



Article

Machine Learning Classification Algorithms for Predicting *Karenia brevis* Blooms on the West Florida Shelf

Marvin F. Li ¹, Patricia M. Glibert ^{2,*}  and Vyacheslav Lyubchich ³ 

¹ Harvard College, Harvard University, 86 Brattle Street, Cambridge, MA 02138, USA; marvinli@college.harvard.edu

² Horn Point Laboratory, University of Maryland Center for Environmental Science, P.O. Box 775, Cambridge, MD 21613, USA

³ Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, P.O. Box 38, Solomons, MD 20688, USA; lyubchich@umces.edu

* Correspondence: glibert@umces.edu; Tel.: +1-410-221-8422

Abstract: Harmful algal blooms (HABs), events that kill fish, impact human health in multiple ways, and contaminate water supplies, have increased in frequency, magnitude, and impacts in numerous marine and freshwaters around the world. Blooms of the toxic dinoflagellate *Karenia brevis* have resulted in thousands of tons of dead fish, deaths to many other marine organisms, numerous respiratory-related hospitalizations, and tens to hundreds of millions of dollars in economic damage along the West Florida coast in recent years. Four types of machine learning algorithms, Support Vector Machine (SVM), Relevance Vector Machine (RVM), Naïve Bayes classifier (NB), and Artificial Neural Network (ANN), were developed and compared in their ability to predict these blooms. Comparing the 21 year monitoring dataset of *K. brevis* abundance, RVM and NB were found to have better skills in bloom prediction than the other two approaches. The importance of upwelling-favorable northerly winds in increasing *K. brevis* probability, and of onshore westerly winds in preventing blooms from dispersing offshore, were quantified using RVM, and all models were used to explore the importance of large river flows and the nutrients they supply in regulating blooms. These models provide new tools for management of these devastating algal blooms.

Keywords: harmful algal bloom; *Karenia brevis*; machine learning; Support Vector Machine; Relevance Vector Machine; Naïve Bayes classifier; Artificial Neural Network



Citation: Li, M.F.; Glibert, P.M.; Lyubchich, V. Machine Learning Classification Algorithms for Predicting *Karenia brevis* Blooms on the West Florida Shelf. *J. Mar. Sci. Eng.* **2021**, *9*, 999. <https://doi.org/10.3390/jmse9090999>

Academic Editor: Carmela Caroppo

Received: 14 July 2021

Accepted: 9 September 2021

Published: 13 September 2021

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



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

1. Introduction

The incidence of harmful algal blooms (HABs) has increased globally. HABs are now occurring more frequently, in new and different places, and often last longer, having a wide range of environmental and toxic impacts and in numerous fresh, estuarine, and marine waters [1–3]. Both nutrient pollution and climate change are now recognized to play important roles in this expansion [4–8]. Nutrient runoff is increasing with increases in human population and associated changes in diets and the food supply chain, and rising temperatures and climate changes are leading to changes in precipitation patterns and increased intensity or frequency of storm events, that, in turn, alter coastal runoff and physical processes, such as upwelling and stratification [8]. From local to global scales, environmental conditions supporting HABs are changing, leading to increasing challenges for understanding—and modeling—the habitats that support and stimulate them.

Blooms of the toxic dinoflagellate *Karenia brevis* occur almost annually on the West Florida Shelf (WFS), and historical accounts show that they have occurred since at least the 16th century [9]. However, recent analyses suggest that bloom events have increased 15-fold from the 1950s to the 1990s, although quantifying patterns and trends is complicated by the inconsistency of data collection over this period [10]. During 2017–2019, southwest Florida experienced an unusually prolonged (18 months) *K. brevis* bloom. At its maximum, this

bloom covered a region with a coastline of more than 250 km, encompassing recreational beaches and numerous commercial and recreational shellfish beds, causing both ecological and economic harm (Figure 1) [11–13].

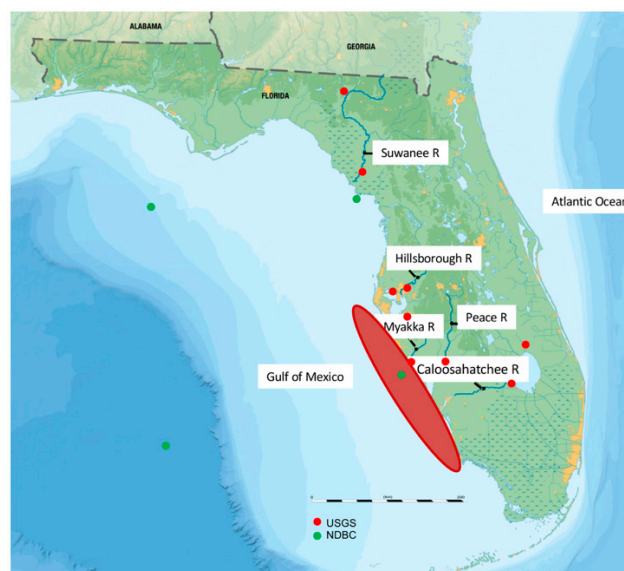


Figure 1. Map of the West Florida Shelf (WFS) showing the region in red where *Karenia brevis* blooms are most frequently observed, and the major rivers that discharge onto the WFS and for which nutrient (total nitrogen and total phosphorus) data are available. Red dots mark the United States Geological Survey (USGS) stations and green dots mark the National Data Buoy Center stations (Station 42,039, 28.787N 86.007W; Cedar Key, 29.12N 83.02 W and Fort Myers 26.65N, 81.88W) from which wind and temperature data were acquired.

Although *K. brevis* is typically thought of as a coastal bloom species, blooms are generally initiated offshore and then transported to coastal waters where they flourish and persist for months in nutrient-rich waters [9]. Blooms of *K. brevis* usually begin in the late summer or early fall, and can persist until the late fall or winter [14]. Upwelling transports *K. brevis* cells to the coast [15–17], but strong upwelling over the shelf break may actually suppress *K. brevis* blooms or favor competing taxa such as diatoms [18,19]. It is thought that northerly wind generates the coastal upwelling that transports *K. brevis* from offshore regions to coastal waters, producing favorable conditions for growth. Once *K. brevis* reaches coastal waters, the westerly wind keeps populations near the coast and prevents them from dispersing offshore. In the current study, the effects of winds were further explored herein.

The nutrient sources, pathways, and processes supporting and maintaining *K. brevis* blooms not only include upwelling, but also riverine nutrient inputs that bring wastewater effluent and agricultural runoff [14]. Other nutrient sources include benthic nutrient fluxes, atmospheric deposition, nutrients released by other phytoplankton and decaying fish from fish deaths, submarine groundwater discharge, and mixotrophic grazing—suggesting complex environmental interactions of nutrients with bloom occurrence and strength [14,20–25]. Nutrient relationships and riverine flow were also tested herein.

The massive bloom of 2017–2019, and another recent, large-scale bloom, observed in 2005, were likely propelled by unusual events. Hu et al. [20] suggested that nutrient inputs resulting from a series of hurricanes in southwest Florida in 2004 were linked with the severity of the 2005 bloom. Hurricanes can accelerate the yield of new sources of land-based nutrients from high riverine flow. Similarly, Hurricanes Irma (2017) and Michael (2018), and Tropical Storm Gordon (2018), are suspected of contributing to the severity of the 2017–2019 *K. brevis* bloom [11]. Moreover, long periods of wet weather through 2018, combined with increased discharges from Lake Okeechobee and the Caloosahatchee River,

added nutrients to coastal waters, sustaining large *K. brevis* blooms through early 2019. In late 2020, another hurricane, Eta, appears to have played a similar role in helping to sustain a very recent large bloom that has lasted through the winter and spring of 2020–2021.

