLI images. *Environ. Sci. Pollut. Res.* **2020**, *27*, 14977–14990.
18. Ananias, P.H.M.; Negri, R.G. Anomalous behaviour detection using one-class support vector machine and remote sensing images: A case study of algal bloom occurrence in inland waters. *Int. J. Digit. Earth* **2021**, *14*, 921–942.
19. Silveira Kupssinskü, L.; Thomassim Guimarães, T.; Menezes de Souza, E.; Zanotta, C.D.; Roberto Veronez, M.; Gonzaga, L.; Mauad, F.F. A Method for Chlorophyll-a and Suspended Solids Prediction through Remote Sensing and Machine Learning. *Sensors* **2020**, *20*, 2125.
20. Cho, H.; Choi, U.; Park, H. Deep learning application to time-series prediction of daily chlorophyll-a concentration. *WIT Trans. Ecol. Environ.* **2018**, *215*, 157–163.
21. Lee, M.S.; Park, K.A.; Chae, J.; Park, J.E.; Lee, J.S.; Lee, J.H. Red tide detection using deep learning and high-spatial resolution optical satellite imagery. *Int. J. Remote Sens.* **2020**, *41*, 5838–5860.
22. Barzegar, R.; Aalami, M.T.; Adamowski, J. Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 415–433.
23. Yu, Z.; Yang, K.; Luo, Y.; Shang, C. Spatial-temporal process simulation and prediction of chlorophyll-a concentration in Dianchi Lake based on wavelet analysis and long-short term memory network. *J. Hydrol.* **2020**, *582*, 124488. [CrossRef]
24. Körting, T.S.; Fonseca, L.M.G.; Castejon, E.F.; Namikawa, L.M. Improvements in Sample Selection Methods for Image Classification. *Remote Sens.* **2014**, *6*, 7580–7591.
25. Wang, X.; Yan, H.; Huo, C.; Yu, J.; Pant, C. Enhancing Pix2Pix for Remote Sensing Image Classification. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2332–2336.
26. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
27. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
28. Vapnik, V. *Estimation of Dependences Based on Empirical Data*; Springer: Berlin/Heidelberg, Germany, 1982. [CrossRef]
29. Mountrakis, G.; Im, J.; Ogole, C. Support Vector Machines in Remote Sensing: A review. *ISPRS J. Photogramm. Remote Sens. Soc.* **2011**, *66*, 247–259.
30. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, UK, 2004.
31. Bruzzone, L.; Persello, C. A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabeled Training Samples. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2142–2154.
32. Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*, 4th ed.; Academic Press: San Diego, CA, USA, 2008; p. 984.
33. Dietterich, T.G. Ensemble learning. In *The Handbook of Brain Theory and Neural Networks*; MIT Press: Cambridge, MA, USA, 2002; Volume 2, pp. 110–125.
34. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

35. McFeeeters, S.K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **1996**, *17*, 1425–1432.
36. Gao, B.C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **1996**, *58*, 257–266.
37. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033.
38. Rouse, J.W., Jr.; Haas, R.H.; Schell, J.; Deering, D. *Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation*; Texas A&M University: College Station, TX, USA, 1973.
39. Huete, A. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309.
40. Liu, H.Q.; Huete, A. A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. *IEEE Trans. Geosci. Remote Sens.* **1995**, *33*, 457–465.
41. Wang, Z.; Liu, C.; Huete, A. From AVHRR-NDVI to MODIS-EVI: Advances in vegetation index research. *Acta Ecol. Sin.* **2003**, *23*, 979–987.
42. Zhao, D. Application of NDVI to detecting algal bloom in the Bohai Sea of China from AVHRR. In *Ocean Remote Sensing and Applications, Proceedings of the Third International Asia-Pacific Environmental Remote Sensing Remote Sensing of the Atmosphere, Ocean, Environment, and Space, Hangzhou, China, 24–26 October 2002* ; SPIE: Bellingham, WA, USA, 2003; Volume 4892, pp. 241–246.
