4-2022

A Remote Sensing and Machine Learning-Based Approach to Forecast the Onset of Harmful Algal Bloom (Red tides)

Moein Izadi
Western Michigan University, moeinizadi@gmail.com

A REMOTE SENSING AND MACHINE LEARNING-BASED APPROACH TO FORECAST THE ONSET OF HARMFUL ALGAL BLOOM

Moein Izadi, Ph.D.

Western Michigan University, 2022

In the last few decades, harmful algal blooms (HABs, also known as “red tides”) have become one of the most detrimental natural phenomena all around the world especially in Florida’s coastal areas due to local environmental factors and global warming in a larger scale. Karenia brevis produces toxins that have harmful effects on humans, fisheries, and ecosystems. In this study, I 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 such as Euphotic Depth (m) and Secchi disk depth, Chlorophyll-a (mg/m3), Diffuse attenuation coefficient (Kd_490; m−1), Sea surface temperature (C°) , Fluorescence line-height, … 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 three consecutive day satellite datasets, with a forecasting span from one 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.

A REMOTE SENSING AND MACHINE LEARNING-BASED APPROACH TO FORECAST THE ONSET OF HARMFUL ALGAL BLOOM

by

Moein Izadi

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Geological and Environmental Sciences
Western Michigan University
April 2022

Doctoral Committee:

Mohamed Sultan, Ph.D., Chair
Matt Reeves, Ph.D.
Racha El Kadiri, Ph.D.

ACKNOWLEDGMENTS

I offer sincere gratitude to my advisor Dr. Mohamed Sultan, without him, I would not have been able to complete this research and achieve other accomplishments that followed it such as my two internships at Precisely and Merck companies. I would like to remember all my lab mates at the Earth Science Remote Sensing Laboratory who were there to encourage me in their own ways! I am forever thankful to my parents and brothers for the unconditional love they poured over me and for their support for all the decisions, I made throughout the years.

At last, I would like to thank someone very special who came the last but provided the most meaning and impact to my work. I thank myself, for being such and incredible dedicated and determined person on his goals! For accepting his life biggest challenge and put behind all his friends, family, and memories to come to the land of opportunities to born and grow up again!

Moein Izadi

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUTION

For the last few decades, Harmful Algal Bloom (HAB; Karenia Brevis formerly known as Gymnodinium breve and Ptychodiscus brevis) has exponentially become one of the most deteriorative natural phenomena in Charlotte County, southwestern Florida, and even globally. Algae can adversely affect fresh and saltwater ecosystems and produce toxins that have harmful effects on human's health, fish industry, marine mammals, birds, and local economies. In other words, Algal blooms can significantly change the water bodies' quality like color, odor, and taste. This requires taking costly measurements like the closure of beaches and conducting costly filtration processes and decontamination activities. There are several environmental variables contributing to the propagation and exponentially growth of Algae's. For example, ever-growing adjacent agricultural activities are being transported into prone water bodies like bay areas by hydrodynamic processes such as infiltration and run off water. These provides Algae with favorable nutrients and as a result can adversely affect the biodiversity and habitats of aquatic ecosystems. Algal Bloom socio-economic importance is not needed more to be emphasized. Algal Bloom is directly affecting people's lives. That is why we need to make society aware of this problem and its urgency. This phenomenon is not a local problem and you can find Algae everywhere these days. Thus, there is a crucial need for mapping and forecasting HAB.

Earlier studies to address this problem are using different approaches. Some used real-time field monitoring of chlorophyll and dissolved oxygen, some used wind-driven and hydrodynamic variables in water current models, and some used rate and volume of

flow, and upwelling-down welling pulses. These models were used mostly for same-day mapping and to model onset of blooms. Later, they were developed to make early-warning systems for HAB forecasting and mapping. Such models depend on whether continuous real-time and archival field data or remotely sensed data which means they are not always available shortly after data acquisition for marine and coastal areas. Some more recent studies have used remotely sensed satellite data like Moderate Resolution Imaging Spectroradiometer (MODIS)-derived fluorescence along with field data to make data-driven statistical models (e, g,. Multiple Linear Regression model) to identify factors controlling HAB propagation. They provided a same-day distribution (now casting) and forecast their occurrences up to three days ahead of time. However, with such models there might be some deficiencies in model interpretation because of not addressing different lag times of different variables contributing to Algal Bloom. Sometimes addressing multicollinearity stay a challenge in linear models. Moreover, such models in data collection strategy and training phase more consider spatial variations within their variables (Spatio-Temporal vs Spatial). There is always a need to have a good statistical metrics to evaluate the model's performances and having different statistical models and comparing them can add values and reliability to the study results.

Monitoring HAB requires extensive field-based observation and measurements that are not available in most of the vulnerable areas. Fortunately, recent advances in remote sensing hold the promise to address these inadequacies. Models in Earth and Ocean Sciences need to be interpretable to show the significance and the correlations between controlling factors. For example, Artificial Neural Networks models are powerful functions for modeling real-world problems. However, it performs as a black box and the

neuron connections, their weights and different layers cannot be associated much with the concept of your physical problem in hand. On top of that, since the paucity of data has always been one of the challenges in front of researchers, a good model should work and be trained with limited dataset. The advantage of the proposed method is that it is addressing the before mentioned shortcomings and it can give a good range of forecasting time because of its spatio-temporal features.

CHAPTER 2

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 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].