There is a strong need to advance predictions of *K. brevis*, and other HABs more generally, to protect human health, fisheries, and economies, but there are many challenges in modeling discrete HAB species [26–31]. Several types of models have been developed and are currently in operation for predicting *K. brevis* blooms [9,32–34]. Weisberg et al. [34] developed a high-resolution coastal ocean circulation model to track the movement of water particles associated with *K. brevis* populations. An operational forecasting modeling system, maintained by the National Oceanic and Atmospheric Administration, provides 3–5 day outlooks of *K. brevis* blooms, using satellite remote sensing of chlorophyll *a*, in situ sampling of *K. brevis* cell density, and wind buoy data [32,33]. The main goal of these forecasts is to inform managers and the public in coastal areas where public health may be compromised [33]. Walsh et al. [35], in addition to Weisberg and He [15], used a three-dimensional (3D) biophysical-coupled model to hindcast bloom initiation and explore the impact of individual forcing functions [35–38]. However, these models utilize many biochemical and physiological parameters, some of which have not been well characterized either in situ or in the laboratory. Furthermore, these 3D-coupled biophysical models are computationally expensive.

Due to their powerful nonlinear modeling capability, machine learning methods are beginning to be used to predict HAB events, including *K. brevis* events [39]. An Artificial Neural Network (ANN) model was used to predict algal blooms in Hong Kong coastal waters [40] and to predict outbreaks of the dinoflagellate *Dinophysis acuminata* in southern Spain [41]. More recently, a Neural Network (NN) approach was used to predict the presence/absence and abundance of the dinoflagellate *Karlodinium* and the diatom *Pseudo-nitzschia* in Alfacs Bay in the northwest Mediterranean Sea [42], and Support Vector Machine (SVM) models were used to predict blooms in freshwater reservoirs [43]. Shen et al. [44] used SVM models to simulate algal blooms in the tidal freshwater of James River in response to riverine nutrient loading.

Machine learning approaches have also been used in predicting HABs in the Gulf of Mexico, but with different objectives. Liu and Weisberg [16] used such approaches to demonstrate the role of deep-ocean forcing on WFS in the major bloom occurrences. Weisberg et al. [18] reported that the position of the Loop Current can affect blooms. When the Loop Current is in its southern position, it creates an upwelling of deep nutrients and fosters a diatom bloom that may outcompete any nascent *K. brevis* blooms. Liu et al. [19] used Self-Organizing Maps to classify spatial patterns of the sea surface height anomalies associated with the Loop Current and found no bloom developed when the Loop Current was in the southern position. That work focused exclusively on the potential effects of the Loop Current on *K. brevis* blooms and did not consider other factors such as river flows and riverine nutrient loading. Gokaraju et al. [45] proposed a machine learning based spatiotemporal data mining approach to detect HABs from SeaWiFS (Sea-viewing Wide Field-of-view) and MODIS (Moderate Resolution Imaging Spectroradiometer)-Aqua space-borne sensor measurements. Recently, Hill et al. [46] used satellite remote sensing of chlorophyll *a* from 2003 to 2018, sea surface temperature, and bathymetry as inputs to a convolutional NN (designed for spatial data) to detect the presence of *K. brevis* blooms on WFS, achieving a maximum detection accuracy of 91%. Such approaches have yet to be used to assess the effects of winds, river flows, and river nutrient discharge on the likelihood of *K. brevis* blooms.

The State of Florida, along with numerous other states and government entities around the world, has established, or is working to establish, nutrient reduction targets to mitigate water quality problems in their water bodies. Continued monitoring and assessment methods will be essential, and improved approaches for establishing criteria for additional waters and to manage water quality across greater regions will continue to be required.

In this research, four machine learning algorithms were used to predict *K. brevis* on the WFS over a 21 year period. Specifically, the performance of these different machine learning approaches was assessed with regard to forecasting the probability of *K. brevis* blooms with changing wind, discharge from different rivers, differing nutrient loads, and sea surface height (as a proxy for temperature and upwelling strength). New modeling approaches will provide new tools for defining scientifically defensible protective nutrient loads in future re-evaluations of Florida's water quality criteria and in predicting blooms to protect human health and commercial interests.

2. Materials and Methods

2.1. The Dataset and Preparation of Explanatory and Dependent Variables

To develop the machine learning models, in situ data of *K. brevis* cell densities (cells L⁻¹) over a 21 year period (1998–2018) on the WFS were obtained from the database available from the Florida Fish and Wildlife Conservation Commission (FFWRCC) [47]. All data represent near-surface cell counts taken using light microscopy; samples are collected by various state and county agencies, private research institutions, and university researchers and routinely reported to FFWRCC. These data represent water samples collected during regular monitoring along the Florida coast and during suspected or confirmed *K. brevis* events. Although data are available for earlier decades, since 1998 more consistent inshore and offshore stations have been sampled, and thus this analysis is limited to data post-1998.

The data used herein were limited to water samples collected between latitudes of 25.85 degrees north (Marco Island) and 29.14 degrees north (Mouth of Suwanee River) and at most 9 km from the coast because most of the *K. brevis* blooms occurred within this area. Moreover, setting a fixed area for the data analysis ensured data consistency [48]. Because the *K. brevis* measurements were largely collected if and when blooms were documented, and not made on a continuous or regular basis, the database has an undersampling of *K. brevis* under low cell density conditions. To overcome the spatial and temporal inconsistency in the data, the 5 highest cell counts across the fixed area were averaged for each week to produce a weekly mean, following the approach used in previous studies [19]. These cell numbers were discretized and weekly averages were combined into a binary variable, with mean cell densities greater than 10⁵ cells L⁻¹ counting as *K. brevis* events, consistent with the commonly used threshold for *K. brevis* blooms.

Streamflow data were obtained from United States Geological Survey (USGS) stations in the major rivers that discharge onto the WFS [49] (Figure 1). The USGS stations used included: Tampa Bay (USGS 2306647), Peace River (USGS 2296750), Lake Okeechobee (USGS 2274325), Suwanee River (USGS 2323500), Withlacoochee River (USGS 2319000), Hillsborough River (USGS 2303330), Little Manatee River (USGS 2300500), Myakka River (USGS 2298830), and Caloosahatchee Canal (USGS 2292000). Nutrient data from the major rivers, including the total nitrogen (TN) and total phosphorus (TP) concentrations, were downloaded from the Tampa Bay and Charlotte Harbor Water Atlas (University of South Florida Water Institute) [50]. No nutrient concentration data were available for the Suwanee River. Weekly averaged nutrient concentrations were multiplied by weekly averaged streamflow to estimate weekly TN and TP loads.

Hourly wind and temperature data were obtained from the National Data Buoy Center (NDBC) stations [51] (Figure 1) across the WFS. The hourly wind speeds were used to calculate weekly averages using a simple vector average. The hourly temperature was used to calculate weekly averages. Satellite altimetry from the GLOBAL-REANALYSIS-PHY-001-030 reanalysis product, provided by the E.U. Copernicus Marine Service Monitoring Service (CMEMS) [52] was used to calculate the difference in sea surface height at two locations to quantify the strength of the deep-sea coastal upwelling caused by the Loop Current, following Maze et al. [48].

Data were aggregated into the following form; each row $i = 1, \dots, 1083$ of the dataset is $\{x_1^i, x_2^i, x_3^i, \dots, x_{33}^i, y_i\}$, where $x_1^i, x_2^i, x_3^i, \dots, x_{33}^i$ are the explanatory variables of river

discharge, nutrient concentration, wind speed and direction, temperature, and sea surface height difference, and y_i is the dependent variable of discretized *K. brevis* cell densities. Machine learning algorithms aim to map $x_1^i, x_2^i, x_3^i, \dots, x_{33}^i$ to y_i . The number of HAB events (318, based on the criteria of cell density $>10^5$ cells L^{-1}) was less than half of the number of events without HABs (765, cell density $<10^5$ cells L^{-1}), resulting in an imbalanced classification problem [53]. Machine learning algorithms for classification predictive models are designed assuming an equal distribution of classes. Because there are fewer examples of the minority class (non-HAB events) than the majority class (HAB events), it becomes harder to predict the non-event periods.

