




## Article

# A Remote Sensing and Machine Learning-Based Approach to Forecast the Onset of Harmful Algal Bloom

Moein Izadi <sup>1</sup>, Mohamed Sultan <sup>1,\*</sup>, Racha El Kadiri <sup>2</sup>, Amin Ghannadi <sup>3</sup> and Karem Abdelmohsen <sup>1,4</sup>

<sup>1</sup> Department of Geological and Environmental Sciences, Western Michigan University, Kalamazoo, MI 49008, USA; moein.izadi@wmich.edu (M.I.); karem.abdelmohsen@wmich.edu (K.A.)

<sup>2</sup> Department of Geosciences, Middle Tennessee State University, Murfreesboro, TN 37132, USA; racha.elkadiri@mtsu.edu

<sup>3</sup> Department of Surveying Engineering, Arak University of Technology, Arak 39455-38138, Iran; m.ghannadi@arakut.ac.ir

<sup>4</sup> Geodynamics Department, National Research Institute of Astronomy and Geophysics (NRIAG), Helwan, Cairo 11421, Egypt

\* Correspondence: mohamed.sultan@wmich.edu

**Abstract:** In the last few decades, harmful algal blooms (HABs, also known as “red tides”) have become one of the most detrimental natural phenomena in Florida’s coastal areas. *Karenia brevis* produces toxins that have harmful effects on humans, fisheries, and ecosystems. In this study, we developed and compared the efficiency of state-of-the-art machine learning models (e.g., XGBoost, Random Forest, and Support Vector Machine) in predicting the occurrence of HABs. In the proposed models the *K. brevis* abundance is used as the target, and 10 level-02 ocean color products extracted from daily archival MODIS satellite data are used as controlling factors. The adopted approach addresses two main shortcomings of earlier models: (1) the paucity of satellite data due to cloudy scenes and (2) the lag time between the period at which a variable reaches its highest correlation with the target and the time the bloom occurs. Eleven spatio-temporal models were generated, each from 3 consecutive day satellite datasets, with a forecasting span from 1 to 11 days. The 3-day models addressed the potential variations in lag time for some of the temporal variables. One or more of the generated 11 models could be used to predict HAB occurrences depending on availability of the cloud-free consecutive days. Findings indicate that XGBoost outperformed the other methods, and the forecasting models of 5–9 days achieved the best results. The most reliable model can forecast eight days ahead of time with balanced overall accuracy, Kappa coefficient, F-Score, and AUC of 96%, 0.93, 0.97, and 0.98 respectively. The euphotic depth, sea surface temperature, and chlorophyll-a are always among the most significant controlling factors. The proposed models could potentially be used to develop an “early warning system” for HABs in southwest Florida.

**Keywords:** harmful algal bloom forecasting; remote sensing; data mining; machine learning



**Citation:** Izadi, M.; Sultan, M.; Kadiri, R.E.; Ghannadi, A.; Abdelmohsen, K. A Remote Sensing and Machine Learning-Based Approach to Forecast the Onset of Harmful Algal Bloom. *Remote Sens.* **2021**, *13*, 3863. <https://doi.org/10.3390/rs13193863>

Academic Editor: Raphael M. Kudela

Received: 5 August 2021

Accepted: 22 September 2021

Published: 27 September 2021

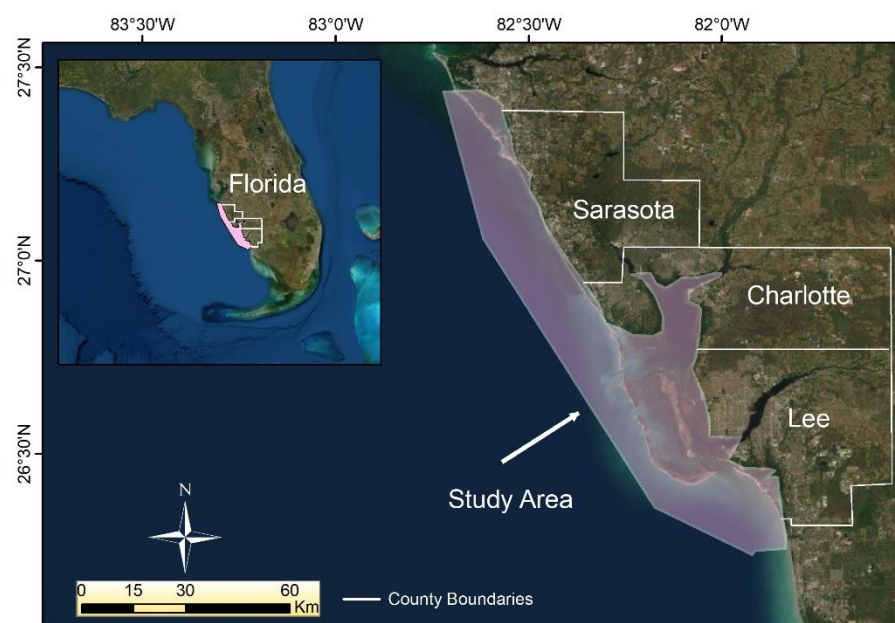
**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Harmful algal blooms (HABs) in saline waters, often referred to as “red tide” events, have been reported in many areas around the world, including the Gulf of St. Lawrence in Canada, Tampa Bay in the Gulf of Mexico, the Bay of Bengal, the Arabian Sea, the Bay of Biscay in Spain, Paracas Bay in Peru, Lisbon Bay in Portugal, and the Persian Gulf [1–8]. In the USA, red tides have been reported in many locations, including the Gulf Coast of Florida, the Gulf of Maine, and Monterey Bay in California [9–12]. The reddish color of the ocean water is most often caused by the proliferation of a microscopic photosynthetic organism called *Karenia brevis* (formerly known as *Gymnodinium breve* and *Ptychodiscus brevis*) [13–15]. Over the last few decades, *K. brevis* has become the predominant HAB phytoplankton species among 5000 known species and one of the most harmful natural phenomena in many areas in the Gulf of Mexico [13,16,17] including our study area, Charlotte County in southwest Florida (Figure 1) [18].



**Figure 1.** Location map for the study area covering coastal waters (width: ~10–60 km; length: 180 km) of Charlotte County in southwest Florida.

HABs can significantly change the quality (i.e., color, odor, and taste) of bodies of water, adversely affect fresh and saltwater ecosystems, and produce neurotoxins called brevetoxins [19] that have harmful effects on human health, the fishing industry, marine mammals, seabirds, and local economies [20–22]. These algae-related adverse effects require local authorities to take costly measurements and/or remedies including closure of beaches and conducting costly filtration processes and decontamination activities [23,24]. Several environmental factors contribute to the growth and propagation of algae, such as nutrients introduced from agricultural activities, lighting condition (low irradiance), salinity, and water temperature [9,25]. *K. brevis* consumes both inorganic and organic nitrogen and phosphorus compounds [26]. Nitrogen-based fertilizers are the main source of nitrogen in the Gulf of Mexico [27,28], where the nutrients of these fertilizers are transported from the agricultural fields within the Mississippi-Atchafalaya River basin [29] into prone water bodies (e.g., bay areas) by surface runoff and infiltration of nutrient-rich waters and groundwater flow towards neighboring water bodies [30,31].

There has been a long standing desire, and a need for, forecasting and mapping HABs [24] given their adverse effects on human health [21,22] and on the biodiversity and habitats of aquatic ecosystems [16,17]. The majority of earlier attempts to detect, map, and forecast HABs can be lumped in two groups: ones that rely heavily on the utilization of satellite remotely acquired data and ones that do not [9,32]. The latter research activities entail the acquisition of in situ real-time field monitoring of relevant parameters, such as chlorophyll-a concentration, dissolved oxygen, and nutrients [33]. Real-time nucleic acid sequence-based amplification assays and simple test kits have been used to detect and quantify the red tide dinoflagellate *K. brevis* [32]. The HAB Program of Florida Fish and Wildlife Research Institute (FWC) is one such program that designs and employs light and electron microscopy and genetic tools to identify and/or quantify HABs [24]. Hydro-meteorological variables (e.g., sea surface temperature [SST], wind speed, cloud amount, salinity, and rainfall) were used in statistical models (fuzzy reasoning and the ensemble method classifiers) to predict HABs occurrences [34,35].

Additional approaches involved the construction of wind-driven models or three-dimensional physical hydrodynamic models to forecast the dominant regional physical processes that result in water exchange events and bloom propagation [33,34]. Most of these models were designed for same-day mapping and modeling the onset of blooms. Although the above-mentioned field-based approaches have been shown to be successful

in detecting HABs [24], their application in many parts of the world has been hindered by their spatiotemporal limitations, high cost, and labor-intensive operational procedures [36].

The footprint of many of the remote-sensing sensors cover large areas with high temporal resolution; thus, they can potentially capture the spatial and temporal variabilities of HABs, as evidenced by the extensive literature describing the detection, monitoring, and forecasting of HABs using remote sensing-based techniques and sensors [9]. Investigations utilizing moderate-resolution imaging spectroradiometers (MODIS-Aqua and MODIS-Terra), SeaWiFS, MERIS, Sentinel-2, and unmanned aerial vehicles have contributed the most to these studies [9,37–42]. The more recent and advanced satellites (e.g., Sentinel-3, launched in February 2016) provide added valuable resources for ocean color products, yet their recent deployment and, hence, their short record of historical data compared to earlier operational satellites (e.g., MODIS: 1999–present) puts them on the waiting list for future machine learning-based forecasting projects.

Many of the earlier attempts for HAB detection used reflectance-based classification algorithms and targeted chlorophyll-a, a good proxy for phytoplankton biomass [43–45]. These include classifications based on chlorophyll-a concentration [46], band ratios (e.g., blue–green band ratios) [47], ocean color band difference algorithms such as fluorescence line height (FLH), and maximum chlorophyll index. Chlorophyll-a concentrations derived from Landsat-8 (OLI) images over inland lakes in China using machine learning techniques (XGBoost) were shown to be more reliable than outputs from band ratio algorithms [48]. A comprehensive review of all these remote sensing-based methods was compiled by [43]. The use of these simple and straightforward indices and measurements, although successful, often introduces uncertainties, including false positive detections [49].

One approach to reduce these false positives is to develop statistical models that use more of the available ocean color products [50]. Using remotely sensed data, a number of machine learning studies were conducted to detect, monitor, and forecast HABs. Applying artificial neural networks and multiple linear regression (MLR) techniques in Kuwait Bay, a hybrid method showed a correlation between a variety of spatial and temporal ocean color products and HAB propagation and growth in the bay [50]. In early machine learning (ML) studies, and using remote sensing data over Monterey Bay, random forest (RF) and support vector machines (SVMs) were applied to build a decision support system for predicting the distribution of algal blooms in the bay [51]. A machine learning-based spatio-temporal data mining approach using kernel-based SVM was applied to detect HAB events in the Gulf of Mexico [52]. In a red tide detection study, a deep learning method was applied to Landsat 8 Operational Land Imager data acquired over the southern coastal region of the Korean Peninsula [53]. Additionally, in a recent study, spatiotemporal SVM, RF, and deep learning long- and short-term memory methods were adopted to develop an HAB detection and forecasting system for the whole west coast of Florida [54]. These methods apply a state-of-the-art machine learning algorithm; however, most of them use only a limited number of variables (one to five) and do not consider as one of the main targets of their investigations the lag time between the onset of a bloom and the time it takes for a variable to have a maximum impact on bloom propagation.