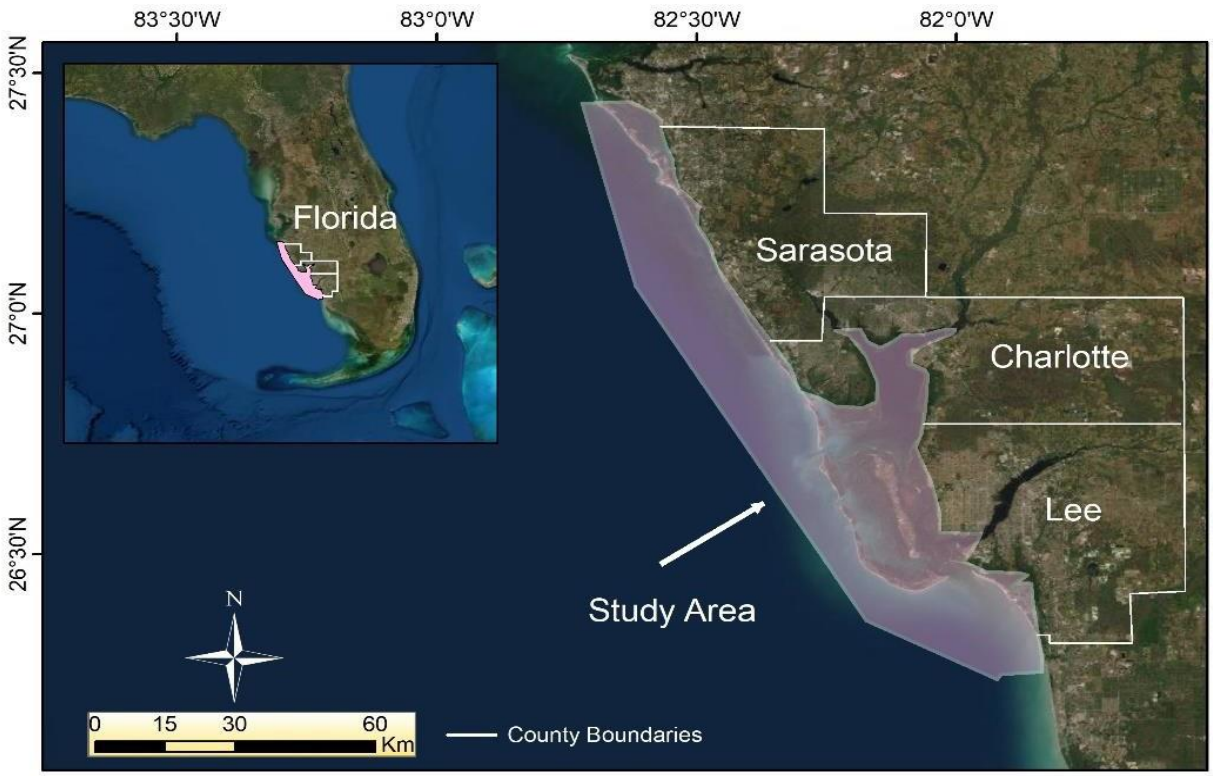

Figure 1. Location map for the study area covering coastal waters (width: ~10–60 km; length: 180 km) of Charlotte County in southwest Florida.

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 a research team mentioned in the reference section [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, first I 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, then I 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.

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.

Data and Methods

In this study, I 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.

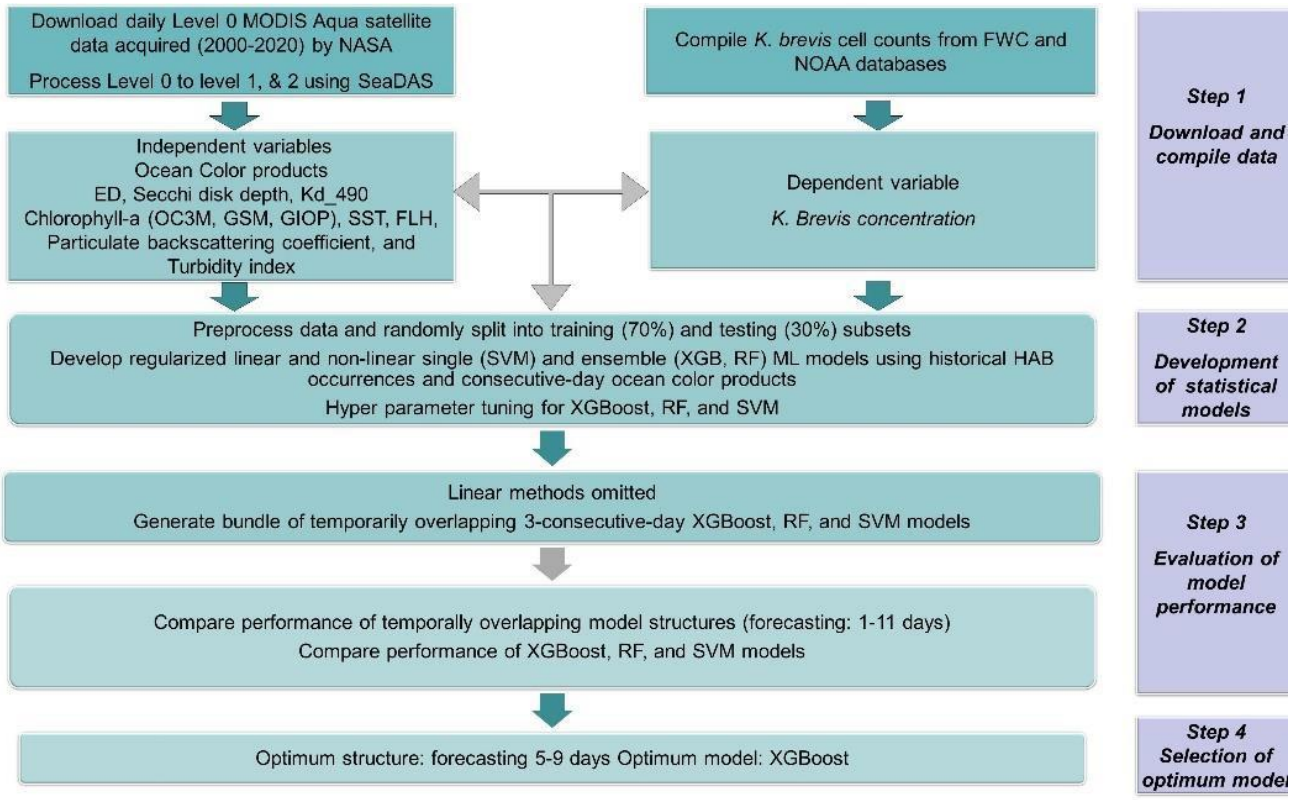

Figure 2. Flowchart describing the adopted methodology.

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.

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 backscattering 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.

1.1.1.1. 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.

1.1.1.2. Chlorophyll-a (mg/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].

1.1.1.3. Diffuse attenuation coefficient (Kd_490; 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 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 Kd_490 were compared; a high correlation was observed between chlorophyll-a and Kd_490 during red tides [86]. The Kd_490 was calculated using the technique described in [87].

1.1.1.4. Sea surface temperature (°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].

1.1.1.5. 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].

1.1.1.6. 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].

1.1.1.7. 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].

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].

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.

For data quality control, I tried to keep the dataset size the same for all different models when it comes to comparison among the models to avoid any bias towards model bias and variance.

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, I follow a common practice in which I utilize, compare, and contrast a set of statistical models described in the following sections.

Linear Models

We chose linear models because they have advantage in interpretability. By stacking up the same variables for different days (multicolinearity alert), I significantly increase the feature space (e.g., three consecutive days and 10 predictors for each day). 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].

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, I 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 ... X_n) are utilized to predict the corresponding target values Y (Y_1 ... 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].

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

where L is a differentiable convex loss function to measure the difference between the predicted values $\mathrm{F}(x_i)$ and the real target values ($_{Yi}$).

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

Then, we fit a classification tree, h, to $\frac{\partial \mathrm{j}}{\partial \mathrm{F}(xi)}$. The new updated (Xi) is $(Xi) + \gamma h$, where γ 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:

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

where

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

T is the number of tree leaves, and ω is the weights of those leaves. The function Ω penalizes the model complexity. The optimal weight ω of leaf j was calculated using (eq.4):

$$\omega_j = \frac{\sum_j gj}{\sum_j hj + \lambda}, \quad (4)$$

where $g_j = -\frac{\partial j}{\partial F(Xi)}$ and h is the j^{th} classification tree fitted to g_j. The optimal value of the loss function was calculated using (eq.5)[105]

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

The additional regularization term was added to avoid overfitting [104][105].