Several approaches have been developed to address this issue and two different approaches were applied herein: (1) the minority class of the training data was randomly oversampled such that the sample size of events with and without HABs were roughly equal in the synthetic training dataset [54]; and (2) the minority class was oversampled by generating new synthetic data using a synthetic minority oversampling technique (SMOTE) preprocessing algorithm [55–57].

2.2. Machine Learning Algorithms

To predict *K. brevis* cell density and test the strength of various explanatory variables, the following machine learning algorithms were used: a) Support Vector Machine (SVM), b) Relevance Vector Machine (RVM, a modification of SVM), c) Naïve Bayes (NB), and d) Artificial NN (ANN). These approaches represent a range of machine learning algorithms with different methodologies and varying complexity. SVM belongs to a class of algorithms called kernel methods. RVM has an identical functional form to SVM but provides probabilistic classification. NB is a family of simple probabilistic classifiers. ANN is based on a system of connected nodes to mirror neurons in a biological brain.

2.2.1. Support Vector Machine

The SVM model is a supervised machine learning algorithm that seeks the hyperplane that best separates two labeled classes from each other; the optimal hyperplane maximizes the marginal distance from the nearest support vector for each class [45] (Figure 2a):

$$f(x) = \text{sign}(\langle w, x \rangle + b) \tag{1}$$

where x represents the vector for the explanatory variables, $\langle w, x \rangle + b = 0$ is the hyperplane that separates the two classes, w is the slope of the hyperplane, b is the intercept, and $f(x)$ is the classifier output which takes the value of either 1 or -1 . SVM seeks a solution for the hyperplane by maximizing the width of the gap between the two data clouds, represented by the cost function (CF, Equation (2), Figure 2a):

$$CF = \|w\|^2. \tag{2}$$

Sometimes the SVM cannot achieve a perfect separation. The soft-margin loss formulation allows some data points to lie within the margin of tolerance but penalizes them in the cost function [58] as follows (Equation (3)):

$$\text{Minimize } CF = C \|w\|^2 + \frac{1}{N} \sum_{i=1}^N \xi_i \tag{3}$$

where in the slack variables, $\xi_i = \max(0, 1 - y_i \cdot (\langle w, x_i \rangle + b))$, y_i is either 1 or -1 , and C is a hyperparameter which determines the trade-off between maximizing the margin width and minimizing the associated error (Figure 2a). This new cost function is then optimized, yielding the linear support vector expansion for the classifier (Equation (4)):

$$f(x) = \text{sign}(b + \sum_{i=1}^N \alpha_i y_i (x_i - x)) \tag{4}$$

where α_i are the Lagrangian multipliers and w is rewritten as a linear combination of the training patterns [59]. The constant b can be found with the Karush–Kuhn–Tucker Conditions [60,61]:

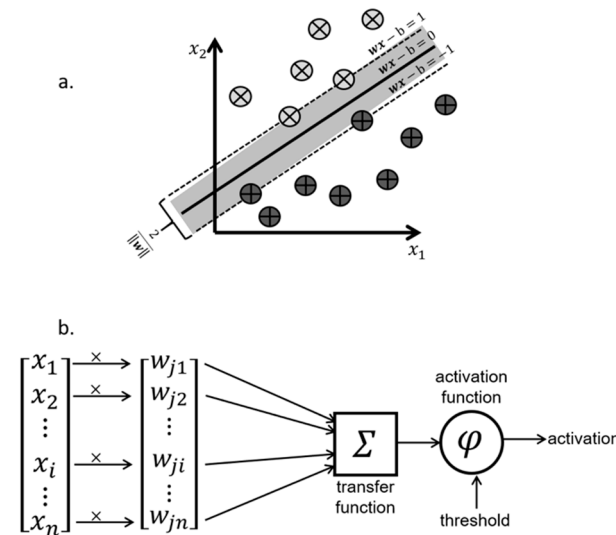


Figure 2. (a) A schematic diagram of the Support Vector Machine (SVM) classifier. The SVM model is a supervised machine learning algorithm that seeks a hyperplane that best separates two labeled classes from each other. The SVM maximizes the width of the gap between the two data clouds. In some cases, not all of the data points can be fitted into the two data clouds outside the shaded gap region. In the soft margin formulation of the SVM, points are allowed inside the gap but penalized in the cost function. (b) A schematic diagram of the Artificial Neural Network (ANN) model. The ANN is based on the feedforward multilayer perceptron architecture, consisting of an input layer, one or more sets of hidden layers, and one output layer. The ANN can be turned into a classifier by discretizing the network’s output. The basic substructure of the ANN is perceptron. For all but the input layer, the perceptron has an input (the outputs of the previous layer). The vectors of inputs and the neuron’s weights are multiplied by a dot product. Then, a transfer function is applied to the sum, giving an output for the next layer of perceptrons.

The linear support vector expansion cannot be used to describe nonlinear relationships between the explanatory and dependent variables. To describe nonlinear datasets, kernel functions are used to map the data to higher dimensions where they exhibit linear patterns and the linear model can be applied in that feature space [62,63]. The Gaussian radial basis function was chosen as the kernel function because of its computational efficiency (Equation (5)),

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2) \tag{5}$$

where γ is the kernel parameter controlling the sensitivity of the kernel function.

SVM has two hyperparameters that cannot be determined from optimization: C and γ . Both were determined with a grid search method on the training data.

2.2.2. Relevance Vector Machine

The RVM model has an identical functional form to SVM but uses a Bayesian probabilistic framework to estimate the parameters [64,65]. To obtain the maximum likelihood estimate of w , and to avoid overfitting, the Bayesian approach is taken to constrain the parameters by defining an explicit prior probability distribution over them. The prior probability distribution is chosen to be a Gaussian distribution, and RVM introduces a vector to enforce a preference for smoothness. Then, the posterior estimate of the unknown parameters given the data is obtained using Bayes’ rule. Because the posterior probability can be evaluated exactly, RVM seeks to maximize the marginal likelihood with respect to

the hyperparameters. RVM typically uses considerably fewer basis functions than SVM. RVM was applied herein using the radial basis function as the kernel function.

2.2.3. Naïve Bayes

The NB classifier is a simple probabilistic classifier based on the Bayes' rule and requires strong "naïve" independence between the features [66,67]. Given a new observation \mathbf{x} , it finds the class C_k that maximizes the conditional probability $p(C_k|\mathbf{x})$, the likelihood of a class given the observation. Using Bayes' theorem, the conditional probability can be calculated as follows (Equation (6)):

$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})} \tag{6}$$

where $p(C_k)$ is the prior probability of observing a class C_k , $p(\mathbf{x}|C_k)$ is the likelihood of observing \mathbf{x} given C_k , and $p(\mathbf{x})$ is the probability of observing \mathbf{x} . Assuming strong naïve independence, the probabilistic chain rule can be used to transform the likelihood $p(\mathbf{x}|C_k)$ of \mathbf{x} into the probabilities of each of the features of \mathbf{x} given a class (Equation (7)):

$$p(\mathbf{x}|C_k) = \prod_{i=1}^N p(x_i|C_k) \tag{7}$$

This study used the Gaussian NB in which the Gaussian distribution (Equation (8)),

$$p(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}} \tag{8}$$

is assumed to underlie the sample distribution. To train the NB classifier, the data were segmented by the classes, and the mean and standard deviation of each of the features for each of the classes were calculated, giving a probability distribution for each of the classes.