MODIS-derived ocean color products, along with field data, were used to develop data-driven statistical models based on MLR expressions to identify factors controlling HAB propagation [55] and to forecast bloom occurrences up to three days in advance. These models assumed a unified lag time for the significant variables, an assumption that does not adequately portray the complex interactions between the controlling factors, leading to the propagation of the HABs [56–58]. Addressing this problem will lead to the development of more realistic modeling structures that can better account for the HAB growth patterns [59]. This could be accomplished by allowing each of the independent variables to have different lag times and the model to select the significant variables, each with its optimum lag time. In practice, the more satellite data and lag time choices we provide, the better and more comprehensive the model.

There are advantages for selecting statistical models that portray the relative significance of, and the correlation between, the factors controlling the onset of HABs. For example, artificial neural networks and deep learning (DL) models are powerful functions for modeling real-world problems [60,61]. The growth of HABs was successfully predicted using historical data and ecological informatics and applying DL methods [62]. The DL methods were also used to predict algal growth in rivers [63–66], lakes [67,68], and coastal areas [69,70]. However, these methods function as black boxes, and the neuron connections, their weights, and different layers cannot be associated much with the concept of the physical problem at hand [71]. The paucity of data has always been one of the challenges facing researchers—a good model should work and be trained with limited datasets [71,72]. Different machine learning models (e.g., linear versus non-linear and tree-based versus non-tree-based models) need to be adopted to compare and contrast the results in terms of consistency and model performance. The proposed approach addresses the aforementioned shortcomings, and provides an adequate forecasting period (up to 9 days) because of its spatio-temporal features [9].

In this manuscript, we first demonstrate the enhanced predictive power of the statistical model when: (1) multiple day (>2 days) satellite data acquisitions are utilized instead of single and 2-day models, (2) the optimum forecasting period is identified and the variations in lag times for the independent variables are accommodated, and (3) multiple statistical models are tested and the optimum predictive model is selected. In light of our findings, we then identify and use the optimum predictive statistical model and data structure to develop multiple sequential forecasting models that utilize available cloud-free scenes that span a period ranging from 1 to 11 days ahead of the onset of the HAB bloom. In doing so, our approach addresses the paucity and temporal discontinuity of satellite ocean color products due to cloud coverage or missing values in areas close to shorelines due to masking and processing data (from levels 0 to 2) and random and systematic errors during data acquisition. Additionally, each of the individual models allows for variations in lag times of up to 2 days for the independent variables. In addition, we utilize statistical models that portray the relative significance of the factors controlling the onset of HABs.

## 2. Study Area

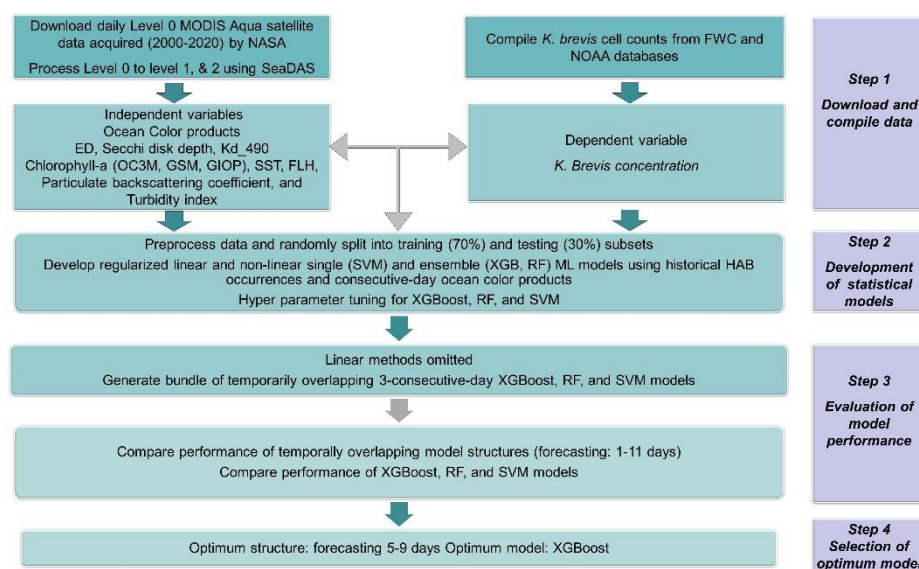
The study area incorporates the coastal areas (width: ~10–60 km, length: 180 km) of Charlotte County in southwest Florida and nearby estuaries, where freshwater and seawater mix (Figure 1). Like many other coastal zones within the Gulf of Mexico, there have been persistent HAB outbreaks that pose serious environmental challenges to the tourism and fishery industries in the county. Charlotte County has a relatively high density of septic systems in areas where the water table is often less than 2 feet below land surface. Shallow ground water and defective septic systems cause seepage of septic effluent into the water table. The introduction of nitrogen from septic systems into lakes, estuaries, and coastal areas is of concern given that nitrogen is one of the primary nutrients responsible for algal blooms occurrences. The majority of the samples are close to the shoreline and the sampling density decreases as we move away from the shoreline towards the ocean. Unfortunately, the study area lacks comprehensive, continuous, organized field-based monitoring systems.

## 3. Data and Methods

In this study, we apply an approach that addresses the paucity of continuous satellite temporal data and lag time variations among the controlling factors, provides predictions for bloom occurrences up to 9 days in advance, and provides insights into the factors controlling the onset of HABs, while predicting optimum solutions. Our approach takes advantage of remote sensing datasets, GIS technologies, and machine learning data-driven modeling. These data-driven models recognize hidden patterns in the collected data (dependent and independent variables) and provide insights into the behavior patterns of the observed ecosystems, namely the factors controlling the onset of HABs. In our



case, the factors controlling, or correlating with, HAB bloom growth and propagation are represented by the independent variables and the HAB occurrences are the dependent or the response variable. The workflow (Figure 2) involved four major steps: (1) downloading and processing of daily MODIS data; (2) developing statistical linear and non-linear models based on historical HAB occurrences and ocean color products derived from consecutive day MODIS data; (3) comparison of the performance of the models, and (4) selection of the optimum model and structure.



**Figure 2.** Flowchart describing the adopted methodology.

### 3.1. Data

Two types of data were used to construct our data-driven models for the period from late 2000 to March 2020. First were the independent variables—daily ocean color satellite products acquired by the National Aeronautics and Space Administration (NASA) MODIS Aqua satellite. Automatic selection of cloud-free (<10%) MODIS data (2905 scenes) was performed and used for this study. Only a small fraction (5%) of the omitted cloudy scenes was found to be cloud-free over the study area based on visual inspection of a subset of those scenes. Second was the dependent variable, daily *K. brevis* abundance (cells/L) observations from the National Oceanic and Atmospheric Administration (NOAA) and the Florida Fish and Wildlife Conservation Commission (FWC).

The independent variables were extracted from NASA's ocean color website (<https://oceancolor.gsfc.nasa.gov/>) for the daily acquired user-defined region of interest (ROI). Southwestern Florida was selected as ROI, and MODIS in Aqua mode as the source of data. The automatic data downloading was scheduled within the Linux environment. Following the download of Level 0 data, it was processed to Level 1, then to Level 2 using SeaDAS (NASA, Greenbelt, MD, USA, version 7.4) Ocean Color Science Software. Radiometric and geometric calibrations were performed to correct for differences in scene acquisition geometries (level 1 processing), and ocean color products were generated (level 2 processing).

The dependent variable (historical occurrences of *K. brevis* and their cell count) was compiled from two resources, namely from FWC and NOAA. The FWC and NOAA datasets contain daily observation of *K. brevis*, and both cover the period from 2000 to 2020.

#### 3.1.1. Independent Variables

Daily ocean color satellite products were automatically downloaded and processed. These include euphotic depth (ED), Secchi disk depth, chlorophyll-a, chlorophyll-gsm, chlorophyll-giop, diffuse attenuation coefficient (Kd\_490), SST, FLH, particulate backscat-

tering coefficient at 547 nm (bbp\_547\_giop), and turbidity index. Because previous work has shown that one or more of these variables could affect, or correlate with, the onset of HABs, each of these potential controlling factors was included in the statistical analysis; the individual variables are described below with their potential contribution to HABs' growth and propagation.

- Euphotic depth (m) and Secchi disk depth

The ED, represents the depth at which about 1 percent of the total incoming light on the ocean's surface can reach [73]. Beyond ED, light cannot penetrate, net photosynthesis and productivity decreases, and nutrients and algae diminish [74]. ED varies with change in season and latitude from only a few centimeters in highly turbid eutrophic waters to around 300 m in the open ocean. Low EDs can represent high nutrient content and provide desirable conditions for HAB growth and propagation [75,76]. The ED was calculated using the approach described in [77]. The Secchi disk depth has a similar concept; it is the depth at which a disk with alternating black and white quadrants disappears as it is lowered in the water column, and thus, it is a measure of the water transparency.

- Chlorophyll-a ( $\text{mg}/\text{m}^3$ )

Three common pigments (chlorophyll-a, -b, and -c) can be found in HABs, but the former (chlorophyll-a) was found to be the best proxy for measuring algal growth in aquatic environments [78,79]. Three different semi-analytical algorithms were developed to compute the chlorophyll-a concentration: chlorophyll-a OC3M (ocean chlorophyll three-band algorithm for MODIS [80]), chlorophyll-a GSM (Garver-Siegel-Maritorena [81]), and chlorophyll-a GIOP (Generalized Inherent Optical Property [82]). These three chlorophyll-a measurements products are highly correlated with HAB cell count, yet they are not redundant; often, one of these algorithms can best estimate the chlorophyll-a concentration in a particular optically complex estuarine environment [83]. In general, the increase in chlorophyll-a concentration has been found to have a strong correlation with the HAB distribution [66].

- Diffuse attenuation coefficient ( $\text{Kd}_{490}$ ;  $\text{m}^{-1}$ )

The diffuse attenuation coefficient of downwelling irradiance at 490 nm reflects the attenuation of the light in blue to green wavelength regions for turbid water and is one of the most important optical properties of ocean water [84]. In one study the  $\text{Kd}_{490}$  coefficient was used as a proxy for the growth of phytoplankton in turbid coastal waters, where the light attenuation was shown to be controlled by the concentration of scattering particles, HABs being one of them [85]. In another study under normal and red tide outbreak conditions in the Persian Gulf, the MODIS Chlorophyll-a normalized line fluorescence height, and  $\text{Kd}_{490}$  were compared; a high correlation was observed between chlorophyll-a and  $\text{Kd}_{490}$  during red tides [86]. The  $\text{Kd}_{490}$  was calculated using the technique described in [87].

- Sea surface temperature ( $^{\circ}\text{C}$ )

Phytoplankton and HAB growth and productivity is directly correlated with SST. The HABs can thrive under specific habitat characteristics and temperature range. The temperature controls the survival of the HABs and the availability and solubility of nutrients that are vital for the growth of HABs as well [88,89]. The correlation between SST and algal bloom growth and its distributions has been successfully demonstrated in various settings worldwide [89–92].

- Fluorescence line height

FLH provides a standard method for measuring radiance, leaving the coastal and ocean surface in the chlorophyll fluorescence emission band (676 nm) [44]. A strong positive correlation was reported between chlorophyll-a concentration and the FLH in ocean waters containing HABs [40]. FLH alone and together with backscattering coefficient have been

successfully used in the detection of chlorophyll-a and *K. brevis* distribution in the Charlotte Harbor Estuary in Florida and in the Gulf of Mexico [30,93–96].