To achieve the optimum structure for the XGB model, parameters such as the number of boosting iterations, gamma, maximum depth, and learning rate (eta) were tuned. Gamma is a pseudo-regularization hyper parameter 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. I found that the optimum hyperparameters 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. But 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 [106]. Typically, *M* is calculated using eq. 6

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

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 the making of the forest and to enhance the variance-bias tradeoff [89][108]. 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. The OOB error was utilized to calculate the prediction error and to evaluate the variable importance measures [109]. 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 [92]. SVM proposes a nonparametric approach to finding linear discriminant functions. It tries to find a unique hyperplane between each pair of the classes in a multidimensional feature space [111].

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

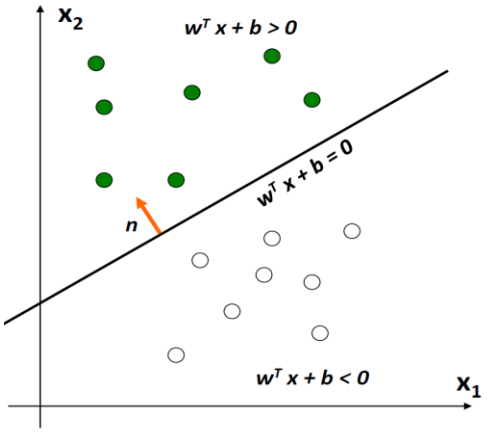

Figure 3. Two-dimensional feature space with SVM linear discrimination function.

Here, x_1 and x_2 are two features and the green and white classes are separated by the line $W^T x + b = 0$, n is normal vector of the hyperplane ($n = \frac{w}{||w||}$); where $\|w\|$ is the Euclidian distance between $\vec{w}$ and the origin. The main objective of SVM is to find a set of weights that specify two hyperplanes (eq. 7)

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

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

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: $Y_i (w^t xi + b) \geq 1$

$$\begin{cases} \min\left(\frac{1}{2}\, w^t\, \mathrm{w}\right) \\ y_i\,(\,w^t\, xi\, +\, b) \geq\, 1 \end{cases} \qquad (8)$$

By solving this optimization equation for *w* and *b*, each of the y_i data points is 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 [γ] 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 γ increase and underfits as their values decrease. The average number of support vectors, optimum γ, and optimum c were selected at 95, 0.5, and 0.1, respectively.

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, the 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 [106]. These criteria are calculated as follows [106]:

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

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

$$\text{Balanced\_a}curacy = \frac{\text{Specificity+Sensitivity}}{2} \quad (11)$$

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

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

where

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

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 one 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 were 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 (−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 (−7, −8, −9) model will be referred to hereafter as a 7-day model, a 3-consecutive (−9, −10, −11) model as a 9-day model, and a 3-consecutive (−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 serially; 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, each tree of an RF is not 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 sub-setting 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.

Results

4.1. Model structure comparison and selection of optimum model structure

Using the XGB model, I compared the performance of one 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 outperforms the single-day models. A similar exercise was conducted using the SVM model, and again the 3-consecutive 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 an 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**
Kappa	0.77	0.76	0.76	0.74	0.76	**0.80**
F-score	0.89	0.88	0.88	0.85	0.88	**0.96**
Precision	0.84	0.88	0.88	0.92	0.88	**0.94**
B. accuracy	87.0	86.0	86.0	83.0	86.0	**88.0**
Combination (2nd SVM)	**−2**	**−3**	**−4**	**−2, −3**	**−3, −4**	**−2, −3, −4**
Kappa	0.35	0.4	0.37	0.4	0.4	0.50
F-score	0.51	0.52	0.48	0.52	0.52	0.60
Precision	0.56	0.72	0.81	0.73	0.73	0.82
B. accuracy	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 3 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 20% (562 scenes), the 3-consecutive day scenes 9% (260 scenes), the 4-consecutive day scenes constituted <2% (57

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 days for model training purposes, I 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 provide 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 before the onset of a HAB occurrence; the last model uses data acquired 10, 11, and 12 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).

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

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

4.2. Comparison of performance of statistical models and selection of optimum model

A variety of machine learning algorithms were adopted based on the nature of data and the problem at 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 the 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 a HAB occurrence. The best metrics among the three models (XGB, RF, 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) indicate 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 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	**0.74**
									X	X	X		73.9	**0.65**	**0.82**	0.85
								X	X	X			**58.0**	**0.27**	**0.63**	0.67
							X	X	X				**76.4**	0.58	**0.78**	0.84
						X	X	X					**83.9**	**0.0.7**	**0.87**	**0.88**
					X	X	X						**92.0**	**0.86**	**0.95**	**0.97**
				X	X	X							87.6	**0.81**	**0.96**	**0.98**
			X	X	X								**96.2**	**0.93**	**0.98**	**0.98**
		X	X	X									**87.4**	**0.76**	**0.92**	**0.91**
	X	X	X										**83.6**	**0.71**	**0.88**	0.81
X	X	X											79.6	**0.68**	**0.80**	**0.80**
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	**0.74**	0.73
									X	X	X		**77.2**	0.63	0.71	**0.86**
								X	X	X			54.7	0.13	0.20	0.74
							X	X	X				76.1	0.60	0.73	**0.87**
						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							**91.4**	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	**0.87**
X	X	X											**80.7**	0.62	0.71	0.79
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			56.3	0.20	0.27	**0.79**
							X	X	X				73.7	**0.63**	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

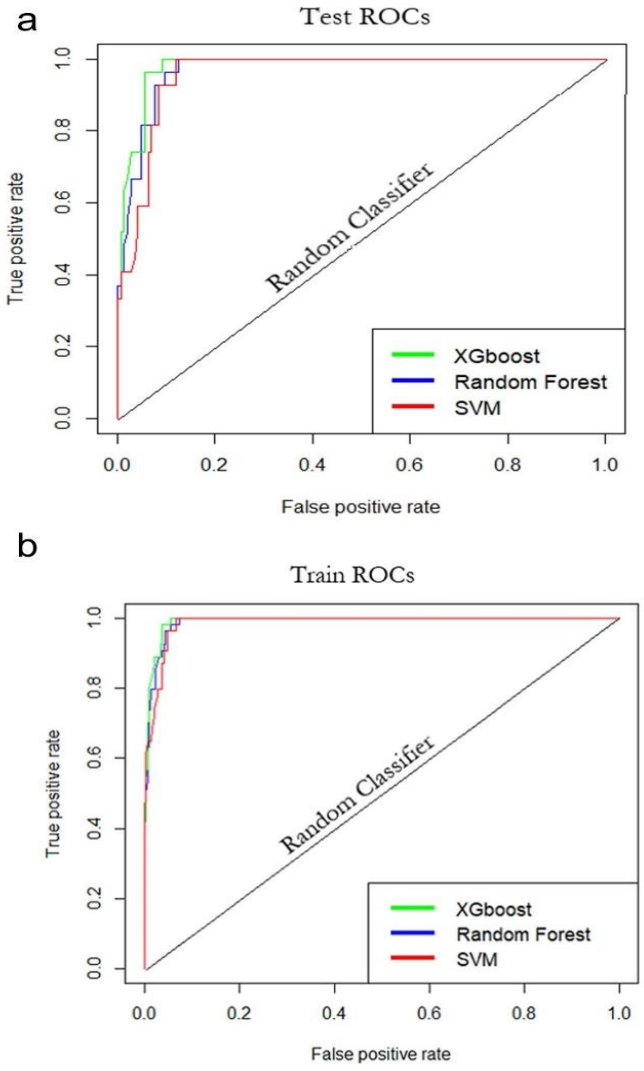

Figure 4. Machine Learning models ROCs for Test ROC (a) and Train ROC (b).