2.2.4. Artificial Neural Network

ANN is based on the feedforward multilayer perceptron architecture, consisting of an input layer, one or more sets of hidden layers, and one output layer [68,69]. ANN can be turned into a classifier by discretizing the network's output. The basic substructure of ANN is a perceptron (Figure 2b). Each perceptron has an input (the outputs of the previous layer), a series of weights, a transfer function, and an output. A transfer function is applied to the dot product of the inputs and weights for each perceptron, giving an output for the next layer. The output $y_j^{(l)}$ for node j in layer l is as follows (Equation (9)):

$$y_j^{(l)}(\mathbf{x}) = \varphi\left(\sum_{i=1}^n w_{ji}^{(l)} y_i^{(l-1)}(\mathbf{x})\right) \tag{9}$$

where \mathbf{x} are the input variables, $y_j^{(l-1)}$ is the output at layer $(l-1)$, w_{ji} are the synaptic weights, and φ is the activation function.

Initially, random numbers are assigned to the synaptic weights. The weights are adjusted with the training data. There are two main steps to the training of the ANN: forward computation and back propagation. In forward propagation, input signals are propagated through the network, layer by layer. In back propagation, the error for the entire network is calculated [70]. Then, the errors are computed for each neuron, and then the local gradients for the synaptic weights of the network are calculated. Gradient descent is then used to adjust the synaptic weights. These steps are repeated until the error falls below a desired threshold. Herein, two hidden layers with 20 and 10 neurons were used in the ANN model, because rules of thumb suggest that two hidden layers are sufficient

given the number of explanatory variables in this classification problem and that each layer should have approximately half the number of nodes as the preceding layer [71]. Hyperparameter tuning was undertaken by a nested k -fold cross-validation procedure, which is described in detail in the following section.

To implement these machine learning algorithms, open-source R packages were used: raster 3.0–7 [72], doParallel 1.0.15 [73], Kernlab 0.9–29 [74], DMwR 0.4.1 [75], PBSmapping 2.72.1 [76], e1071 1.7–2 [77], neuralnet 1.44.2 [78], ggplot2 3.3.5 [79] in R 3.6.1 [80].

2.3. Model Evaluation and Metrics

The predictive skill of the machine learning algorithms was evaluated using two approaches. First, a k -fold cross-validation approach that has been widely used in machine learning classification [81,82] was applied. In this approach, the data are randomly divided into k disjointed subsets of equal size. Then, for each combination of $k-1$ of the k subsets, one of the k models is trained, and the test statistic for that model is evaluated on the remaining subset [83,84]. The mean of the test statistics over all k models is called the cross-validation estimate. In this study, $k = 10$ so that each subset spanned 2 years of data. The data are assumed to be independent during the k -fold cross-validation. However, this assumption may be inappropriate for time series that may be auto-correlated. Thus, the data herein were further validated by block cross-validation [85–88]. To do this, the data were divided by chronological order into 10 subsets of 2 years each: 1998–1999, 2000–2001, . . . , 2017–2018. In one iteration of the cross-validation procedure, the models were trained on the data from 1998–2016 and then tested on data from 2017–2018. This procedure was repeated for all the 2 year blocks.

Four metrics were used to evaluate the performance of the machine learning classifiers in predicting *K. brevis* blooms [53,89]. Accuracy measures the overall accuracy of the prediction (Equation (10)):

$$A = \frac{TrPos + TrNeg}{TrPos + FNeg + TrNeg + FPos} \quad (10)$$

where TrPos is the number of weeks with blooms predicted correctly (true positives), FNeg is the number of weeks with blooms predicted to be non-HAB weeks (false negatives), TrNeg is the number of non-HAB weeks predicted correctly (true negatives), and FPos is the number of non-HAB weeks predicted to be weeks with HABs (false positives). A is the measure of all the correctly identified cases. Recall (R) is the ratio of the correctly-predicted HAB weeks to the total number of the observed HAB weeks (Equation (11)):

$$R = \frac{TrPos}{TrPos + FNeg} \quad (11)$$

Precision (P) is the ratio of the correctly-predicted HAB weeks to the total number of the predicted HAB weeks (Equation (12)):

$$P = \frac{TrPos}{TrPos + FPos} \quad (12)$$

F1 measures the balance between precision and recall (Equation 13):

$$F1 = 2 \times \frac{R * P}{R + P} \quad (13)$$

Although A is most often used when there are similar amounts of each class, $F1$ score is a better metric where there are imbalanced classes. The testing metrics were averaged for both the k -fold or block validation procedures. To further test the models' predictions, a time series of the cross-validation predictions was created.

In addition to the four metrics, the robustness of each model, such as sparsity, was examined. The mean number of support vectors (SVs)/relevance vectors (RVs) was calculated for the SVM and RVM, and the Akaike Information Criterion (AIC) was determined for the ANN.

2.4. Sensitivity Analysis

In order to determine how environmental factors affect the probability of *K. brevis*, Platt scaling [90] was applied. This uses a logistic transformation to convert classifier predictions into probability distributions over the classes. Platt scaling is ideal for this study because of its simplicity and the size of the training dataset. First, each machine learning algorithm was trained on the entire dataset. Platt scaling was then applied to calculate the probability of *K. brevis* blooms (Equation (14)):

$$P(y_i = C_{+1}|x) = \frac{1}{1 + \exp(Af(x) + B)} \quad (14)$$

where y_i is a sample, C_{+1} is one of the classes, $f(x)$ is the classifier output, and A and B are scalar constants [91].

2.5. In Silico Experiments

Using the various models, the impacts of different environmental variables on probabilities of *K. brevis* occurrence were assessed. To do so, each explanatory variable or variables were varied by 1–2 standard deviations around its mean whereas the other variables were set to their respective annual mean values. Line plots and contour diagrams of HAB probability as a function of explanatory variables were created by varying one or two explanatory variables at a time.

The probability of *K. brevis* blooms as a function of wind speed components in the north–south direction (negative for northerly wind) and the east–west direction (negative for easterly wind) were examined using SVM. To do so, the wind components were varied 1–2 standard deviations above and below the long-term mean while holding other factors constant.

The probabilities of *K. brevis* outbreaks as a function of discharge from the Suwanee, Hillsborough, Myakka, Peace, and Caloosahatchee Rivers—all of which discharge into the WFS—were examined using SVM, RVM, and NB. Using the same three models, the probabilities of blooms for each river as a function of their TN or TP loads were also estimated based on variations of 1–2 standard deviations from the mean (and setting other features to the mean).

3. Results

3.1. Overall Model Performance

The predictability of *K. brevis* blooms over the 21 year time series (1998–2018) was tested relative to the observed *K. brevis* cell concentrations along the WFS using all four machine learning approaches. All models captured the general time series of *K. brevis* events, encompassing both prolonged blooms with high cell counts, and periods of only a short duration with relatively low cell counts, but the RVM model was the most robust (Figures 3 and 4).