- Particulate backscattering coefficient

This factor represents the backscattering coefficient of water particles at 547 nm. Earlier studies have shown its utility in identifying HABs distribution, particularly the *K. brevis* in the Gulf of Mexico [96]. In two different studies at the West Florida shelf, the particulate backscattering coefficient at 551 nm, in conjunction with fluorescence (in the first study) and chlorophyll-a (in the second) was utilized to detect *K. brevis* [94,97]. The backscatter coefficient of particles at 547 nm was calculated using an algorithm provided in [98].

- Turbidity index

The turbidity index is based on the reflectance in the green part of the spectrum. It provides a measure of the water clarity based on the amount of the scattered light caused by water-suspended particles [99]. When it is low, water is clearer, and more light can penetrate down into the water column, providing favorable living and growing conditions for HABs [100]. On the other hand, HAB growth increase turbidity, per se. Turbidity alone is not a direct indicator of HAB concentration, but it can be used in conjunction with other aforementioned factors. It has been successfully used to estimate the severity of HABs and to identify phytoplankton blooms [4,101]. The turbidity index was calculated using the method described in a previous study [102].

### 3.1.2. Target Variable

The number of *K. brevis* (cells/L) in shallow (depth: 0.5 m) waters is considered to be the target dependent variable (response variable). A threshold of 10,000 cells/L was adopted for classification purposes, because at concentrations exceeding 10,000 cells/L, respiratory irritation and fish kills are more likely to occur (<https://myfwc.com/research/redtide/statewide/>) and the chlorophyll-a concentration is high enough to enable the detection of HABs from satellite data [103]. Moreover, the adopted cell count groupings in this study are those used by the HABs Observing System (+ve: >10,000 cells/L; −ve: <10,000 cells/L) [12].

### 3.1.3. Data Preparation

In the proposed models, the number of *K. brevis* cells (cells/L) is used as the response variable. For the classification application a threshold of 10,000 cells/L was adopted to separate cell counts into two classes of positive and negative events. Ten level-02 ocean color products are used as controlling factors. Chromophoric dissolved organic matter index was manually removed from the list of level 02 products due to the discontinuous and patchy nature of this variable over the investigated period. In the generation of the models, the dataset was randomly split into train (70%) and test (30%). All the models were tested on roughly 300 positive and negative events that have not been seen by the models covering the observation time period from 2000 to 2020.

## 3.2. Machine Learning Modeling

We developed data-driven machine learning models to address the problem. The adopted state-of-the-art machine learning models are discussed in two categories: linear versus non-linear, and tree-based versus non-tree-based models. Shrinkage methods were adopted as an example of regularized linear models, SVM for non-tree-based models, and XGBoost and RF as examples for non-linear tree-based models. Due to data size, data distribution, and the complexity of patterns in data, we follow a common practice in which we utilize, compare, and contrast a set of statistical models described in the following sections.

### 3.2.1. Linear Models

We chose linear models because they have advantage in interpretability. By stacking up the same variables for different days (multicollinearity alert), we significantly increase the feature space (e.g., three consecutive days and 10 predictors for each day). We used the shrinkage method that applies a penalty term to the loss function embedded in the linear regression (LR) to avoid over- and underfitting. Shrinkage models shrink insignificant variables (coefficient estimates) into zero, which leaves us with the most significant variables to address the lag times [72].

### 3.2.2. Tree-Based Models (Non-Linear)

Since we are stacking up a few sets of the same variables in consecutive days, the correlation among the same variables is very high. This imposes unsolvable multicollinearity to linear models. Therefore, we resorted to nonlinear models, such as tree-based models, to address the nonlinearity content of the problem. These models provide variable importance plots and can handle limited training datasets, which is the case in our investigation. Trees can be non-robust with high variance, which is why we considered ensemble models such as extreme gradient boosting (XGB) and RF to improve the prediction accuracy and lower the variance [72,104].

- Extreme gradient boosting

XGB is a scalable learning algorithm designed for higher speed and performance. It uses a regularized model formalization to control overfitting.

In gradient boosting, the input predictors  $X$  ( $X_1 \dots X_n$ ) are utilized to predict the corresponding target values  $Y$  ( $Y_1 \dots Y_n$ ). In fact, we need to minimize the sum of the loss function ( $J$ ) by improving the model  $F(X)$ . Equations (1)–(5) are from [105,106]

$$J = \sum_{i=1}^n L(y_i, F(x_i)) \quad (1)$$

where  $L$  is a differentiable convex loss function to measure the difference between the predicted values  $F(x_i)$  and the real target values ( $y_i$ ).

In applying the XGB, we went through the following iterations. We first calculated the negative gradients of  $J$  with respect to  $(x_i)$ , which is  $-\frac{\partial J}{\partial F(x_i)}$ .

Then, we fit a classification tree,  $h$ , to  $\frac{\partial J}{\partial F(x_i)}$ . The new updated  $(Xi)$  is  $(Xi) + \gamma h$ , where  $\gamma$  is the step size to reach the estimated minimum of  $J$ .

The iteration continues to the point at which we achieve the minimum difference between prediction and observation. In XGB, the loss function is:

$$J = \sum_{i=1}^n L(y_i, F(x_i)) = 1 + \Omega(h) \quad (2)$$

where:

$$\Omega(h) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (3)$$

where  $T$  is the number of tree leaves and  $\omega$  is the weights of those leaves. The function  $\Omega$  penalizes the model complexity. The optimal weight  $\omega$  of leaf  $j$  was calculated using Equation (4):

$$\omega_j = \frac{\sum_j g_j}{\sum_j h_j + \lambda}, \quad (4)$$

where  $g_j = -\frac{\partial J}{\partial F(x_i)}$  and  $h$  is the  $j^{\text{th}}$  classification tree fitted to  $g_j$ . The optimal value of the loss function was calculated using Equation (5) [106]:

$$J = -\frac{1}{2} \sum_{j=1}^T \frac{(\sum_j g_j)^2}{\sum_j h_j + \lambda} + \gamma T \quad (5)$$



The additional regularization term was added to avoid overfitting [105,106].

To achieve the optimum structure for the XGB model, parameters such as number of boosting iterations, gamma, maximum depth, and learning rate (eta) were tuned. Gamma is a pseudo-regularization hyperparameter in gradient boosting (complexity control). The higher the gamma is, the higher the regularization, and the more conservative the algorithm will become. The eta specifies the participation of each tree and reduces overfitting. Maximum depth determines the maximum number of end nodes in each leaf of the trees. These hyperparameters were calculated based on grid search and cross validation in R. We found that the optimum hyper parameters for the XGB including the number of boosting iterations, gamma, maximum depth, and learning rate (eta) were 100, 0, 10, and 0.05, respectively.

- Random forest

In RF, we build hundreds of trees on bootstrapped training samples. However, each time we generate an individual tree and a split in a tree is considered, a random fresh selection of  $M$  predictors is chosen as split candidates from the full set of the  $P$  predictors to avoid the strongest predictor always being utilized in the process [107]. Typically,  $M$  is calculated using Equation (6):

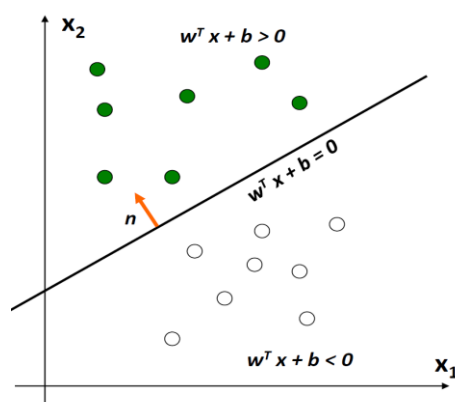
$$M = \sqrt{P} \quad (6)$$

At each split a new sample of predictors is considered according to a user-defined number of predictors ( $M_{try}$ ). Another user-specified hyperparameter is the number of trees ( $N_{tree}$ ). Small values were avoided to enable making of the forest and to enhance variance-bias tradeoff [108,109]. For global optimum, two-third of the samples were used for training, and the remaining out-of-bag (OOB) were used to cross-validate the RF model. In fact, the OOB error was utilized to calculate the prediction error and to evaluate the variable importance measures [110]. The RF hyperparameters ( $M_{try}$  and  $N_{tree}$ ) that were used in this study to optimize the model performance were 6 and 1000, respectively.

- Support vector machines

To evaluate the reliability of results, a non-parametric (insensitive to the distribution of data) non-tree-based supervised learning method called SVM was adopted [111]. SVM proposes a nonparametric approach of finding linear discriminant functions. It tries to find a unique hyperplane between each pair of the classes in a multidimensional feature space [112].

The linear function general formula is  $g(x) = W^T \cdot x + b$ , which is a hyperplane in higher dimensions and is represented in Figure 3.



**Figure 3.** Two-dimensional feature space with SVM linear discrimination function.

In Figure 3,  $x_1$  and  $x_2$  are two features, the green and white classes are separated by the line  $W^T x + b = 0$ , and  $n$  is the normal vector of the hyperplane ( $n = \frac{w}{\|w\|}$ ), where  $\|w\|$

is the Euclidian distance between  $w^{\rightarrow}$  and the origin. The main objective of SVM is to find a set of weights that specify two hyperplanes Equation (7)

Given a set of data points  $\{(x_i, y_i)\}, i = 1, 2 \dots n$ ; where:

$$\begin{cases} y_i = +1, w^t x_i + b \geq k \\ y_i = -1, w^t x_i + b \leq k \end{cases} \quad (7)$$

$k = 1$  after scale transformation on both  $w$  and  $b$ .

We have infinite possible discrimination functions. One way to find the optimal hyperplane is by maximizing the width of the margin (margin width:  $\frac{2}{\|w\|}$ ) or minimizing  $\frac{1}{2} \|w\|^2 = \frac{1}{2} w^t w$  such that:  $y_i (w^t x_i + b) \geq 1$ :

$$\begin{cases} \min \left( \frac{1}{2} w^t w \right) \\ y_i (w^t x_i + b) \geq 1 \end{cases} \quad (8)$$

By solving this optimization equation for  $w$  and  $b$ , each of the  $y_i$  data points are correctly classified. In fact,  $y_i$  indicates the class value (transformed either to +1 or −1).

In the adopted SVM, a radial basis function kernel yielded a better performance and was applied to address nonlinearity and overfitting. Model optimization was performed using the tune function in the R software package on the SVM hyperparameters (gamma  $[\gamma]$  and cost  $[c]$ ). The cost hyperparameter specifies the cost of a violation to the margin; at small cost values, margins will be wide and many support vectors will be available, and vice versa at high cost values. The model overfits data as the values of  $c$  or  $\gamma$  increase and underfits as their values decrease. The average number of support vectors, optimum  $\gamma$ , and optimum  $c$  were selected at 95, 0.5, and 0.1, respectively.

### 3.2.3. Assessment of Models

The performance of the models was evaluated using the test data with binary classes of low and high concentration of *K. brevis* and applying a confusion matrix.