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 concentrations 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%.

XGBoost (the top sub table) 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.

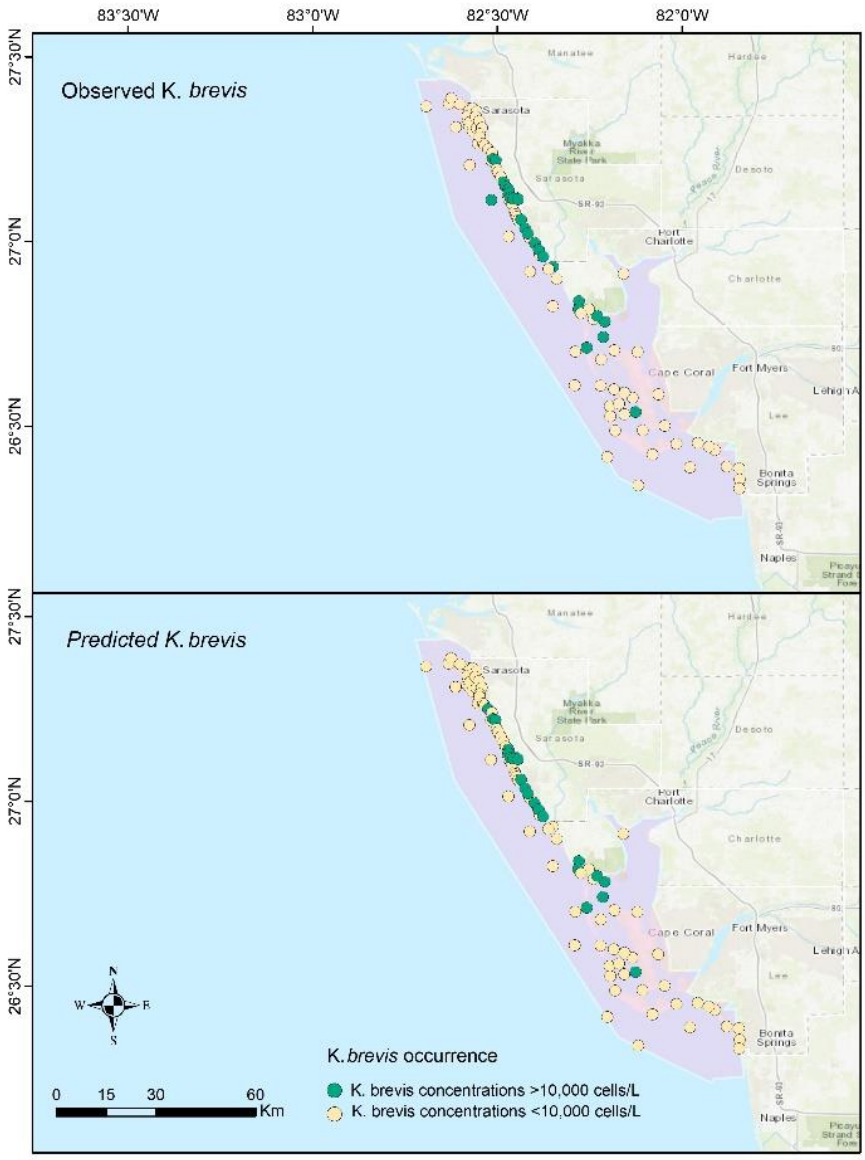

Figure 5. Comparison between the reported K. brevis concentration in 170 randomly selected test samples within the study area (top) with their predicted concentrations from an 8-day RF model (bottom). The samples have been classified correctly with 95% accuracy.

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 2.2). 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 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.

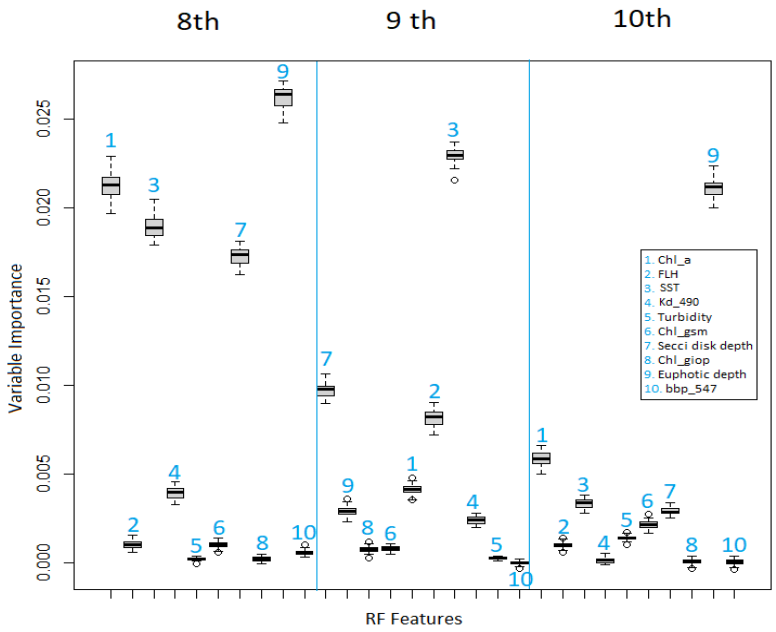

Figure 6. Variable importance (VI) boxplot for the best RF model. The VI metrics are on the y-axis and the 10 variables for each of the 3-consecutive days are numbered.

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. These are (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 single-day models [54]. 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 a 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 single-day or two-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 single-day model outputs [54] over the study area reveal enhanced overall performance for our 3-consecutive day models. The single-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_490, 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 day- 11 models. 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 before 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 [52][57][112]. In the Persian Sea, phytoplankton biomass was found to correlate with SST with a lag time of 2–8 days [57]; in Santa Monica, it was 5 days [113]. In addition, a 3–6 day lag time was reported between the introduction of phosphorus nutrients and the occurrence of HAB events [114].

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 three 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 data set) 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, 3- day, 4-day, 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.

Conclusion

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 was 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 the 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 wide scale applications of the developed concepts in areas outside of the study area require the 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 before feeding the data into the ML algorithms. The consecutive raster images potential for data imputation, interpolation for coming up with more consecutive days should also be investigated for the future works.