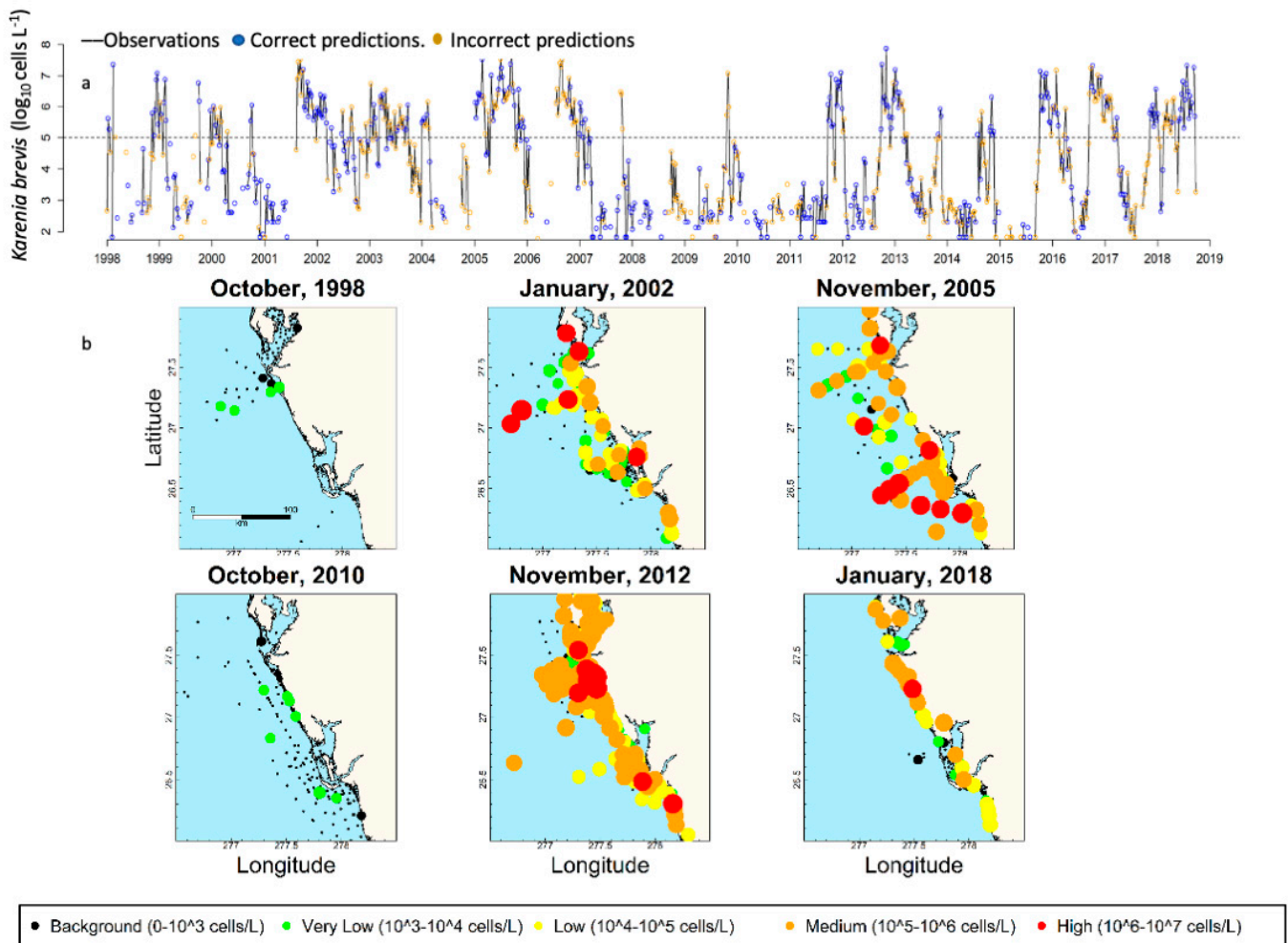


Figure 3. Comparison of Relevance Vector Machine (RVM) output and observational data of *Karenia brevis*. (a) Time series of the observed (black line) and predicted (blue and orange dots) area-averaged *K. brevis* concentrations from 1998 to 2018. Cell counts above 5 (log₁₀ cells L⁻¹; dashed line) are herein considered bloom conditions. (b) Snapshots of the observed *K. brevis* distribution in selected months. The twenty-one year timespan includes many years with blooms (2002, 2005, 2012, 2018) and without blooms (1998, 2010).

Results from the random oversampling and SMOTE sampling method were similar for all four approaches (Table 1). According to the block cross-validation, SVM and ANN achieved significantly higher prediction accuracy (0.62 and 0.61, respectively, from random oversampling) than RVM and NB (0.55 and 0.47, respectively). In contrast, when comparing the recall values using the same block cross-validation with random oversampling, NB had the highest recall (0.72), followed by RVM (0.58), implying that these models correctly predicted 72% and 58% of the prior *K. brevis* blooms, respectively, whereas SVM and ANN had much lower recall values (0.27 and 0.33 respectively). All models predicted a similar number of false positives, as shown by their precision values ranging between 0.32 and 0.35. The F1 score, the balance between recall and precision, was highest for RVM and NB, at 0.43 and 0.45 respectively.

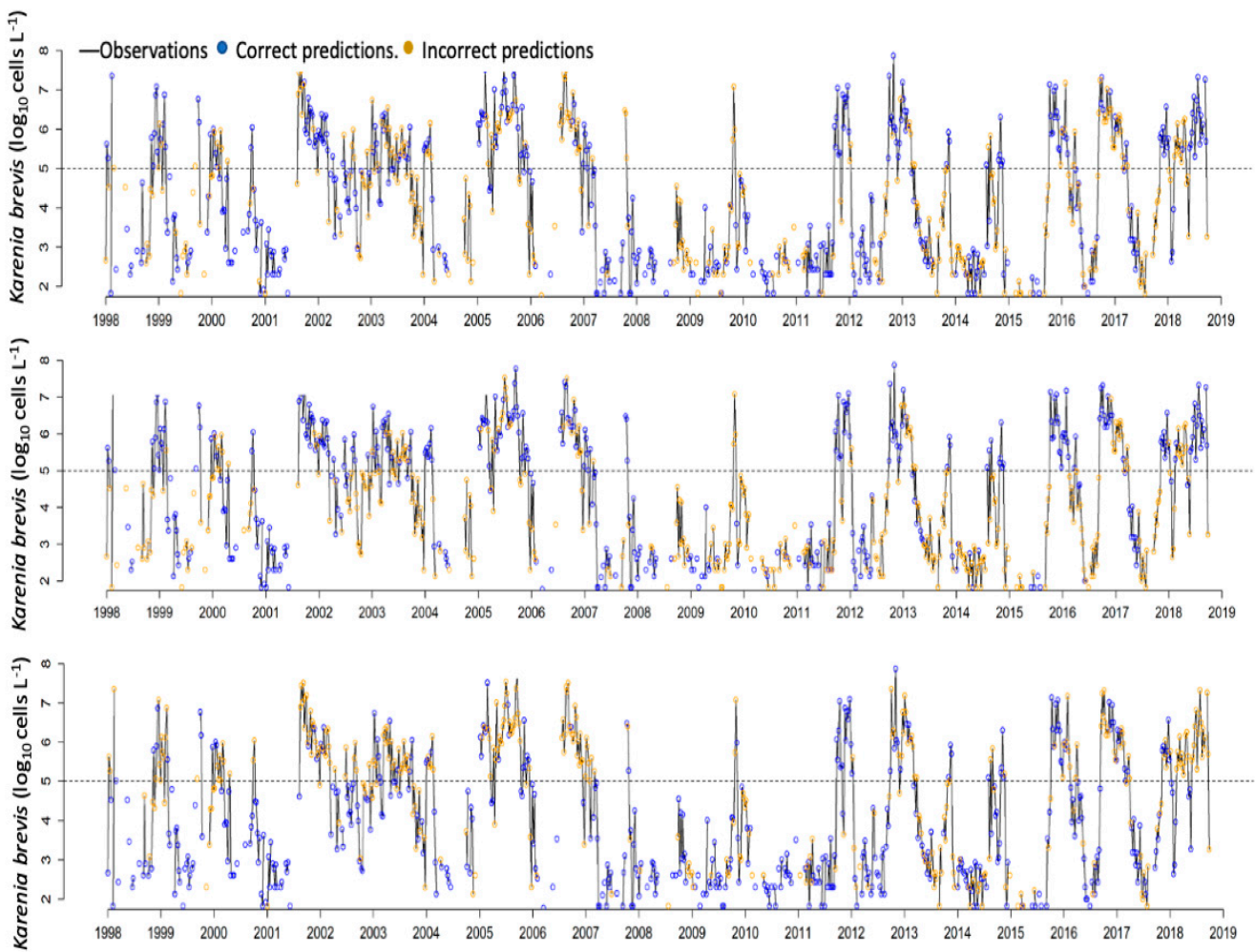


Figure 4. Comparative model outputs of the observed (black line) and predicted (blue and orange dots) area-averaged *K. brevis* concentrations from 1998 to 2018 for the Relevance Vector Machine (top panel), Naïve Bayes (middle panel), and Artificial Neural Network (bottom panel). Cell counts above 5 (\log_{10} cells L^{-1} ; dashed line) are herein considered bloom conditions.