Some of the important metrics of confusion matrix appropriate for imbalanced datasets are Cohen's kappa, balanced accuracy, and F-score. On top of that, receiver operating characteristic (ROC) curve and the area under the curve (AUC) were calculated to compare different classifiers. ROC is a graphical plot that represents false and true positive rates on the x and y axes, respectively. ROC indicates a model's diagnostic ability when the class discrimination threshold varies [107]. These criteria are calculated as follows [107]:

$$Specificity = \frac{TN}{TN + FP} \quad (9)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (10)$$

$$Balanced\_accuracy = \frac{Specificity + Sensitivity}{2} \quad (11)$$

$$F\text{-Measure} = \frac{TP}{TP + \frac{1}{2} (FP + FN)} \quad (12)$$

$$Kappa = \frac{Po - Pe}{1 - Pe} \quad (13)$$

where:

$$Po = \frac{TP + TN}{n} \text{ and } Pe = \frac{1}{\sqrt{N}} ((TP + FN)(TP + FP) + (FP + TN)(FN + TN)). \quad (14)$$

In the above equations,  $N$  is the total number of cases,  $n$  points to the number of accurately categorized incidents or non-incidents,  $TP$ ,  $TN$ ,  $FP$ , and  $FN$  refer to true positive, true negative, false positive, and false negative, respectively.

The performance of 1-day (e.g.,  $-7$ ;  $-8$ ;  $-9$ ), 2 consecutive day (e.g.,  $-7$ ,  $-8$ ;  $-8$ ,  $-9$ ), and 3 consecutive day (e.g.,  $-7$ ,  $-8$ ,  $-9$ ) models was measured using four performance metrics (kappa, F-score, precision, and balanced accuracy). In this case and throughout the text, the “-ve” sign and the numbers refer to the number of days ahead of a bloom onset. For example, the 3 consecutive day ( $-7$ ,  $-8$ ,  $-9$ ) model refers to a model that uses satellite data acquired on three consecutive days, 7, 8, and 9 days ahead of a bloom occurrence. For simplification purposes, a 3 consecutive day ( $-7$ ,  $-8$ ,  $-9$ ) model will be referred to hereafter as a 7-day model, a 3 consecutive day ( $-9$ ,  $-10$ ,  $-11$ ) model as a 9-day model, and a 3 consecutive day ( $-10$ ,  $-11$ ,  $-12$ ) model as a 10-day model.

There are two measures of variable importance in the RF models. The first is based on how much the accuracy decreases when we exclude the variable. The second measure is based on the decrease in Gini impurity when a variable is selected to split a node. For boosting-based models (XGBoost), learning is done in a serial way; when there are several correlated features (as in our case), boosting will tend to choose one and use it in several trees (if necessary), and the use of other correlated features will be limited. On the other hand, not each tree of an RF is built from the same features (there is a random selection of features to use for each tree). Therefore, RF has the most intuitive feature importance for our case. In addition, each time we run the forest-based classification, we get slightly different results due to both randomness introduced in the model to avoid overfitting and the random subsetting of validation data. Therefore, instead of a bar chart, we get a variable importance box-plot that shows the distribution of importance across many runs.

#### 4. Results

##### 4.1. Model Structure Comparison and Selection of Optimum Model Structure

Using the XGB model, we compared the performance of 1-day ( $-7$ ;  $-8$ ;  $-9$ ), 2 consecutive day ( $-7$ ,  $-8$ ;  $-8$ ,  $-9$ ), and 3 consecutive day ( $-7$ ,  $-8$ ,  $-9$ ) models (Table 1). Examination of Table 1 shows that the 3 consecutive day structure outperforms the 2 consecutive day models, which in turn outperform the 1-day models. A similar exercise was conducted using the SVM model, and again the 3 consecutive day structure was found to outperform the 2 consecutive day models, which in turn outperform the single day. Although not shown, we observe a general enhancement in the performance of each of the remaining three models (XGBoost, RF, and SVM) with increasing number of consecutive days. Thus, there are added benefits for increasing the number of consecutive day entries to our models given the same number of variables.

**Table 1.** Comparison between the performance of single and 2 and 3 consecutive day models.

Combination (7th XGB)	$-7$	$-8$	$-9$	$-7, -8$	$-8, -9$	$-7, -8, -9$
<b>Kappa</b>	0.77	0.76	0.76	0.74	0.76	<b>0.80</b>
<b>F-score</b>	0.89	0.88	0.88	0.85	0.88	<b>0.96</b>
<b>Precision</b>	0.84	0.88	0.88	0.92	0.88	<b>0.94</b>
<b>B. accuracy</b>	87.0	86.0	86.0	83.0	86.0	<b>88.0</b>
Combination (2nd SVM)	$-2$	$-3$	$-4$	$-2, -3$	$-3, -4$	$-2, -3, -4$
<b>Kappa</b>	0.35	0.4	0.37	0.4	0.4	0.50
<b>F-score</b>	0.51	0.52	0.48	0.52	0.52	0.60
<b>Precision</b>	0.56	0.72	0.81	0.73	0.73	0.82
<b>B. accuracy</b>	66.0	66.0	64.0	67.0	66.0	71.0

Unfortunately, the availability of cloud-free (<10%) MODIS data over the study area limits our ability to develop models that utilize more than three consecutive days. Table 2

shows that out of a total of 2905 scenes that were acquired over the study area and period, the single scenes constituted 36% (1039) of the cloud-free scenes, the 2 consecutive day scenes constituted 20% (562 scenes) of the cloud-free scenes, the 3 consecutive day scenes constituted 9% (260 scenes) of the cloud-free scenes, the 4 consecutive day scenes constituted <2% (57 scenes) of the cloud-free scenes, and each of the 5, 6, and 7 consecutive scenes constituted less than 1% of the cloud-free scenes.

**Table 2.** Availability of cloud-free (<10%) MODIS data (2905 scenes) acquired over the study area and period (2000–2020).

Days (2000–2020) < 10% Cloud	Scenes	Frequency
Total	2905	
single days	1039	36%
2 consecutive days	562	20%
3 consecutive days	260	9.0%
4 consecutive days	57	1.9%
5 consecutive days	29	0.9%
6 consecutive days	21	0.7%
7 consecutive days	14	0.4%
8 consecutive days	6	0.2%
9 consecutive days	1	0.03%
10 consecutive days	0	0%

Given the paucity of consecutive MODIS data for periods exceeding three days and the lesser chances for finding field observations (dependent variable) in 4 consecutive day scenes for model training purposes, we chose to develop our models based primarily on the 3 consecutive day structure.

Eleven 3 consecutive day models were generated (Table 3, top). This structure and bundle of temporally overlapped models provides short- to mid-term HAB forecasting through a range of spatio-temporal models, and addresses, at least in part, the differences in optimum lag times between each of the individual independent variables and the onset of HABs. For example, the first model uses ocean color products acquired a day prior, two days prior, and three days prior to onset of an HAB occurrence; the last model uses data acquired 11, 12, and 13 days in advance. As described earlier, the 3 consecutive day models produce better results than the 2-day models. For comparison purposes, the structure of 11 2 consecutive day models is shown in Table 3 (bottom).

#### 4.2. Comparison of the Performance of Statistical Models and Selection of the Optimum Model

A variety of machine learning algorithms were adopted based on the nature of data and the problem in hand. The Lasso regression analysis was first adopted, but all the variables were found to shrink to zero due to very high multicollinearity among the variable sets. Tree-based models (XGBoost, RF, and SVM) were then applied. The ROC curve (AUC), balanced accuracy, kappa, and F-score derived from confusion matrix were adopted to evaluate the performance of models on the test dataset. The comparison of forecasting models is displayed in Table 4 and Figure 4. The model structures are represented by numbers ranging from −1 to −13, representing the days in advance of an HAB occurrence. The best metrics among the three models (XGB, RF, and SVM) are boldfaced. For example, XGBoost achieved the highest performance among the models for eight (−8) days forecasting with all four metrics (accuracy: 96%; Kappa: 0.93; F-score: 0.97; and AUC: 0.98). Figure 4 displays the ROC plots for the train and test datasets for 8-day SVM, RF, and XGBoost models. The three model ROC curves and the area under them (AUC) indicates a slightly higher performance for the XGBoost compared to the other models (AUC: XGBoost, 0.98; RF, 0.96; SVM, 0.94). XGBoost was selected as the optimum model.

**Table 3.** Temporal modeling structure for 2 and 3 consecutive day models.

3 Day Models													Day
−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	
										X	X	X	Bloom
									X	X	X		Bloom
								X	X	X			Bloom
							X	X	X				Bloom
						X	X	X					Bloom
				X	X	X							Bloom
			X	X	X								Bloom
		X	X	X									Bloom
	X	X	X										Bloom
X	X	X											Bloom

2 Day Models													Day
−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	
											X	X	Bloom
										X	X		Bloom
									X	X			Bloom
								X	X				Bloom
							X	X					Bloom
						X	X						Bloom
				X	X								Bloom
			X	X									Bloom
		X	X										Bloom
	X	X											Bloom

**Table 4.** Comparison between the performance of 3 consecutive day models using XGBoost, RF, and SVM. The best metrics are face bolded.

XGBoost													Model Performance			
−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	Accuracy	Kappa	F-Score	AUC
										X	X	X	73.1	0.52	0.64	<b>0.74</b>
									X	X	X		73.9	<b>0.65</b>	<b>0.82</b>	0.85
								X	X	X			<b>58.0</b>	<b>0.27</b>	<b>0.63</b>	0.67
							X	X	X				76.4	0.58	<b>0.78</b>	0.84
						X	X	X					<b>83.9</b>	<b>0.07</b>	<b>0.87</b>	<b>0.88</b>
				X	X	X							<b>92.0</b>	<b>0.86</b>	<b>0.95</b>	<b>0.97</b>
			X	X	X								87.6	<b>0.81</b>	<b>0.96</b>	<b>0.98</b>
			X	X									<b>96.2</b>	<b>0.93</b>	<b>0.98</b>	<b>0.98</b>
		X	X	X									<b>87.4</b>	<b>0.76</b>	<b>0.92</b>	<b>0.91</b>
	X	X	X										<b>83.6</b>	<b>0.71</b>	<b>0.88</b>	0.81
X	X	X											79.6	<b>0.68</b>	<b>0.80</b>	<b>0.80</b>

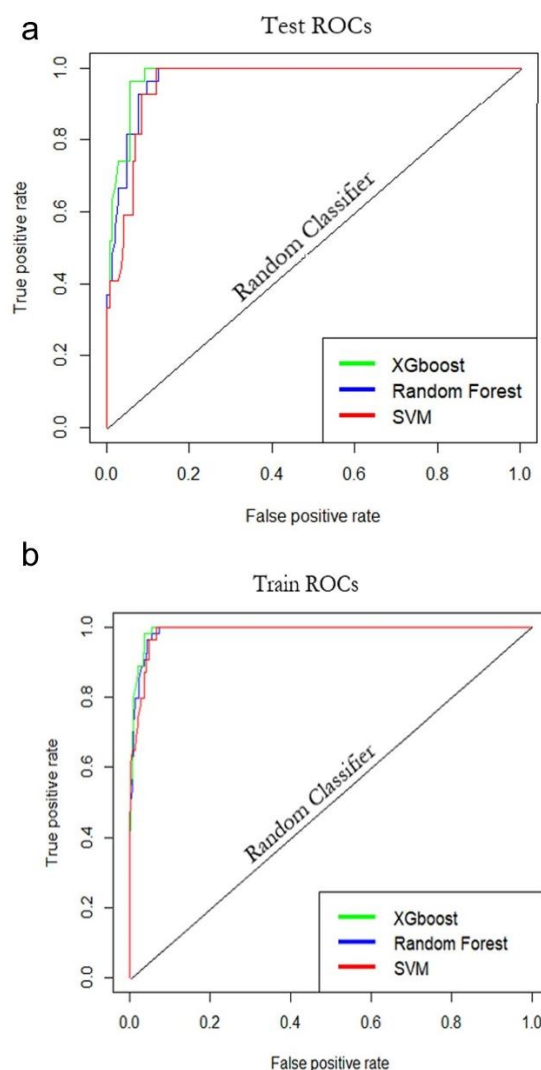
  

RF													Model Performance			
−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	Accuracy	Kappa	F-Score	AUC
										X	X	X	65.6	0.40	<b>0.74</b>	0.73
									X	X	X		<b>77.2</b>	0.63	0.71	<b>0.86</b>
								X	X	X			54.7	0.13	0.20	0.74
							X	X	X				76.1	0.60	0.73	<b>0.87</b>
						X	X	X					83.5	0.67	0.80	0.83
					X	X	X						89.2	0.82	0.84	0.95
				X	X	X							<b>91.4</b>	0.75	0.84	0.96
			X	X	X								95.2	0.92	0.95	0.96
		X	X	X									78.3	0.73	0.88	0.80
	X	X	X										81.7	0.67	0.78	<b>0.87</b>
X	X	X											<b>80.7</b>	0.62	0.71	0.79



Table 4. Cont.