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

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

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**, doi:10.1016/S0278-4343(97)00022-8.

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. Bio. Ecol.* **2013**, doi:10.1016/j.jembe.2013.03.015.

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**, doi:10.1016/S0278-4343(03)00088-8.

4. Kahru, M.; Mitchell, B.G.; Diaz, A.; Miura, M. MODIS detects a devastating algal bloom in Paracas Bay, Peru. *Eos (Washington. DC).* **2004**, doi:10.1029/2004EO450002.

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**, doi:10.1016/j.pocean.2009.07.030.

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**, doi:10.1016/j.csr.2008.11.006.

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 the Persian Sea. *Remote Sens. Environ.* **2011**,

doi:10.1016/j.rse.2011.04.028.

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**, doi:10.1080/01431161.2010.545449.

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, doi:10.1016/j.pocean.2013.12.008.

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**, doi:10.1093/plankt/fbq086.

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**, doi:10.1016/S0924-7963(03)00030-7.

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**, doi:10.1016/j.hal.2010.02.002.

13. Magaña, H.A.; Contreras, C.; Villareal, T.A. A historical assessment of Karenia Brevis in the western Gulf of Mexico. *Harmful Algae* **2003**, doi:10.1016/S1568-9883(03)00026-X.

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, doi:10.1080/01431160802562305.

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**, doi:10.4319/lo.2006.51.6.2646.

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**, doi:10.1364/oe.17.009126.

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**, doi:10.1111/j.0022-3646.2004.02-149.x.

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 (2)*, 99–115, doi:10.1016/j.hal.2003.08.005.

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**, doi:10.1016/j.hal.2009.09.001.

20. Landsberg, J.H. The effects of harmful algal blooms on aquatic organisms. *Rev. Fish. Sci.* **2002**, *2*, 113–390, doi:10.1080/20026491051695.

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 (2)*, 224–233, doi:10.1016/j.hal.2010.08.006.

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**, doi:10.1016/j.marenvres.2014.05.010.

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**, doi:10.1016/j.hal.2009.11.003.

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*, doi:10.3389/fmars.2019.00399.

25. Thomas, A.C.; Townsend, D.W.; Weatherbee, R. Satellite-measured phytoplankton variability in the Gulf of Maine. *Cont. Shelf Res.* **2003**, doi:10.1016/S0278-4343(03)00086-4.

26. Vargo, G.A. A 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, doi:10.1016/j.hal.2008.11.002.

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**, doi:10.1007/BF02179825.

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 (Washington. DC).* **2009**, doi:10.1029/2009EO140001.

29. Pinet, P.R. *Invitation to Oceanography*; 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**, doi:10.1016/j.rse.2008.11.003.

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**, doi:10.1029/2004GL020382.

32. Anderson, D.M. Approaches to monitoring, control, and management of harmful algal blooms (HABs). *Ocean Coast. Manag.* **2009**, doi:10.1016/j.ocecoaman.2009.04.006.

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**, doi:10.1016/j.ecss.2005.06.005.

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. with Appl.* **2021**, *5*, 100038, doi:10.1016/j.mlwa.2021.100038.

35. Park, S.; Lee, S.R. Red tides prediction system using fuzzy reasoning and the ensemble method. *Appl. Intell.* **2014**, doi:10.1007/s10489-013-0457-1.

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**, doi:10.1364/AO.45.005414.

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(4)*, 35, doi:10.3390/drones2040035.

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-like, Japan. *Water* **2019**, *11*, 1044, doi:10.3390/w11051044.

39. Klemas, V. Remote Sensing of Algal Blooms: An Overview with Case Studies. *J. Coast. Res.* **2012**, *278*, 34–43, doi:10.2112/JCOASTRES-D-11-00051.1.

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*, 1–30, doi:10.3390/rs13020220.

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**, *0*, 1–15, doi:10.1080/10106049.2020.1815866.

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*, doi:10.1049/iet-rsn.2020.0058.

43. Gower, J.; King, S.; Goncalves, P. Global monitoring of plankton blooms using MERIS MCI. *Int. J. Remote Sens.* **2008**, *29*, 6209–6216, doi:10.1080/01431160802178110.

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.

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**, doi:10.1016/j.rse.2012.05.032.

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**, doi:10.1016/j.rse.2012.11.003.

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 European Space Agency, (Special Publication) ESA SP; 2005.

48. 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**, doi:10.1016/j.rse.2009.07.016.

49. 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, doi:10.1109/JSTARS.2016.2555898.

50. 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**, doi:10.3390/rs71013564.

51. 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**, doi:10.1109/JSTARS.2010.2103927.

52. 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**, doi:10.1080/01431161.2019.1706011.

53. 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, doi:10.1109/JSTARS.2020.3001445.

54. 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**, doi:10.3390/rs10101656.

55. Franks, P.J.S. Models of harmful algal blooms. *Limnol. Oceanogr.* **1997**, doi:10.4319/lo.1997.42.5_part_2.1273.

56. Trombetta, T.; Vidussi, F.; Mas, S.; Parin, D.; Simier, M.; Mostajir, B. Water temperature drives phytoplankton blooms in coastal waters. *PLoS One* **2019**, doi:10.1371/journal.pone.0214933.

57. Lotliker, A.A.; Baliarsingh, S.K.; Samanta, A.; Varaprasad, V. Growth and Decay of High-Biomass Algal Bloom in the Northern Persian Sea. *J. Indian Soc. Remote Sens.* **2020**, doi:10.1007/s12524-019-01094-3.

58. 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; Vol. 2020, pp. IN011--09.

59. Zolfaghari, A.; Izadi, M. Burst Pressure Prediction of Cylindrical Vessels Using Artificial Neural Network. *J. Press. Vessel Technol. Trans. ASME* **2020**, *142*, 1–7, doi:10.1115/1.4045729.

60. 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, doi:10.1007/s12524-017-0660-3.

61. Recknagel, F.; Michener, W. Ecological informatics: Data management and knowledge discovery. *Springer* **2017**.

62. 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 Env. Res* **2012**, *6*, 95–108.

63. Kim, S. A multiple process univariate model for the prediction of chlorophyll-a concentration in river systems. *Int J Limnol* **2016**, *56*, 137–150, doi:10.1051/limn/2016003.

64. Cho, H.; Choi, U.; Park, H. Deep learning application to time-series prediction of daily chlorophyll-a concentration. *Wit Trans Ecol Envir* **2018**, *215*, 175–163, doi:10.2495/EID180141.

65. Lee, S.; Lee, D. Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models. *Int J Env. Res Public Heal.* **2018**, *15(7)*, 1–15, doi:10.3390/ijerph15071322.

66. 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(2)*, 211–227, doi:10.1080/02705060.2011.635883.

67. Daghighi, A. Harmful algae bloom prediction model for Western Lake Erie using stepwise multiple regression and genetic programming. *Master's thesis, Cleveland, State, Univ.* **2017**.