43. Han-Qiu, X. A study on information extraction of water body with the modified normalized difference water index (MNDWI). *J. Remote Sens.* **2005**, *5*, 589–595.
44. Alawadi, F. Detection of surface algal blooms using the newly developed algorithm surface algal bloom index (SABI). In *Remote Sensing of the Ocean, Sea Ice, and Large Water Regions 2010, Proceedings of the SPIE Remote Sensing, Toulouse, France, 20 September 2010,* SPIE: Bellingham, WA, USA, 2010; Volume 7825, pp. 45–58.
45. Oyama, Y.; Fukushima, T.; Matsushita, B.; Matsuzaki, H.; Kamiya, K.; Kobinata, H. Monitoring levels of cyanobacterial blooms using the visual cyanobacteria index (VCI) and floating algae index (FAI). *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *38*, 335–348.
46. Rodell, M.; Houser, P.R.; Jambor, U.; Gottschalck, J.; Mitchell, K.; Meng, C.J.; Arsenault, K.; Cosgrove, B.; Radakovich, J.; Bosilovich, M.; et al. The Global Land Data Assimilation System. *Bull. Am. Meteorol. Soc.* **2004**, *85*, 381–394.
47. Lehner, B.; Verdin, K.; Jarvis, A. New Global Hydrography Derived From Spaceborne Elevation Data. *Eos Trans. Am. Geophys. Union* **2008**, *89*, 93–94.
48. Takaku, J.; Tadono, T.; Doutsu, M.; Ohgushi, F.; Kai, H. Updates OF 'AW3D30' Alos Global Digital Surface Model with Other Open Access Datasets. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *43*, 183–189.
49. Wan, Z.; Hook, S.; Hulley, G. *MOD11A1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006*; Type: Dataset; NASA EOSDIS Land Processes DAAC: Sioux Falls, SD, USA, 2015.
50. McKinney, W. *Python for Data Analysis: Data Wrangling with PANDAS, NumPy, and IPython*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2012.
51. van Rossum, G.; Drake, F.L. *The Python Language Reference Manual*; Network Theory Ltd. : Godalming, UK, 2011.
52. van der Walt, S.; Colbert, S.C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.
53. McKinney, W. Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; Volume 445, pp. 51–56.
54. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
55. Ketkar, N. Introduction to keras. In *Deep learning with Python*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 97–111.
56. GEE-API. Google Earth Engine API. 2019. Available online: https://developers.google.com/earth-engine (accessed on 24 July 2022).
57. USGS. MODIS/Terra Surface Reflectance Daily L2G Global 1 km and 500 m. 2020. Available online: https://lpdaac.usgs.gov/products/mod09gav006/ (accessed on 24 July 2022).
58. Rastrigin, L. The convergence of the random search method in the extremal control of a many parameter system. *Autom. Remote Control.* **1963**, *24*, 1337–1342.
59. Baba, N. Convergence of a random optimization method for constrained optimization problems. *J. Optim. Theory Appl.* **1981**, *33*, 451–461.
60. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
61. Jolliffe, I. Principal component analysis. *Encycl. Stat. Behav. Sci.* **2005** . [CrossRef]
62. Congalton, R.G.; Green, K. *Assessing the Accuracy of Remotely Sensed Data*; CRC Press: Boca Raton, FL, USA, 2009; p. 183.
63. van Rijsbergen, C.J. *Information Retrieval*, 2nd ed.; Butterworths: London, UK, 1979.
64. Draper, N.R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 1998; Volume 326.
65. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
66. Bolsenga, S.J.; Herdendorf, C.E. *Lake Erie and Lake St. Clair Handbook*; Wayne State University Press: Detroit, MI, USA, 1993.
67. Zhu, M.; Zhu, G.; Zhao, L.; Yao, X.; Zhang, Y.; Gao, G.; Qin, B. Influence of algal bloom degradation on nutrient release at the sediment–water interface in Lake Taihu, China. *Environ. Sci. Pollut. Res.* **2012**, *20*, 1803–1811.