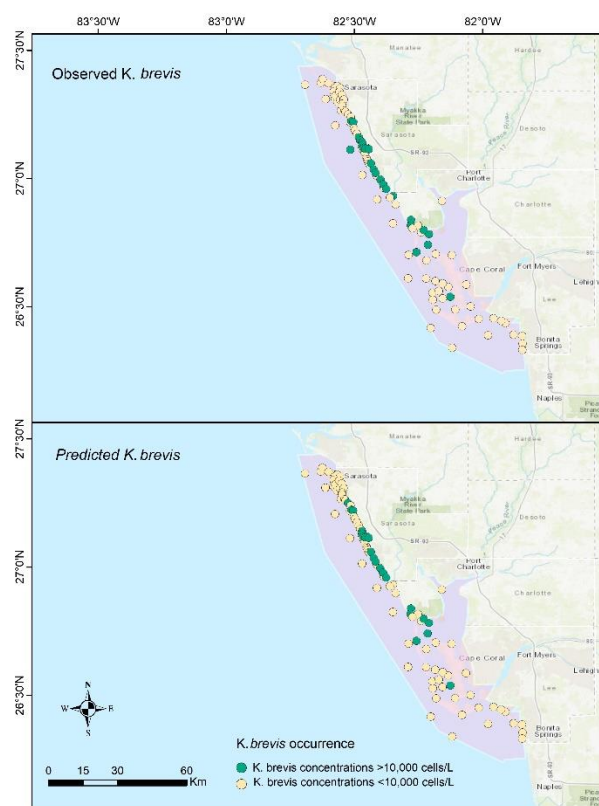
SVM													Model Performance			
−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	Accuracy	Kappa	F-Score	AUC
										X	X	X	62.4	0.35	0.72	0.69
									X	X	X		71.2	0.50	0.60	0.80
							X	X	X	X			56.3	0.20	0.27	<b>0.79</b>
							X	X	X				73.7	<b>0.63</b>	0.75	0.84
						X	X	X					83.6	0.67	0.81	0.81
					X	X	X						87.0	0.66	0.77	0.94
				X	X	X							91.1	0.72	0.79	0.90
			X	X	X								88.2	0.83	0.86	0.93
		X	X	X									74.1	0.62	0.63	0.82
	X	X	X										63.0	0.32	0.74	0.80
X	X	X											61.0	0.59	0.70	0.80

**Figure 4.** Machine Learning models ROCs for Test ROC (a) and Train ROC (b).

#### 4.3. Comparison of Lag Times and Selection of Optimum Lag Time

Inspection of Table 4 reveals that, in general, the forecasting models of ~5–9 days in advance (5- to 9-day models) achieved relatively more reliable and comparable results, with the 7- and 8-day forecasting models being the optimum models. For example, Figure 5 shows a good correspondence between the reported concentration of *K. brevis* in 170 random test samples within the study area with the predicted concentration for each of those

samples from an 8-day RF model. The samples have been classified correctly with an accuracy of 95%.

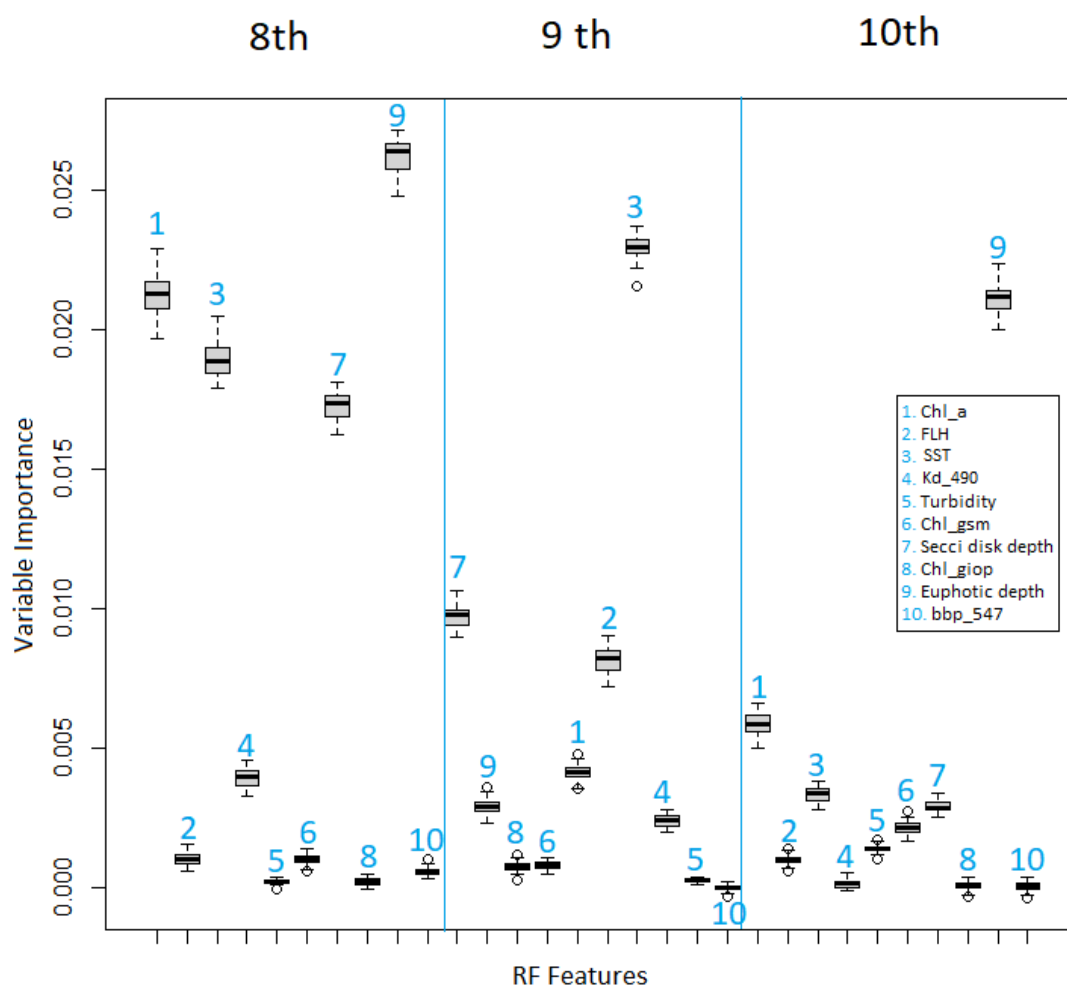


**Figure 5.** Comparison between the reported *K. brevis* concentration in 170 randomly selected test samples within the study area (**top**) with their predicted concentrations from an 8-day RF model (**bottom**). The samples have been classified correctly with 95% accuracy.

XGBoost (the top subtable) also demonstrated a more uniform superiority performance in this interval, which is portrayed by boldfaced figures. The models out of this range (5–9 days) in general showed a relatively lower performance, especially on the kappa metric, which is an indicator of model performance in comparison to a random guess (random classifier). For example, in general, the 1-, 2-, 3-, 4-, 10-, and 11-day models in all three ML methods showed low kappa values compared to the 5-, 6-, 7-, 8-, and 9-day models.

#### 4.4. Identification of Controlling Factor Importance

An RF feature importance plot was selected for the depiction of the significant variables because it provides the most intuitive display of variable importance (refer to Section 3.2.3). Figure 6 shows the variable importance boxplot for the 8-day RF model. The x-axis displays 30 controlling factors, 10 factors for each of the 3 days (days 8, 9, and 10) in three sets; the y-axis is the scaled variable importance. Boxplots show the range of variations in variable importance in different model runs ( $n = 100$ ). Each set of 10 variables is separated by a vertical blue line. Chlorophyll-a, SST, Secchi disc depth, and ED from the 8th day are amongst the most significant variables. Although not shown, XGBoost showed generally similar, yet not identical results in feature importance.



**Figure 6.** Variable importance (VI) boxplot for the best RF model. The VI metrics are on the y-axis and the 10 variables for each of the 3 consecutive days are numbered.

## 5. Discussion

This study was intended to provide guidelines for the development of comprehensive predictive HABs models using temporal remotely acquired data in ways that can address, at least in part, two of the main shortcomings of remote sensing-based HAB predictive models: (1) the paucity of satellite data due to cloudy scenes and other systematic and random missing data that prevent us from making reliable models and (2) the differences in lag time between the period at which the individual variables reach their highest correlation with the target and the time the bloom occurs.

Our findings suggest that these shortcomings could be addressed by using multiple sequential, consecutive day models, as opposed to 1-day models [55]. The larger the number of consecutive days, the better the results. In our case, the paucity of consecutive cloud-free data limited our analysis to 2 and 3 consecutive day models (11 models). The 3 consecutive day models predict the onset of HABs from 1 to 11 days in advance and accommodate differences in lag time of up to 2 days for the independent variables. Three consecutive day models increased the variability within the models and increased the overall performance of all models (SVM, XGBoost, and RF) in comparison with the 1-day or 2-day models. In absence of 3 consecutive day data, one can resort to the use of 2 consecutive day models. Comparisons of our 3-day XGB model outputs with previous 1-day model outputs [54] over the study area reveal enhanced overall performance for our 3 consecutive day models. The 1-day model achieved 65% accuracy for the one day in advance model, and 72.1% for the two-day model [54]. The overall performance of our 1-day and 2-day forecasting XGBoost models adjusted for imbalance yielded an accuracy of

73.0% and 73.9, respectively. In both studies, the SST, chlorophyll-a, KD<sub>490</sub>, and euphotic depth, were among the most important controlling factors.

Following the general trend in almost all four metrics, the performance of the models was enhanced with the ~5–9-day models, and the best results were those obtained from the 7–8-day models. For example, in the XGBoost analysis (Table 4), the kappa ranged from 0.5 to 0.6 for day 1, 2, 3, and 4 models, rose to up to 0.93 in the day-8 model, then decreased to 0.68 in the 11-day model. The mid-term forecasting of 5–9 days is not only more accurate, but is also more functional compared to short-term forecasting models. They provide enough time to execute appropriate warning and mitigation steps prior to the onset of HAB occurrences. One explanation for the above-mentioned findings is that better (−5 to −9) and optimum lag times (−7, −8) are met within these time frames. The VI boxplot shows that the ED and Secchi disc depth were found to be highly significant factors, probably due to the spatial variability of these factors throughout the study area and their correlation with the distribution and concentration of HABs. Chlorophyll-a and SST were found to be significant factors as well. Our findings are consistent with previous non-data-driven studies [57,58,113]. In the Arabian Sea, phytoplankton biomass was found to correlate with SST with a lag time of 2–8 days [58]; in Santa Monica, it was 5 days [114]. In addition, a 3–6 day lag time was reported between the introduction of phosphorus nutrients and the occurrence of HAB events [115].