68. 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, doi:10.1016/j.knosys.2017.03.027.

69. McGowan, J.; Deyle, E.; Ye, H.; Carter, M.; Perretti, C.; Seger, K.; De, V.; A, S. Predicting coastal algal blooms in southern California. *Ecology* **2017**, *98(5)*, 1419–1433, doi:10.1002/ecy.1804.

70. Sheykhmousa, 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, doi:10.1109/JSTARS.2020.3026724.

71. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to Statistical Learning*; 2000; ISBN 978-1-4614-7137-0.

72. 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**, doi:10.1029/2006JC003802.

73. Behrenfeld, M.J.; Falkowski, P.G. A consumer's guide to phytoplankton primary productivity models. *Limnol. Oceanogr.* **1997**, doi:10.4319/lo.1997.42.7.1479.

74. Behrenfeld, M.J.; Boss, E.; Siegel, D.A.; Shea, D.M. Carbon-based ocean productivity and phytoplankton physiology from space. *Global Biogeochem. Cycles* **2005**, doi:10.1029/2004GB002299.

75. Anderson, D.M.; Glibert, P.M.; Burkholder, J.M. Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. *Estuaries* **2002**, doi:10.1007/BF02804901.

76. 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**, doi:10.1016/j.rse.2007.03.012.

77. 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(12)*, 1309, doi:10.3390/rs9121309.

78. Hoepffner, N.; Sathyendranath, S. Effect of pigment composition on absorption properties

of phytoplankton. *Mar. Ecol. Prog. Ser.* **1991**, doi:10.3354/meps073011.

79. O'Reilly, J.; Maritorena, S. Ocean color chlorophyll a algorithms for SeaWiFS, OC2, and OC4: Version 4. *SeaWiFS postlaunch …* **2000**.

80. Maritorena, S.; Siegel, D.A.; Peterson, A.R. Optimization of a semianalytical ocean color model for global-scale applications. *Appl. Opt.* **2002**, doi:10.1364/ao.41.002705.

81. 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**, doi:10.3390/rs10070987.

82. 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, doi:10.5194/bg-11-269-2014.

83. 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**, doi:10.1016/j.isprsjprs.2005.09.003.

84. 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**, doi:10.1016/j.rse.2013.08.031.

85. Ghanea, M.; Moradi, M.; Kabiri, K. A novel method for characterizing harmful algal blooms in the Persian Gulf using MODIS measurements. *Adv. Sp. Res.* **2016**, doi:10.1016/j.asr.2016.06.005.

86. Lee, Z.P.; Du, K.P.; Arnone, R. A model for the diffuse attenuation coefficient of downwelling irradiance. *J. Geophys. Res. C Ocean.* **2005**, doi:10.1029/2004JC002275.

87. Goldman, J.C.; Carpenter, E.J. A kinetic approach to the effect of temperature on algal growth. *Limnol. Oceanogr.* **1974**, doi:10.4319/lo.1974.19.5.0756.

88. Hallegraeff, G.M. Ocean climate change, phytoplankton community responses, and harmful algal blooms: a formidable predictive challenge. *J. Phycol.* **2010**, *46*, 220–235, doi:10.1111/j.1529-8817.2010.00815.x.

89. 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, doi:10.1016/S0034-4257(01)00335-2.

90. Errera, R.M.; Yvon-Lewis, S.; Kessler, J.D.; Campbell, L. Reponses of the dinoflagellate Karenia brevis to climate change: PCO2 and sea surface temperatures. *Harmful Algae* **2014**, doi:10.1016/j.hal.2014.05.012.

91. Sarma, Y.V.B.; Al-Hashmi, K.; L. Smith, S. Sea Surface Warming and its Implications for Harmful Algal Blooms off Oman. *Int. J. Mar. Sci.* **2013**, doi:10.5376/ijms.2013.03.0008.

92. 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**, doi:10.1016/j.rse.2005.05.013.

93. 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**, doi:10.3390/rs8050377.

94. Neville, R.A.; Gower, J.F.R. Passive remote sensing of phytoplankton via chlorophyll α fluorescence. *J. Geophys. Res.* **1977**, *82*, 3487–3493, doi:10.1029/JC082i024p03487.

95. 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**, doi:10.1016/j.hal.2013.07.004.

96. 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**, doi:10.1016/j.hal.2009.05.001.

97. 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**, doi:10.1364/ao.41.005755.

98. Davies-Colley, R.J.; Smith, D.G. Turbidity, suspended sediment, and water clarity: A review. *J. Am. Water Resour. Assoc.* **2001**, *37*, 1085–1101, doi:10.1111/j.1752-1688.2001.tb03624.x.

99. Roelke, D.; Buyukates, Y. The diversity of harmful algal bloom-triggering mechanisms and the complexity of bloom initiation. *Hum. Ecol. Risk Assess.* **2001**, doi:10.1080/20018091095041.

100. 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**, doi:10.3354/meps254111.

101. Morel, A.; Bélanger, S. Improved detection of turbid waters from ocean color sensors information. *Remote Sens. Environ.* **2006**, doi:10.1016/j.rse.2006.01.022.

102. Brand, L.E.; Compton, A. Long-term increase in Karenia brevis abundance along the Southwest Florida coast. *Harmful Algae* **2007**, *6*, 232–252, doi:10.1016/j.hal.2006.08.005.

103. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, doi:10.1007/bf00058655.

104. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016.

105. Chen, T.; He, T.; Benesty, M. xgboost : eXtreme gradient boosting. *R Packag. version 0.71-2* **2018**, 1–4.

106. Hastie, T.; Tibshirani, R.; James, G.; Witten, D. *An Introduction to Statistical Learning, Springer Texts*; 2006; ISBN 9780387781884.

107. Klusowski, J.M. Complete Analysis of a Random Forest Model. *ArXiv* 2018.

108. Breiman, L. Random forests. *Mach. Learn.* **2001**, doi:10.1023/A:1010933404324.

109. Ho, T.K. Random decision forests. *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR* **1995**, doi:10.5555/844379.844681.

110. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Networks* **1999**, *10*, 988–999, doi:10.1109/72.788640.

111. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, doi:10.1023/A:1022627411411.

112. Kim, T.-J. Prevention of Harmful Algal Blooms by Control of Growth Parameters. *Adv. Biosci. Biotechnol.* **2018**, doi:10.4236/abb.2018.911043.

113. 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 (Switzerland)* **2019**, doi:10.3390/w11071504.

114. Jones, M. Forecasting Algal Bloom Lags and Stability in a Watershed. *SIAM Undergrad. Res. Online* **2018**, doi:10.1137/18s016643.