3.2. Role of Wind

Winds differed in years with and without a bloom. During years with blooms, stronger northerly and westerly winds occurred (Figure 5). Using RVM, the probability of *K. brevis* blooms as a function of wind speed components was examined in the north–south direction (negative for northerly wind) and the east–west direction (negative for easterly wind). Bloom probability was much higher under northerly winds than under southerly winds (Figure 6). Bloom probability reached a maximum of 0.57 under northerly wind, whereas strong southerly wind reduced bloom probability to <0.3 . Northerly winds drive coastal upwelling, thereby transporting *K. brevis* from the offshore waters to coastal waters. Westerly winds corresponded to higher probability (up to 0.53) of a bloom, compared with easterly winds, with bloom probability as low as 0.36 for the strongest easterly winds (Figure 6). Once *K. brevis* reaches nearshore locations, westerly winds help hold *K. brevis* blooms against the shore where they can access nutrient sources from land and rivers.

Table 1. Comparison of the four machine learning approaches applied herein (Support Vector Machine, SVM; Relevance Vector Machine, RVM; Naïve Bayes, NB; and Artificial Neural Network, ANN), as validated using *k*-fold cross-validation and block cross-validation. See text for equations applied. Best value for each metric in each column is in bold.

Model	Performance Metric	<i>k</i> -Fold Cross-Validation (Random Oversampling)	<i>k</i> -Fold Cross-Validation (SMOTE)	Block Cross-Validation (Random Oversampling)	Block Cross-Validation (SMOTE)
SVM	Accuracy	0.79	0.79	0.62	0.62
	Recall	0.63	0.63	0.27	0.26
	Precision	0.64	0.65	0.32	0.32
	F1	0.64	0.64	0.29	0.29
RVM	Accuracy	0.62	0.76	0.55	0.59
	Recall	0.73	0.72	0.58	0.47
	Precision	0.42	0.58	0.35	0.35
	F1	0.53	0.64	0.43	0.40
NB	Accuracy	0.52	0.54	0.47	0.47
	Recall	0.85	0.78	0.72	0.73
	Precision	0.37	0.37	0.33	0.32
	F1	0.52	0.50	0.45	0.45
ANN	Accuracy	0.74	0.71	0.61	0.60
	Recall	0.57	0.56	0.33	0.40
	Precision	0.55	0.51	0.34	0.34
	F1	0.56	0.53	0.34	0.37

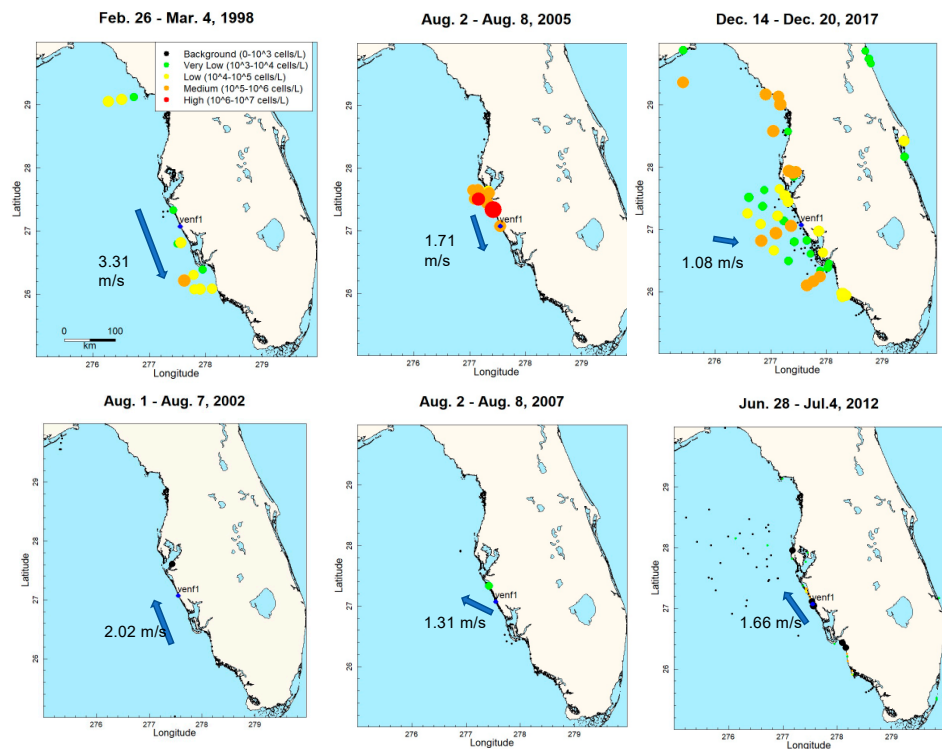


Figure 5. Examples of winds associated with blooms (top panels) and winds for periods without blooms (lower panels). Arrows denote wind direction.

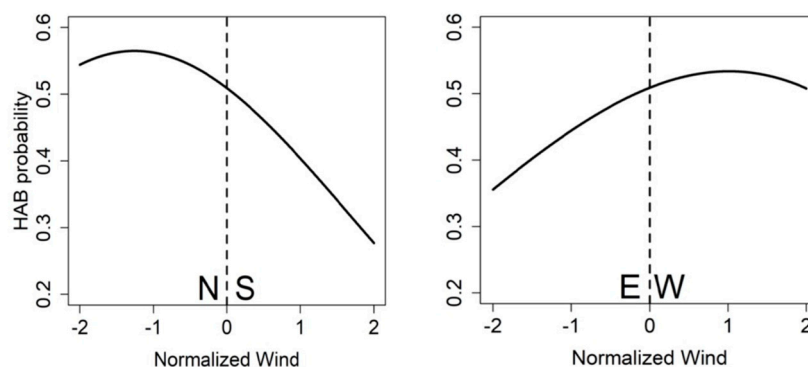


Figure 6. Probability of *Karenia brevis* as a function of wind speed and direction, obtained from the sensitivity analysis using the RVM.

3.3. Role of River Flow and Associated Nutrients

Seasonally, riverine flow typically increases for all rivers in all years during the summer months, but interannual variability is high (illustrated for the 21 year record for the Suwanee, Peace, Myakka, and Hillsborough Rivers, Figure 7a). For all rivers, and across all discharge levels examined, the probability of *K. brevis* blooms increased with river flow (Figure 7b).

For all rivers, bloom probability predicted by the NB was always greater than that predicted by the RVM, which was always larger than that predicted by the SVM (Figure 7b). This result can be explained by Table 1: the NB had the highest recall value; the RVM ranked second; the SVM had the lowest value. The ANN was excluded from the sensitivity analysis because the ANN yields probability predictions that are either very close to 0 (<0.01) or 1 (>0.99). This presents numerical instability issues when computing the weights of the logistic function (Equation (14)) in Platt's scaling, yielding ANN probability curves with discontinuous jumps that are either very close to 0 (<0.01) or 1 (>0.99), like the initial ANN predictions.