The XGBoost model outperformed the SVM and RF models. It combines the advantages of RF and gradient boosting. The high performance of the XGBoost model can be attributed to the specific data patterns and size, data imbalance, and more robustness of the model to noisy data and outliers due to its loss function flexibility, which has a regularization term to reduce the complexity of the tree functions.

Our method has its limitations. The application of our methodology in an area will ultimately depend on the availability of consecutive cloud-free data. In arid parts of the world, that should not be a major problem, but in temperate areas, cloud-free data for the application of the optimum 3-day models (5- to 9-day models) might not be available, and in such cases, less accurate 3-day models (1–4, 10, and 11-day models) or even 2-day models will have to be used instead. The generated models are specific for the investigated area, and thus, similar models have to be tailored to individual areas. Moreover, the proposed approach is labor-intensive, since it requires the development of many models—in our case, some 20 models, including extensive data engineering, data blending, and data wrangling. Using archival field and satellite data (as a training dataset) for areas of interest, future work should concentrate on the development of automated systems that can construct tens to hundreds of models at various combinations of consecutive days, (2-day models, 3-day models, 4-day models, 5-day models, etc.), evaluate their performance using test data, rank the models based on their performance, and, depending on availability of cloud-free ocean color data, select and apply the model with the highest performance.

## 6. Conclusions

We developed, compared, and contrasted the efficiency of state-of-the-art data-driven machine learning models (XGBoost, RF, and SVM) in predicting the occurrence of HABs. The number of *K. brevis* cells in surface water samples collected during red tides over the past 20 years were used as a binary response to the environmental controlling factors (target variable) and 10 level-02 ocean color products extracted from daily archival MODIS satellite data were used as environmental controlling factors.

Two main shortcomings of earlier models were addressed: (1) the paucity of satellite data due to cloudy scenes and other systematic and random missing data and (2) the lag time between the period at which a variable reaches its highest correlation with the target and the time the bloom occurs. Eleven spatio-temporal models were generated, each from three consecutive days' satellite datasets, with a forecasting span of 1 to 11 days. One or more of the generated 11 models could be used to predict HAB occurrences with acceptable performance, depending on availability of the cloud-free consecutive days.

Findings indicate: (1) XGBoost outperformed the remaining methods, (2) the forecasting models of 5–9 days achieved the best and most reliable results, (3) the most reliable model can forecast eight days ahead of time, and (4) ED, SST, and chlorophyll-a are always among the most significant variables.

The findings from this study could serve as guidelines for the development of remote sensing-based early warning systems for HABs in southwest Florida with short- to mid-term forecasting capabilities. As described above, the generated models are specific to the study area and their development is labor intensive. Thus, speedy and widescale applications of the developed concepts in areas outside of the study area requires development of fully automated algorithms that will accomplish the following functions: downloading of daily data acquisition for the desired study area from a big data platform, designing and maintaining a database management system for data blending and query-based data engineering, online ML modeling, and interactive model evaluation.

In southwest Florida, it was difficult to get cloud-free ocean color acquisition for more than three consecutive days. There are many other coastal areas around the world, especially in arid areas in which cloud free scenes are more available, where the developed methodologies could be readily applied. The development of automated systems that can construct many models at various combinations of consecutive days could facilitate the application of the advocated methods over areas where cloud-free data is limited.

Additional approaches to address the paucity of cloud-free consecutive day data should be explored and their performance evaluated. For example, additional statistical models that rely on non-consecutive day data could be generated; alternatively, the values of the missing days could be estimated using forward window averages or data imputation prior to feeding the data into the ML algorithms.

**Author Contributions:** Conceptualization, M.I. and M.S.; methodology, M.S. and M.I.; software, M.I. and A.G.; validation, M.I.; formal analysis, M.I.; investigation, M.S. and R.E.K.; resources, M.S. and K.A.; data curation, M.S. and M.I.; writing, original draft preparation, M.I., M.S., and K.A.; writing, review and editing, M.I., M.S., R.E.K., and K.A.; visualization, M.S. and R.E.K.; supervision, M.S.; project administration, M.S.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This project was supported through Enterprise Charlotte Foundation and Western Michigan University.

**Data Availability Statement:** Data cited in this manuscript are available in registries that are freely accessible to the public.

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

## References

1. Fuentes-Yaco, C.; Vézina, A.F.; Larouche, P.; Gratton, Y.; Gosselin, M. Phytoplankton pigment in the Gulf of St. Lawrence, Canada, as determined by the Coastal Zone Color Scanner Part II: Multivariate analysis. *Cont. Shelf Res.* **1997**, *17*, 1441–1459. [\[CrossRef\]](#)
2. Chari, N.V.H.K.; Keerthi, S.; Sarma, N.S.; Pandi, S.R.; Chiranjeevulu, G.; Kiran, R.; Koduru, U. Fluorescence and absorption characteristics of dissolved organic matter excreted by phytoplankton species of western Bay of Bengal under axenic laboratory condition. *J. Exp. Mar. Biol. Ecol.* **2013**, *445*, 148–155. [\[CrossRef\]](#)
3. Gohin, F.; Lampert, L.; Guillaud, J.F.; Herbland, A.; Nézan, E. Satellite and in situ observations of a late winter phytoplankton bloom, in the northern Bay of Biscay. *Cont. Shelf Res.* **2003**, *23*, 1117–1141. [\[CrossRef\]](#)
4. Kahru, M.; Mitchell, B.G.; Diaz, A.; Miura, M. MODIS detects a devastating algal bloom in Paracas Bay, Peru. *Eos* **2004**, *85*, 465–472. [\[CrossRef\]](#)
5. Oliveira, P.B.; Moita, T.; Silva, A.; Monteiro, I.T.; Sofia Palma, A. Summer diatom and dinoflagellate blooms in Lisbon Bay from 2002 to 2005: Pre-conditions inferred from wind and satellite data. *Prog. Oceanogr.* **2009**, *83*, 270–277. [\[CrossRef\]](#)
6. Ryan, J.P.; Fischer, A.M.; Kudela, R.M.; Gower, J.F.R.; King, S.A.; Marin, R.; Chavez, F.P. Influences of upwelling and downwelling winds on red tide bloom dynamics in Monterey Bay, California. *Cont. Shelf Res.* **2009**, *29*, 785–795. [\[CrossRef\]](#)
7. Tilstone, G.H.; Angel-Benavides, I.M.; Pradhan, Y.; Shutler, J.D.; Groom, S.; Sathyendranath, S. An assessment of chlorophyll-a algorithms available for SeaWiFS in coastal and open areas of the Bay of Bengal and Arabian Sea. *Remote Sens. Environ.* **2011**, *115*, 2277–2291. [\[CrossRef\]](#)



8. Moradi, M.; Kabiri, K. Red tide detection in the Strait of Hormuz (east of the Persian Gulf) using MODIS fluorescence data. *Int. J. Remote Sens.* **2012**, *33*, 1015–1028. [\[CrossRef\]](#)
9. Blondeau-Patissier, D.; Gower, J.F.R.; Dekker, A.G.; Phinn, S.R.; Brando, V.E. A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. *Prog. Oceanogr.* **2014**, *123*, 123–144. [\[CrossRef\]](#)
10. Song, H.; Ji, R.; Stock, C.; Wang, Z. Phenology of phytoplankton blooms in the Nova Scotian Shelf-Gulf of Maine region: Remote sensing and modeling analysis. *J. Plankton Res.* **2010**, *32*, 1485–1499. [\[CrossRef\]](#)
11. Nezlin, N.P.; Li, B.L. Time-series analysis of remote-sensed chlorophyll and environmental factors in the Santa Monica-San Pedro Basin off Southern California. *J. Mar. Syst.* **2003**, *39*, 185–202. [\[CrossRef\]](#)
12. Carvalho, G.A.; Minnett, P.J.; Fleming, L.E.; Banzon, V.F.; Baringer, W. Satellite remote sensing of harmful algal blooms: A new multi-algorithm method for detecting the Florida Red Tide (*Karenia brevis*). *Harmful Algae* **2010**, *9*, 440–448. [\[CrossRef\]](#)
13. Magaña, H.A.; Contreras, C.; Villareal, T.A. A historical assessment of *Karenia brevis* in the western Gulf of Mexico. *Harmful Algae* **2003**, *2*, 163–171. [\[CrossRef\]](#)
14. Kutser, T. Passive optical remote sensing of cyanobacteria and other intense phytoplankton blooms in coastal and inland waters. *Int. J. Remote Sens.* **2009**, *17*, 4401–4425. [\[CrossRef\]](#)
15. Dierssen, H.M.; Kudela, R.M.; Ryan, J.P.; Zimmerman, R.C. Red and black tides: Quantitative analysis of water-leaving radiance and perceived color for phytoplankton, colored dissolved organic matter, and suspended sediments. *Limnol. Oceanogr.* **2006**, *51*, 2646–2659. [\[CrossRef\]](#)
16. Amin, R.; Zhou, J.; Gilerson, A.; Gross, B.; Moshary, F.; Ahmed, S. Novel optical techniques for detecting and classifying toxic dinoflagellate *Karenia brevis* blooms using satellite imagery. *Opt. Express* **2009**, *17*, 9126–9144. [\[CrossRef\]](#)
17. Haywood, A.J.; Steidinger, K.A.; Truby, E.W.; Bergquist, P.R.; Bergquist, P.L.; Adamson, J.; MacKenzie, L. Comparative morphology and molecular phylogenetic analysis of three new species of the genus *Karenia* (Dinophyceae) from New Zealand. *J. Phycol.* **2004**, *40*, 165–179. [\[CrossRef\]](#)
18. Kirkpatrick, B.; Fleming, L.E.; Squicciarini, D.; Backer, L.C.; Clark, R.; Abraham, W.; Benson, J.; Cheng, Y.S.; Johnson, D.; Pierce, R.; et al. Literature review of Florida red tide: Implications for human health effects. *Harmful Algae* **2004**, *3*, 99–115. [\[CrossRef\]](#)
19. Ross, C.; Ritson-Williams, R.; Pierce, R.; Bullington, J.B.; Henry, M.; Paul, V.J. Effects of the Florida red tide dinoflagellate, *Karenia brevis*, on oxidative stress and metamorphosis of larvae of the coral *Porites astreoides*. *Harmful Algae* **2010**, *9*, 173–179. [\[CrossRef\]](#)
20. Landsberg, J.H. The effects of harmful algal blooms on aquatic organisms. *Rev. Fish. Sci.* **2002**, *2*, 113–390. [\[CrossRef\]](#)
21. Fleming, L.E.; Kirkpatrick, B.; Backer, L.C.; Walsh, C.J.; Nierenberg, K.; Clark, J.; Reich, A.; Hollenbeck, J.; Benson, J.; Cheng, Y.S.; et al. Review of Florida red tide and human health effects. *Harmful Algae* **2011**, *10*, 224–233. [\[CrossRef\]](#)
22. Fleming, L.E.; McDonough, N.; Austen, M.; Mee, L.; Moore, M.; Hess, P.; Depledge, M.H.; White, M.; Philippart, K.; Bradbrook, P.; et al. Oceans and human health: A rising tide of challenges and opportunities for Europe. *Mar. Environ. Res.* **2014**, *99*, 16–19. [\[CrossRef\]](#)
23. Dyson, K.; Huppert, D.D. Regional economic impacts of razor clam beach closures due to harmful algal blooms (HABs) on the Pacific coast of Washington. *Harmful Algae* **2010**, *9*, 264–271. [\[CrossRef\]](#)
24. Stauffer, B.A.; Bowers, H.A.; Buckley, E.; Davis, T.W.; Johengen, T.H.; Kudela, R.; McManus, M.A.; Purcell, H.; Smith, G.J.; Vander Woude, A.; et al. Considerations in harmful algal bloom research and monitoring: Perspectives from a consensus-building workshop and technology testing. *Front. Mar. Sci.* **2019**, *6*. [\[CrossRef\]](#)
25. Thomas, A.C.; Townsend, D.W.; Weatherbee, R. Satellite-measured phytoplankton variability in the Gulf of Maine. *Cont. Shelf Res.* **2003**, *23*, 971–989. [\[CrossRef\]](#)
26. Vargo, G.A. A brief summary of the physiology and ecology of *Karenia brevis* Davis (G. Hansen and Moestrup comb. nov.) red tides on the West Florida Shelf and of hypotheses posed for their initiation, growth, maintenance, and termination. *Harmful Algae* **2009**, *8*, 573–584. [\[CrossRef\]](#)
27. Howarth, R.W.; Billen, G.; Swaney, D.; Townsend, A.; Jaworski, N.; Lajtha, K.; Downing, J.A.; Elmgren, R.; Caraco, N.; Jordan, T.; et al. Regional nitrogen budgets and riverine N & P fluxes for the drainages to the North Atlantic Ocean: Natural and human influences. *Biogeochemistry* **1996**, *35*, 75–139. [\[CrossRef\]](#)
28. Boesch, D.F.; Boynton, W.R.; Crowder, L.B.; Diaz, R.J.; Howarth, R.W.; Mee, L.D.; Nixon, S.W.; Rabalais, N.N.; Rosenberg, R.; Sanders, J.G.; et al. Nutrient enrichment drives Gulf of Mexico hypoxia. *Eos* **2009**, *90*, 117–118. [\[CrossRef\]](#)
29. Pinet, P.R. *Invitation to Oceanography*; Jones & Bartlett Publishers: Burlington, MA, USA, 2009; ISBN 1449667988.
30. Tomlinson, M.C.; Wynne, T.T.; Stumpf, R.P. An evaluation of remote sensing techniques for enhanced detection of the toxic dinoflagellate, *Karenia brevis*. *Remote Sens. Environ.* **2009**, *113*, 598–609. [\[CrossRef\]](#)
31. Hu, C.; Muller-Karger, F.E.; Vargo, G.A.; Neely, M.B.; Johns, E. Linkages between coastal runoff and the Florida Keys ecosystem: A study of a dark plume event. *Geophys. Res. Lett.* **2004**, *31*. [\[CrossRef\]](#)
32. Anderson, D.M. Approaches to monitoring, control and management of harmful algal blooms (HABs). *Ocean Coast. Manag.* **2009**, *52*, 342–347. [\[CrossRef\]](#)
33. Lee, J.H.W.; Hodgkiss, I.J.; Wong, K.T.M.; Lam, I.H.Y. Real time observations of coastal algal blooms by an early warning system. *Estuar. Coast. Shelf Sci.* **2005**, *65*, 172–190. [\[CrossRef\]](#)
34. Kamangir, H.; Collins, W.; Tissot, P.; King, S.A.; Dinh, H.T.H.; Durham, N.; Rizzo, J. FogNet: A multiscale 3D CNN with double-branch dense block and attention mechanism for fog prediction. *Mach. Learn. Appl.* **2021**, *5*, 100038. [\[CrossRef\]](#)