Appendix

In this section, I added the main body of code that I used for one of my models.

```
######################
###install.packages###
######################
install.packages('randomForest')
install.packages('gbm')
install.packages('tree')
install.packages('PerformanceAnalytics')
install.packages('tidyverse')
install.packages('caret')
install.packages('pROC')
install.packages('purrr')
install.packages('lattice')
install.packages('ggplot2')
```

```
install.packages('e1071')
install.packages('ROCR')
install.packages('precrec')
install.packages('ROCit')
install.packages("vip")
install.packages("ranger")
install.packages("xgboost")
install.packages('Ckmeans.1d.dp')
install.packages('party')
library(Ckmeans.1d.dp)
library(ggplot2)
library(precrec)
library(vip)
library(ranger)
library(ROCit)
library(precrec)
library(pROC)
library(tidyverse)
library(caret)
library(purrr)
library(lattice)
library(ggplot2)
library(e1071)
library(PerformanceAnalytics)
library(randomForest)
library(gbm)
library(tree)
library(ROCR)
library(xgboost)
library(party)
library(ROCR)
library(caTools)

####################
### Read  #########
####################
set.seed(500)
dataset = read.csv(file = "Cleaned_to_NA_3dys_MICE146.csv")
dataset = read.csv(file = "Cleaned_to_NA_3dys_MICE196.csv")
dataset = read.csv(file = "Cleaned_to_NA_3dys_MICE340.csv")
head(dataset)

### Factorizing #######
#dataset$Karenia_br= cut(dataset$Karenia_br, breaks=c(-1,100000,100000000), labels=c("not
present", "Present"))
dataset$Karenia_br= cut(dataset$Karenia_br, breaks=c(-1,10000,100000000), labels=c("0", "1"))

##### train/test split #######
set.seed(1116)
#split = sample.split(dataset$Karenia_br, SplitRatio = 0.8)
```

```
#training_set = subset(dataset, split == TRUE)
#test_set = subset(dataset, split == FALSE)
#test.Karenia_br <- test_set[111]
train <- sample(1:nrow(dataset), nrow(dataset)/2)
dataset.test <- dataset[-train, ]
test_set=dataset[-train, ]
training_set = dataset[train, ]
test.Karenia_br <- dataset$Karenia_br[-train]
test.Karenia_br <- dataset$Karenia_br

#####################################################################
####################
######### Random Forest (randomForest) ########
################################################
# m = sqrt{p} foR classification AND p/3 for regression
rf.dataset <- randomForest(Karenia_br ~., subset = train, data = dataset, mtry = 6,
                  importance=TRUE,importanceSD=TRUE, localImp=TRUE, ntree=1000)

###### Predictions on the TEST dataset
yhat.rf <- predict(rf.dataset, newdata = dataset[-train,], 'prob')
yhat.rf <- predict(rf.dataset, newdata = dataset, 'prob')
#plot(yhat.rf[,2], test.Karenia_br)
#abline(0, 1)
h=cbind(yhat.rf[,2], test.Karenia_br)[1:170,]
h
###### Variable importance plot 2 #####
#importance(rf.dataset, type=1)
#varImpPlot(rf.dataset)

#########################
## RF ConfusionMatrix ##
#########################
confusionMatrix(yhat.rf, test.Karenia_br, positive = "Present", mode="everything")
#error <- mean(test.Karenia_br != yhat.rf)



#####################################################################
#################################
#############################################################
#################### XGBOOST ##############################
#############################################################
classifier = xgboost(data = as.matrix(training_set[-1]), label =
as.numeric(levels(training_set$Karenia_br))[training_set$Karenia_br], type="response",
                  objective = "binary:logistic", #"binary:hinge",
                  nrounds = 100,
                  max_depth = 6,
                  eta = 0.05,
                  gamma = 0,
                  colsample_bytree = 0.1,
```

```
                min_child_weight =1,
                subsample = 1,
                verbose = 5
               )

# Predicting the Test set results
y_pred = predict(classifier, newdata = as.matrix(test_set[-1]), type="response", outputmargin=F)
y_pred = predict(classifier, newdata = as.matrix(dataset[-1]), type="response", outputmargin=F)


########################
## XGB ConfusionMatrix ##
########################
#confusionMatrix(as.factor(y_pred), as.factor(test.Karenia_br), positive = "Present",
mode="everything")
confusionMatrix(as.factor(y_pred), as.factor(test.Karenia_br), mode="everything")
h=cbind(factor(y_pred), factor(test.Karenia_br))[1:170,]
h
###################################
##### XGBoost Tuning ##############
###################################
set.seed(500)
dataset = read.csv(file = "Cleaned_to_NA_3dys_MICE340.csv")
dataset$Karenia_br= cut(dataset$Karenia_br, breaks=c(-1,10000,100000000),
labels=c("notpresent", "Present"))
##### train/test split #######
set.seed(1116)
split = sample.split(dataset$Karenia_br, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
test.Karenia_br <- test_set$Karenia_br

cv.ctrl <- trainControl(method = "repeatedcv", repeats = 1,number = 3,
                #summaryFunction = twoClassSummary,
                classProbs = TRUE,
                allowParallel=T)

xgb.grid <- expand.grid(nrounds = 100,
                eta = c(0.01,0.05,0.1),
                #eta = c(0.01,0.05),
                max_depth = c(2,4,6,8,10,14),
                #max_depth = c(2,4),
                gamma= c(0, 10),
                colsample_bytree= c(0.1, 0.4),
                min_child_weight= c(1L, 10L) ,
                subsample= c(0.5, 1)
                #colsample_bytree= 0.1,
                #min_child_weight= 1L ,
                #subsample= 0.5
)
```

```
set.seed(45)
xgb_tune <-train(training_set[,-1],
                 training_set$Karenia_br,
                 method="xgbTree",
                 trControl=cv.ctrl,
                 tuneGrid=xgb.grid,
                 verbose=T,
                 metric="AUC",
                 nthread =3
)
xgb_tune



#####################################################################
#################################
#############################################################
#################### Gradiant Bossting #################
#############################################################
aa=dataset[train,]
AAA=dataset[-train,]
AAA=AAA[,-1]
set.seed(102)
bst <- xgboost(
  data = data.matrix(subset(dataset[train,], select = -Karenia_br)),
  label = aa$Karenia_br,
  objective = "reg:linear",
  nrounds = 1000,
  max_depth = 100,
  eta = 0.4,
  verbose = 0  # suppress printing
)

# Predicting the Test set results
y_pred2 = predict(bst, newdata = as.matrix(AAA))
y_pred2 = predict(bst, newdata = as.matrix(dataset[,-1]))
#y_pred2 = (y_pred2 >= 0.5)