As discharge changed, using RVM as the example, the slope in bloom probability was highest with the Hillsborough River, with low discharge yielding a 0.20 probability in blooms, increasing to 0.55 with high discharge (Figure 7b). Increases in discharge from the Peace and Suwanee Rivers also increased bloom probability substantially, from 0.33–0.52 and 0.23–0.54, respectively (with RVM as the example), across the range of typical flows. Changes in discharge from the Myakka River yielded probabilities that changed from 0.34 to 0.55. Across all discharge levels, bloom probability was consistently higher (0.49–0.51 with RVM) with increased Caloosahatchee River discharge than for the other rivers examined, and it increased linearly as river discharge increased. The Caloosahatchee River has the highest discharge of the rivers examined, and it transports the highest amount of nutrients.

The composition of the nutrients discharged by the different rivers also varied and accordingly the probability of blooms varied for their different nutrient loads (Figure 7c,d). Applying the three models SVM, RVM, and NB, with increasing TN, the largest increase in bloom probability was found for the Myakka River, whereas smaller increases were found for the Peace and Caloosahatchee Rivers (Figure 7c). For the Hillsborough River, *K. brevis* probability as a function of the TN loads resembles a parabolic function. For TP, increases in probability were seen for the Peace and Caloosahatchee Rivers, but a parabolic relationship was noted for the Hillsborough and Myakka Rivers (Figure 7d). As nutrient loads increase, it is possible that *K. brevis* may be either outcompeted by a different species or and/or become limited by a different growth factor. However, there were some differences among the three models. For example, the bloom probability versus the Peace River TP load had a slope that was less steep in the NB model than in the RVM and SVM. Although the probability increased slightly with TN in the Caloosahatchee River in the NB and SVM, it decreased in the SVM.

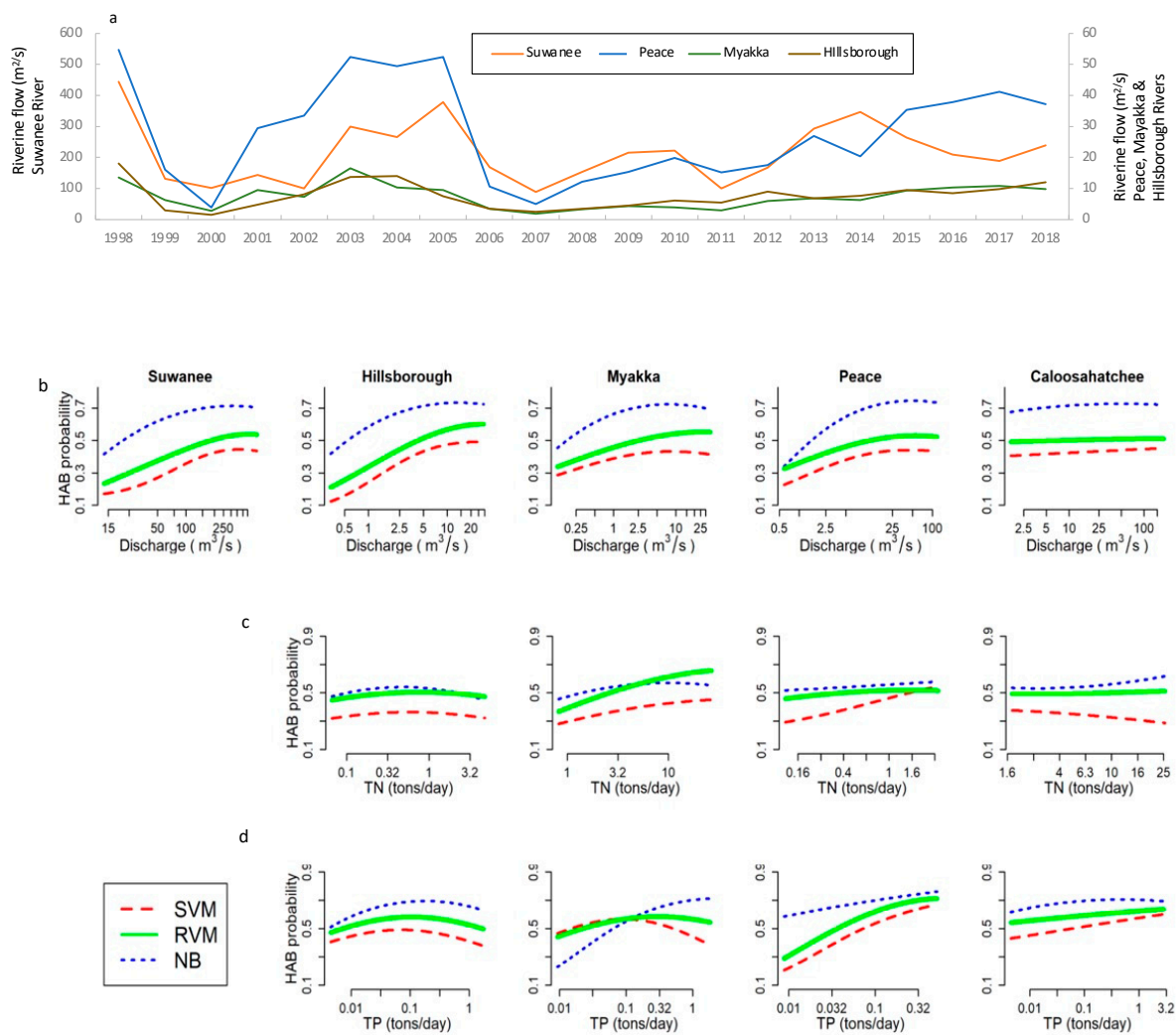


Figure 7. (a) Annual mean riverine flow for rivers indicated for the 21 year time series. (b) Probability of *K. brevis* blooms as a function of riverine discharge. (c,d) Probability of *K. brevis* blooms as a function of total nitrogen loading (TN) and total phosphorous loading (TP). Probabilities were obtained from three machine learning models: RVM (thick green lines); NB (dotted blue lines); SVM (dashed red lines).

By comparing TN and TP discharge from different rivers, it can be seen that large reductions in both nutrients are needed from multiple sources to substantially reduce the frequency of *K. brevis* blooms (Figure 8), based on the results from the RVM. These comparisons, based on variations of 1–2 standard deviations from the mean (and setting other features to the mean), illustrate the magnitude of reductions necessary to reduce the probability of blooms from >0.6 to <0.2.

3.4. Role of Sea Surface Height

Sea surface height difference was chosen as one of the explanatory variables in our machine learning algorithms because previous studies [48] have related this variable to the position of the Loop Current, and the associated temperature and degree of upwelling. Results for the RVM model were nearly identical with or without this explanatory variable (Figure 9) and, given this outcome, this factor was not tested with the other models. This may be due to the fact that the Loop Current is inhibited from penetrating the WFS due to the sloping topography; shelf currents are controlled to a larger extent by local winds [92]. Nevertheless, for 1998, 2002, 2009, 2010, and 2013, when the Loop Current was in its southern position, the RVM model generally had a much lower precision value (0.36, 0.42, 0.00, 0.00, 0.25, and an average of 0.21) versus 0.35 for all years. This suggests other factors

not considered in the explanatory variables may be needed to improve bloom prediction for those years.