68. Stumpf, R.P.; Wynne, T.T.; Baker, D.B.; Fahnenstiel, G.L. Interannual Variability of Cyanobacterial Blooms in Lake Erie. *PLoS ONE* **2012**, *7*, e42444.
69. Sengupta, M.; Anandurai, R.; Nanda, S.; Datti, A.A. Geospatial identification of algal blooms in inland waters: A post cyclone case study of Chilka Lake, Odisha, India. *RASAYAN J. Chem.* **2017**, *10*, 234–239.
70. Panigrahi, J.K. Water Quality, Biodiversity and Livelihood Issues: A Case Study of Chilika Lake, India. In Proceedings of the 2007 Atlanta Conference on Science, Technology and Innovation Policy, Atlanta, GA, USA, 19–20 October 2007; pp. 1–6.
71. Ranjan, R. A forestry-based PES mechanism for enhancing the sustainability of Chilika Lake through reduced siltation loading. *For. Policy Econ.* **2019**, *106*, 101944. [CrossRef]
72. Liu, L.; Dong, Y.; Kong, M.; Zhou, J.; Zhao, H.; Wang, Y.; Zhang, M.; Wang, Z. Towards the comprehensive water quality control in Lake Taihu: Correlating chlorphyll a and water quality parameters with generalized additive model. *Sci. Total Environ.* **2020**, *705*, 135993. [CrossRef] [PubMed]
73. Gao, Y.; Zhu, G.; Paerl, H.W.; Qin, B.; Yu, J.; Song, Y. A study of bioavailable phosphorus in the inflowing rivers of Lake Taihu, China. *Aquat. Sci.* **2020**, *82*, 1.
74. Ma, R. *Lake Taihu Chlorophyll Inversion Product Data Set (2016)*; National Earth System Science Data Center, National Science and Technology Infrastructure of China: Beijing, China, 2016.
75. Ma, R. *Lake Taihu Chlorophyll Inversion Product Data Set (2017)*; National Earth System Science Data Center, National Science and Technology Infrastructure of China: Beijing, China, 2017.
76. Xu, H.; Paerl, H.W.; Qin, B.; Zhu, G.; Hall, N.S.; Wu, Y. Determining Critical Nutrient Thresholds Needed to Control Harmful Cyanobacterial Blooms in Eutrophic Lake Taihu, China. *Environ. Sci. Technol.* **2015**, *49*, 1051–1059. [CrossRef] [PubMed]
77. Wang, M.; Strokal, M.; Burek, P.; Kroeze, C.; Ma, L.; Janssen, A.B. Excess nutrient loads to Lake Taihu: Opportunities for nutrient reduction. *Sci. Total Environ.* **2019**, *664*, 865–873. [CrossRef] [PubMed]
78. Kane, D.D.; Conroy, J.D.; Richards, R.P.; Baker, D.B.; Culver, D.A. Re-eutrophication of Lake Erie: Correlations between tributary nutrient loads and phytoplankton biomass. *J. Great Lakes Res.* **2014**, *40*, 496–501. [CrossRef]
79. Scavia, D.; Allan, J.D.; Arend, K.K.; Bartell, S.; Beletsky, D.; Bosch, N.S.; Brandt, S.B.; Briland, R.D.; Daloğlu, I.; DePinto, J.V.; et al. Assessing and addressing the re-eutrophication of Lake Erie: Central basin hypoxia. *J. Great Lakes Res.* **2014**, *40*, 226–246. [CrossRef]
80. Barik, S.K.; Bramha, S.N.; Mohanty, A.K.; Bastia, T.K.; Behera, D.; Rath, P. Sequential extraction of different forms of phosphorus in the surface sediments of Chilika Lake. *Arab. J. Geosci.* **2016**, *9*, 135. [CrossRef]