#####################################################################
#################################
#############################################################
#################### Support Vector Machine #################
#############################################################
svm.fit2 <- svm(Karenia_br ~ ., data = dataset[train, ], kernel = "radial", gamma = 1, cost = 1,
probability = TRUE)
svm.pred <- predict(svm.fit2, newdata = dataset[-train, ], probability = TRUE, type = "prob")
svm.pred <- predict(svm.fit2, newdata = dataset, probability = TRUE, type = "prob")
```

```
Prob=attr(svm.pred,"probabilities")[,2]



################################
############ ROC  #############
################################

######## ROCR- 2005 ############
#pred <- prediction(yhat.rf[,2], test.Karenia_br)
#perf <- performance(pred,"tpr","fpr")
#plot(perf,colorize=TRUE)

############ pROC - 2010 Data science ROC ################
############ Random Forest ROC
pROC_obj <- roc(test.Karenia_br,yhat.rf[,2],
            smoothed = TRUE,
            # arguments for ci
            ci=TRUE, ci.alpha=0.9, stratified=FALSE,
            # arguments for plot
            plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,
            print.auc=TRUE, show.thres=TRUE)
sens.ci <- ci.se(pROC_obj)
plot(sens.ci, type="shape", col="lightblue")
plot(sens.ci, type="bars")
###### ROCit - 2019 Youden index ########
ROCit_obj <- rocit(score=yhat.rf[,2],class=test.Karenia_br)
plot(ROCit_obj)

############ pROC - 2010 Data science ################
############ XGBoost ROC
pROC_obj <- roc(test.Karenia_br,y_pred,
            smoothed = TRUE,
            # arguments for ci
            ci=TRUE, ci.alpha=0.9, stratified=FALSE,
            # arguments for plot
            plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,
            print.auc=TRUE, show.thres=TRUE)
sens.ci <- ci.se(pROC_obj)
plot(sens.ci, type="shape", col="lightblue")
plot(sens.ci, type="bars")
###### ROCit - 2019 Youden index ########
ROCit_obj <- rocit(score=y_pred,class=test.Karenia_br)
plot(ROCit_obj)

############ pROC - 2010 Data science ################
############ GBM ROC
pROC_obj <- roc(test.Karenia_br,y_pred2,
            smoothed = TRUE,
            # arguments for ci
```

```
        ci=TRUE, ci.alpha=0.9, stratified=FALSE,
        # arguments for plot
        plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,
        print.auc=TRUE, show.thres=TRUE)
sens.ci <- ci.se(pROC_obj)
plot(sens.ci, type="shape", col="lightblue")
plot(sens.ci, type="bars")
###### ROCit - 2019 Youden index ########
ROCit_obj <- rocit(score=y_pred2,class=test.Karenia_br)
plot(ROCit_obj)

############ pROC - 2010 Data science ROC ################
############ SVM ROC
test.Karenia_br <- test_set$Karenia_br
test.Karenia_br=as.numeric(levels(test.Karenia_br))[test.Karenia_br]
pROC_obj <- roc(test.Karenia_br, Prob,
        smoothed = TRUE,
        # arguments for ci
        ci=TRUE, ci.alpha=0.9, stratified=FALSE,
        # arguments for plot
        plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,
        print.auc=TRUE, show.thres=TRUE)
sens.ci <- ci.se(pROC_obj)
plot(sens.ci, type="shape", col="lightblue")
plot(sens.ci, type="bars")
###### ROCit - 2019 Youden index ########
ROCit_obj <- rocit(score=Prob,class=test.Karenia_br)
plot(ROCit_obj)


##### Compare ROCs ########
pred2 <- prediction(y_pred, test.Karenia_br)
pred <- prediction(yhat.rf[,2], test.Karenia_br)
pred3 <- prediction(y_pred2, test.Karenia_br)
pred4 <- prediction(Prob, test.Karenia_br)
perf2 <- performance(pred2, "tpr", "fpr")
perf <- performance( pred, "tpr", "fpr" )
perf3 <- performance(pred3, "tpr", "fpr")
perf4 <- performance(pred4, "tpr", "fpr")
plot( perf, colorize = FALSE, col="blue", lty = 1, lwd = 1.5)
plot(perf2, add = TRUE, colorize = FALSE, col="GREEN", lty = 1, lwd = 1.5)
plot(perf4, add = TRUE, colorize = FALSE, col="RED", lty = 1, lwd = 1.5)
#plot(perf4, add = TRUE, colorize = FALSE, col="PURPLE", lty = 1, lwd = 1.5)
legend(x = "bottomright",
    col = c("GREEN", "blue","red" , "PURPLE"), lty = 1, lwd = 4,
    legend = c('XGboost', 'Random Forest', "GBM", "SVM"), cex=1.1)
legend(x = "bottomright",
    col = c("GREEN", "blue","red"), lty = 1, lwd = 4,
    legend = c('XGboost', 'Random Forest', "SVM"), cex=1.1)
```

```
################################################
############### precrec - 3*3 plots ############
# Random Forest
precrec_obj1 <- evalmod(scores = yhat.rf[,2], labels = test.Karenia_br) #Precision-Recal
precrec_obj2 <- evalmod(scores = yhat.rf[,2], labels = test.Karenia_br, mode="basic") #ALL
autoplot(precrec_obj1)
autoplot(precrec_obj2)


#roc_imp <- filterVarImp(x = training_set[, -ncol(training_set)], y = training_set$Karenia_br)
#head(roc_imp)




################################################
############ Variable Importance #############
################################################
######## VarImpt BOXPLOT (cforest)
set.seed(1)
cf <- cforest(Karenia_br ~., subset = train, data = dataset,
         control = cforest_unbiased(mtry = 6, ntree = 1000))
vi <- t(replicate(50, varimp(cf)))
boxplot(vi)


########## VIP BARplot (ranger)
#set.seed(101)
#rfo <- ranger(Karenia_br ~ ., data = dataset[train,], importance = "impurity")
#(vi_rfo <- rfo$variable.importance)
#barplot(vi_rfo, horiz = TRUE, las = 1)

########## VI plot for XGBOOST Color cluster #########
(vi_bst <- xgb.importance(model = classifier))
xgb.ggplot.importance(vi_bst)
########## VI plot for GMB Color cluster ############
(vi_bst2 <- xgb.importance(model = bst))
xgb.ggplot.importance(vi_bst2)

######## VI for model-specific (not Model-agnostic VI scores)
#vi(rfo)   # rf 1
#vi(cf) # rf 2
vi(rf.dataset) # rf 3
vi(bst)   # GBM
vi(classifier) # XGBOOST
p1 <- vip(rf.dataset, width = 0.5, aesthetics = list(col = "blue1"))   # rf 3
p2 <- vip(bst, aesthetics = list(col = "green2"))    # GBM
p3 <- vip(classifier, aesthetics = list(col = "red2"))   # XGBOOST
grid.arrange(p1, p2, p3, ncol = 3)
```

```
######## SVM Var Importance ######
w = t(svm.fit2$coefs) %*% svm.fit2$SV
ww = as.data.frame(w)
ww=sort(w)
head(svm.fit2$decision.values)
ww=ww[(length(ww)/2):length(ww)]
barplot(ww,
      main = "SVM Variable Importance",
      xlab = "Variable Importance",
      ylab = "Variable",
      names.arg = c("1", "2", "3", "4", "5", "6", "7","8","9"),
      col = "BLACK",
      horiz = TRUE)


# for theme_light() function
# vip(bst, num_features = 5, geom = "point", horizontal = FALSE, aesthetics = list(color = "red",
shape = 17, size = 4)) + theme_light()
```