35. Park, S.; Lee, S.R. Red tides prediction system using fuzzy reasoning and the ensemble method. *Appl. Intell.* **2014**, *40*, 244–255. [\[CrossRef\]](#)
36. Craig, S.E.; Lohrenz, S.E.; Lee, Z.; Mahoney, K.L.; Kirkpatrick, G.J.; Schofield, O.M.; Steward, R.G. Use of hyperspectral remote sensing reflectance for detection and assessment of the harmful alga, *Karenia brevis*. *Appl. Opt.* **2006**, *45*, 5414–5425. [\[CrossRef\]](#)
37. Kislik, C.; Dronova, I.; Kelly, M. UAVs in support of algal bloom research: A review of current applications and future opportunities. *Drones* **2018**, *2*, 35. [\[CrossRef\]](#)
38. Sakuno, Y.; Maeda, A.; Mori, A.; Ono, S.; Ito, A. A simple red tide monitoring method using sentinel-2 data for sustainable management of Brackish Lake Koyama-ike, Japan. *Water* **2019**, *11*, 1044. [\[CrossRef\]](#)
39. Klemas, V. Remote sensing of algal blooms: An overview with case studies. *J. Coast. Res.* **2012**, *278*, 34–43. [\[CrossRef\]](#)
40. Seydi, S.T.; Akhoondzadeh, M.; Amani, M.; Mahdavi, S. Wildfire damage assessment over Australia using Sentinel-2 imagery and modis land cover product within the Google Earth engine cloud platform. *Remote Sens.* **2021**, *13*, 220. [\[CrossRef\]](#)
41. Ghannadi, M.A.; Alebooye, S.; Izadi, M.; Moradi, A. A method for Sentinel-1 DEM outlier removal using 2-D Kalman filter. *Geocarto Int.* **2020**, *35*, 1–15. [\[CrossRef\]](#)
42. Ghannadi, M.A.; SaadatSeresht, M.; Izadi, M.; Alebooye, S. Optimal texture image reconstruction method for improvement of SAR image matching. *IET Radar Sonar Navig.* **2020**, *14*, 1229–1235. [\[CrossRef\]](#)
43. Gower, J.; King, S.; Goncalves, P. Global monitoring of plankton blooms using meris MCI. *Int. J. Remote Sens.* **2008**, *29*, 6209–6216. [\[CrossRef\]](#)
44. Xing, X.G.; Zhao, D.Z.; Liu, Y.G.; Yang, J.H.; Xiu, P.; Wang, L. An overview of remote sensing of chlorophyll fluorescence. *Ocean Sci. J.* **2007**, *42*, 49–59. [\[CrossRef\]](#)
45. 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**, *129*, 637–652. [\[CrossRef\]](#)
46. Siswanto, E.; Ishizaka, J.; Tripathy, S.C.; Miyamura, K. Detection of harmful algal blooms of *Karenia mikimotoi* using MODIS measurements: A case study of Seto-Inland Sea, Japan. *Remote Sens. Environ.* **2013**, *129*, 185–196. [\[CrossRef\]](#)
47. Bernard, S.; Balt, C.; Pitcher, G.; Probyn, T.; Fawcett, A.; Du Randt, A. The use of MERIS for harmful algal bloom monitoring in the Southern Benguela. In Proceedings of the MERIS (A)ATSR Workshop 2005 (ESA SP-597), Frascati, Italy, 26–30 September 2005.
48. Cao, Z.; Ma, R.; Duan, H.; Pahlevan, N.; Melack, J.; Shen, M.; Xue, K. A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes. *Remote Sens. Environ.* **2020**, *248*, 111974. [\[CrossRef\]](#)
49. Moore, T.S.; Campbell, J.W.; Dowell, M.D. A class-based approach to characterizing and mapping the uncertainty of the MODIS ocean chlorophyll product. *Remote Sens. Environ.* **2009**, *113*, 2424–2430. [\[CrossRef\]](#)
50. Elkadiri, R.; Manche, C.; Sultan, M.; Al-Dousari, A.; Uddin, S.; Chouinard, K.; Abotalib, A.Z. Development of a coupled spatiotemporal algal bloom model for coastal areas: A remote sensing and data mining-based approach. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 5159–5171. [\[CrossRef\]](#)
51. Song, W.; Dolan, J.M.; Cline, D.; Xiong, G. Learning-based algal bloom event recognition for oceanographic decision support system using remote sensing data. *Remote Sens.* **2015**, *7*, 13564–13585. [\[CrossRef\]](#)
52. Gokaraju, B.; King, R.L.; Durbha, S.S.; Younan, N.H. A Machine Learning Based Spatio-Temporal Data Mining Approach for Detection of Harmful Algal Blooms in the Gulf of Mexico. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *4*, 710–720. [\[CrossRef\]](#)
53. 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. [\[CrossRef\]](#)
54. Hill, P.R.; Kumar, A.; Temimi, M.; Bull, D.R. HABNet: Machine learning, remote sensing-based detection of harmful algal blooms. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3229–3239. [\[CrossRef\]](#)
55. Karki, S.; Sultan, M.; Elkadiri, R.; Elbayoumi, T. Mapping and forecasting onsets of harmful algal blooms using MODIS data over coastalwaters surrounding charlotte county, Florida. *Remote Sens.* **2018**, *10*, 1656. [\[CrossRef\]](#)
56. Franks, P.J.S. Models of harmful algal blooms. *Limnol. Oceanogr.* **1997**, *42*, 1273–1282. [\[CrossRef\]](#)
57. Trombetta, T.; Vidussi, F.; Mas, S.; Parin, D.; Simier, M.; Mostajir, B. Water temperature drives phytoplankton blooms in coastal waters. *PLoS ONE* **2019**, *14*, e0214933. [\[CrossRef\]](#)
58. Lotliker, A.A.; Baliarsingh, S.K.; Samanta, A.; Varaprasad, V. Growth and decay of high-biomass algal bloom in the Northern Arabian Sea. *J. Indian Soc. Remote Sens.* **2020**, *48*, 465–471. [\[CrossRef\]](#)
59. Izadi, M.; Sultan, M.; Elkadiri, R.; Ghannadi, M.A.; Nikraftar, Z.; Namjoo, F. Remote sensing and statistical learning approach to harmful algal bloom forecasting using MODIS ocean colour parameters. In Proceedings of the AGU Fall Meeting Abstracts; 2020. Available online: <https://ui.adsabs.harvard.edu/abs/2020AGUFMIN011..091/abstract> (accessed on 5 September 2021).
60. Zolfaghari, A.; Izadi, M. Burst Pressure Prediction of Cylindrical Vessels Using Artificial Neural Network. *J. Press. Vessel Technol. Trans. ASME* **2020**, *142*, 1–7. [\[CrossRef\]](#)
61. Izadi, M.; Mohammadzadeh, A.; Haghighattalab, A. A new neuro-fuzzy approach for post-earthquake road damage assessment using GA and SVM classification from QuickBird satellite images. *J. Indian Soc. Remote Sens.* **2017**, *45*, 965–977. [\[CrossRef\]](#)
62. Recknagel, F.; Michener, W. *Ecological Informatics: Data Management and Knowledge Discovery*; Springer: Berlin, Germany, 2017.
63. Kim, D.; Jeong, K.; McKay, R.; Chon, T.; Joo, G. Machine learning for predictive management: Short and long term prediction of phytoplankton biomass using genetic algorithm based recurrent neural networks. *Int. J. Environ. Res.* **2012**, *6*, 95–108.

64. Kim, S. A multiple process univariate model for the prediction of chlorophyll-a concentration in river systems. *Int. J. Limnol.* **2016**, *56*, 137–150. [[CrossRef](#)]
65. Cho, H.; Choi, U.; Park, H. Deep learning application to time-series prediction of daily chlorophyll-a concentration. *Wit. Trans. Ecol. Environ.* **2018**, *215*, 163–175. [[CrossRef](#)]
66. Lee, S.; Lee, D. Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1322. [[CrossRef](#)]
67. Malek, S.; Salleh, A.; Milow, P.; Baba, M.; Sharifah, S. Applying artificial neural network theory to exploring diatom abundance at tropical Putrajaya lake, Malaysia. *J. Freshw. Ecol.* **2012**, *27*, 211–227. [[CrossRef](#)]
68. Daghighi, A. Harmful Algae Bloom Prediction Model for Western Lake Erie Using Stepwise Multiple Regression and Genetic Programming. Master's Thesis, Cleveland State University, Cleveland, OH, USA, 2017.
69. Qin, M.; Li, Z.; Du, Z. Red tide time series forecasting by combining ARIMA and deep belief network. *Knowl.-Based Syst.* **2017**, *125*, 39–52. [[CrossRef](#)]
70. McGowan, J.A.; Deyle, E.R.; Ye, H.; Carter, M.L.; Perretti, C.T.; Seger, K.D.; Verneil, A.; Sugihara, G. Predicting coastal algal blooms in Southern California. *Ecology* **2017**, *98*, 1419–1433. [[CrossRef](#)]
71. Sheykhoum, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6308–6325. [[CrossRef](#)]
72. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2000; ISBN 978-1-4614-7137-0.
73. Lee, Z.P.; Weidemann, A.; Kindle, J.; Arnone, R.; Carder, K.L.; Davis, C. Euphotic zone depth: Its derivation and implication to ocean-color remote sensing. *J. Geophys. Res. Ocean.* **2007**, *112*. [[CrossRef](#)]
74. Behrenfeld, M.J.; Falkowski, P.G. A consumer's guide to phytoplankton primary productivity models. *Limnol. Oceanogr.* **1997**, *42*, 1479–1491. [[CrossRef](#)]
75. Behrenfeld, M.J.; Boss, E.; Siegel, D.A.; Shea, D.M. Carbon-based ocean productivity and phytoplankton physiology from space. *Glob. Biogeochem. Cycles* **2005**, *19*. [[CrossRef](#)]
76. Anderson, D.M.; Glibert, P.M.; Burkholder, J.M. Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. *Estuaries* **2002**, *25*, 704–726. [[CrossRef](#)]
77. Morel, A.; Huot, Y.; Gentili, B.; Werdell, P.J.; Hooker, S.B.; Franz, B.A. Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. *Remote Sens. Environ.* **2007**, *111*, 69–88. [[CrossRef](#)]
78. Wang, G.; Lee, Z.; Mouw, C. Multi-spectral remote sensing of phytoplankton pigment absorption properties in cyanobacteria bloom waters: A regional example in the western basin of Lake Erie. *Remote Sens.* **2017**, *9*, 1309. [[CrossRef](#)]
79. Hoepffner, N.; Sathyendranath, S. Effect of pigment composition on absorption properties of phytoplankton. *Mar. Ecol. Prog. Ser.* **1991**, *73*, 11–23. [[CrossRef](#)]
80. O'Reilly, J.; Maritorena, S. Ocean color chlorophyll a algorithms for SeaWiFS, OC2, and OC4: Version 4. In *SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3*; NASA Center for Aerospace Information: Mountain View, CA, USA, 2000.
81. Maritorena, S.; Siegel, D.A.; Peterson, A.R. Optimization of a semianalytical ocean color model for global-scale applications. *Appl. Opt.* **2002**, *41*, 2705–2714. [[CrossRef](#)]
82. Lacava, T.; Ciancia, E.; Di Polito, C.; Madonia, A.; Pascucci, S.; Pergola, N.; Piermattei, V.; Satriano, V.; Tramutoli, V. Evaluation of MODIS-Aqua chlorophyll-a algorithms in the Basilicata Ionian Coastal waters. *Remote Sens.* **2018**, *10*, 987. [[CrossRef](#)]
83. Shang, S.L.; Dong, Q.; Hu, C.M.; Lin, G.; Li, Y.H.; Shang, S.P. On the consistency of MODIS chlorophyll products in the northern South China Sea. *Biogeosciences* **2014**, *11*, 269–280. [[CrossRef](#)]
84. Mishra, D.R.; Narumalani, S.; Rundquist, D.; Lawson, M. Characterizing the vertical diffuse attenuation coefficient for downwelling irradiance in coastal waters: Implications for water penetration by high resolution satellite data. *ISPRS J. Photogramm. Remote Sens.* **2005**, *60*, 48–64. [[CrossRef](#)]
85. Chen, J.; Cui, T.; Tang, J.; Song, Q. Remote sensing of diffuse attenuation coefficient using MODIS imagery of turbid coastal waters: A case study in Bohai Sea. *Remote Sens. Environ.* **2014**, *140*, 78–93. [[CrossRef](#)]
86. Ghanea, M.; Moradi, M.; Kabiri, K. A novel method for characterizing harmful algal blooms in the Persian Gulf using MODIS measurements. *Adv. Space Res.* **2016**, *58*, 1348–1361. [[CrossRef](#)]
87. Lee, Z.P.; Du, K.P.; Arnone, R. A model for the diffuse attenuation coefficient of downwelling irradiance. *J. Geophys. Res. C Ocean.* **2005**, *110*. [[CrossRef](#)]
88. Goldman, J.C.; Carpenter, E.J. A kinetic approach to the effect of temperature on algal growth. *Limnol. Oceanogr.* **1974**, *19*, 756–766. [[CrossRef](#)]
89. Hallegraeff, G.M. Ocean climate change, phytoplankton community responses, and harmful algal blooms: A formidable predictive challenge. *J. Phycol.* **2010**, *46*, 220–235. [[CrossRef](#)]
90. Bricaud, A.; Bosc, E.; Antoine, D. Algal biomass and sea surface temperature in the Mediterranean Basin: Intercomparison of data from various satellite sensors, and implications for primary production estimates. *Remote Sens. Environ.* **2002**, *81*, 163–178. [[CrossRef](#)]

91. Errera, R.M.; Yvon-Lewis, S.; Kessler, J.D.; Campbell, L. Responses of the dinoflagellate *Karenia brevis* to climate change: PCO<sub>2</sub> and sea surface temperatures. *Harmful Algae* **2014**, *37*, 110–116. [\[CrossRef\]](#)
92. Sarma, Y.V.B.; Al-Hashmi, K.; Smith, S.L. Sea surface warming and its implications for harmful algal blooms off Oman. *Int. J. Mar. Sci.* **2013**, *3*, 65–71. [\[CrossRef\]](#)
93. Hu, C.; Muller-Karger, F.E.; Taylor, C.; Carder, K.L.; Kelble, C.; Johns, E.; Heil, C.A. Red tide detection and tracing using MODIS fluorescence data: A regional example in SW Florida coastal waters. *Remote Sens. Environ.* **2005**, *97*, 311–321. [\[CrossRef\]](#)
94. El-habashi, A.; Ioannou, I.; Tomlinson, M.C.; Stumpf, R.P.; Ahmed, S. Satellite retrievals of *Karenia brevis* harmful algal blooms in the West Florida Shelf using neural networks and comparisons with other techniques. *Remote Sens.* **2016**, *8*, 377. [\[CrossRef\]](#)
95. Neville, R.A.; Gower, J.F.R. Passive remote sensing of phytoplankton via chlorophyll  $\alpha$  fluorescence. *J. Geophys. Res.* **1977**, *82*, 3487–3493. [\[CrossRef\]](#)
96. Zhao, J.; Hu, C.; Lenes, J.M.; Weisberg, R.H.; Lembke, C.; English, D.; Wolny, J.; Zheng, L.; Walsh, J.J.; Kirkpatrick, G. Three-dimensional structure of a *Karenia brevis* bloom: Observations from gliders, satellites, and field measurements. *Harmful Algae* **2013**, *29*, 22–30. [\[CrossRef\]](#)
97. Cannizzaro, J.P.; Hu, C.; English, D.C.; Carder, K.L.; Heil, C.A.; Müller-Karger, F.E. Detection of *Karenia brevis* blooms on the west Florida shelf using in situ backscattering and fluorescence data. *Harmful Algae* **2009**, *8*, 898–909. [\[CrossRef\]](#)
98. Lee, Z.; Carder, K.L.; Arnone, R.A. Deriving inherent optical properties from water color: A multiband quasi-analytical algorithm for optically deep waters. *Appl. Opt.* **2002**, *41*, 5755–5772. [\[CrossRef\]](#)
99. Davies-Colley, R.J.; Smith, D.G. Turbidity, suspended sediment, and water clarity: A review. *J. Am. Water Resour. Assoc.* **2001**, *37*, 1085–1101. [\[CrossRef\]](#)
100. Roelke, D.; Buyukate, Y. The diversity of harmful algal bloom-triggering mechanisms and the complexity of bloom initiation. *Hum. Ecol. Risk Assess.* **2001**, *7*, 1347–1362. [\[CrossRef\]](#)
101. May, C.L.; Koseff, J.R.; Lucas, L.V.; Cloern, J.E.; Schoellhamer, D.H. Effects of spatial and temporal variability of turbidity on phytoplankton blooms. *Mar. Ecol. Prog. Ser.* **2003**, *254*, 111–128. [\[CrossRef\]](#)
102. Morel, A.; Bélanger, S. Improved detection of turbid waters from ocean color sensors information. *Remote Sens. Environ.* **2006**, *102*, 237–249. [\[CrossRef\]](#)
103. Brand, L.E.; Compton, A. Long-term increase in *Karenia brevis* abundance along the Southwest Florida coast. *Harmful Algae* **2007**, *6*, 232–252. [\[CrossRef\]](#)
104. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [\[CrossRef\]](#)
105. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
106. Chen, T.; He, T.; Benesty, M. *Xgboost: eXtreme Gradient Boosting*, R Package version 0.71-2; Grin Verlag: München, Germany, 2018; pp. 1–4.
107. Hastie, T.; Tibshirani, R.; James, G.; Witten, D. *An Introduction to Statistical Learning, Springer Texts*; Springer: Berlin, Germany, 2006; ISBN 9780387781884.
108. Klusowski, J.M. Complete analysis of a random forest model. *arXiv* **2018**, arXiv:1005.0208.
109. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
110. Ho, T.K. Random decision forests. *Proc. Int. Conf. Doc. Anal. Recognit.* **1995**, *1*, 278–282. [\[CrossRef\]](#)
111. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [\[CrossRef\]](#)
112. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
113. Kim, T.-J. Prevention of harmful algal blooms by control of growth parameters. *Adv. Biosci. Biotechnol.* **2018**, *09*, 613–648. [\[CrossRef\]](#)
114. Zhang, M.; Niu, Z.; Cai, Q.; Xu, Y.; Qu, X. Effect of water column stability on surface chlorophyll and time lags under different nutrient backgrounds in a deep reservoir. *Water* **2019**, *11*, 1504. [\[CrossRef\]](#)
115. Jones, M. Forecasting algal bloom lags and stability in a watershed. *SIAM Undergrad. Res. Online* **2018**, *11*, 352–368. [\[CrossRef\]](#)