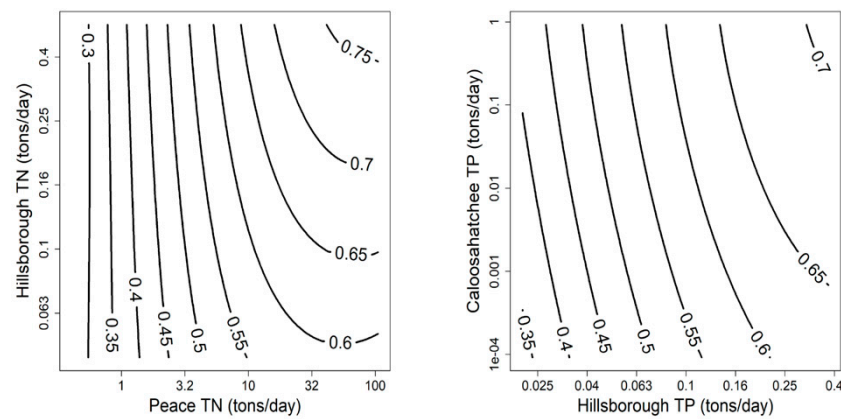


Figure 8. Contour plots of *K. brevis* probability as a function of combinations of riverine nutrient loads, obtained from the sensitivity analysis using the RVM. Left panel: Hillsborough and Peace River total nitrogen (TN). Right panel: Hillsborough and Caloosahatchee total phosphorus (TP).

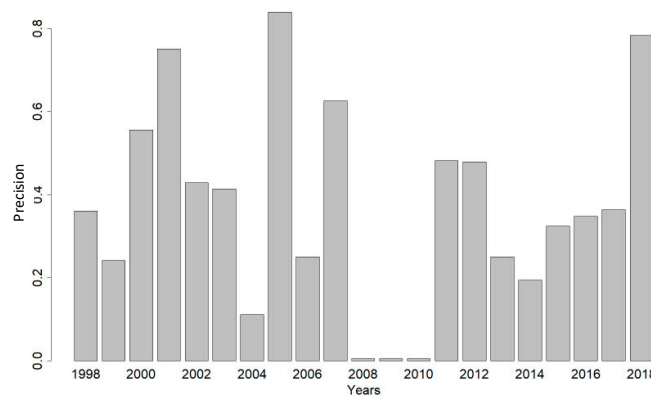


Figure 9. Precision of the RVM model by year.

4. Discussion

Machine learning provides powerful approaches for predicting HAB occurrences and testing their potential change under different conditions. The flexibility of these tools and ability to find data-driven solutions when mechanistic relationships are yet unknown or hard to parameterize are particularly important for advancing research on factors that result in toxin-producing blooms. Prior modeling of HABs accommodating nonlinearity of relationships and non-normality of distributions has been undertaken by statistical regression models, and the power of machine learning techniques is beginning to be recognized [39]. Machine learning has become a powerful tool in water quality assessment, from freshwater to marine waters [93–95].

Models, however imperfect, can be useful for testing the strength of particular factors or variables on outcomes. In silico experiments in which nutrient sources are turned on or off, or climate variables altered, provide clear clues regarding the importance of such factors both for present and future conditions [26,27]. Such in silico experiments are also insightful regarding the potential magnitude of impacts of nutrient reductions if undertaken by management.

Models may yield either false positives (indicating blooms or conditions for blooms when they do not occur) or false negatives (indicating no bloom when in fact they do occur). Both can be problematic in terms of protecting human health and economies. False positives may be preferable if the goal is to protect human health—better to be “safe than sorry”—but false positives can also be more expensive economically [31]. For example, a

fishery may be closed when it was not necessary to do so. False negatives are not protective of human or ecological health.

In this study, four different machine learning classifiers were used to predict the likelihood of *K. brevis* blooms between 1998 and 2018. Comparing the 20 year monitoring dataset of the abundance of this dinoflagellate using all algorithms, RVM and NB were found to have better skills in bloom prediction than the two other approaches. All models were comparable in how frequently false negatives were reported. Because the number of weeks with blooms was about 42% of the number of non-bloom weeks, it required the classifiers to learn from an imbalanced dataset. This challenge was resolved by two different methods that oversample the minority class: random oversampling and generation of synthetic data using SMOTE. The predictive skills were very similar between the two data sampling methods (Table 1). This result is perhaps not surprising because the ratio of the samples in the minority to the samples in the majority class was only 1:2.4, in comparison to models in which the class imbalance reached 1:10 or 1:100 [53].

Another feasible alternative machine learning model is that of the random forest. This approach often outperforms other models and has the option of ranking variables by their individual importance. However, in the current data set for which there are ~1000 data points, any resulting decision tree would be based only on 10s–100s of data points. Thus, the approaches used herein were more suitable to the size of the data set.

Both *k*-fold and block cross-validation methods were used to evaluate the predictive skills of the machine learning classifiers. Although the SVM achieved good scores (recall = 0.63, precision = 0.64, and F1 = 0.64 using random oversampling) during the *k*-fold cross-validation, its performance deteriorated significantly during the block cross-validation (recall = 0.27, precision = 0.32, and F1 = 0.29) (Table 1). A similar deterioration was seen in the ANN between *k*-fold and block cross-validation methods. When trained using the random oversampling approach and tested with the *k*-fold cross-validation procedure, the SVM used 533 support vectors and the ANN had a high Akaike Information Criterion score of 1814. It is possible that these two algorithms overfitted the training data and their predictive skills deteriorated when tested on completely independent data, as undertaken in the block cross-validation analysis. In contrast, the RVM had only 19 relevance vectors, and NB was a simple probabilistic classifier, thus producing more robust results. The accuracy, recall, precision, and F1 scores of RVM and NB remained higher, regardless of the cross-validation methods.

Blooms of *K. brevis* occur almost annually in the eastern Gulf of Mexico, typically initiating in early fall, but varying in intensity and duration. The bloom of 2017–2019 was among the largest and most expensive in recent history. It caused the deaths of hundreds of tons of fish, hundreds of manatees, dolphins, and sea turtles, in addition to many cases of respiratory distress [13]. Fisheries closures, and revenue lost by local businesses, also had economic impacts in the tens to hundreds of millions of dollars [11]. Understanding the links between physical controls (upwelling, river flow), nutrient inputs, and extreme weather events has been a high priority in order to make long-term predictions to protect environmental and human health [14,18,34]. The results reported herein confirm that wind direction, river flow, and nutrient load are important explanatory variables with regard to *K. brevis* probabilities. Although sea surface height (as a proxy for the Loop Current) did not contribute to improved forecasts, the increase in false positives for select years (lower precision values) suggests that the height values alone do not capture the effect of this current adequately, i.e., it was not a sensitive proxy.

Using a convolutional NN approach, Hill et al. [46] achieved high accuracy in detecting blooms of *K. brevis*. There are several differences between the methodology applied herein and the analysis of Hill et al. [46]. The study of Hill et al. [46] used satellite remote sensing of chlorophyll *a* as a proxy to detect *K. brevis*, whereas direct cell counts were used here. Moreover, Hill et al. [46] did not explore the role of wind speed, river flow, or nutrient loads. These results are complementary and show the promise of machine learning approaches not only in modeling various aspects of *K. brevis* blooms, but of HAB events more generally.

Although there have been debates about the extent to which anthropogenic nutrients fuel *K. brevis* blooms [10,14], and references therein], there is no doubt that Florida's continuing population growth has accelerated eutrophication. Florida is working to establish nutrient reduction targets to mitigate water quality problems in the water bodies. With Florida's continuing population growth and its coastal development and dependence on tourism, more people are exposed to *K. brevis* and its toxins than in previous decades, and the prolonged duration of recent blooms is also increasing the period of exposure when blooms do occur [14]. Blooms, which traditionally occur from the late summer until late winter, have been sustained throughout the summer months in recent years, raising important questions about the interplay of physical controls and biological responses, including changing temperature conditions. The nutritional pathways, and sources of nutrients supporting *K. brevis* blooms, are complex [14,21,24,25]. The fact that nutrient loads have increased is, in itself, an insufficient explanation for the expansion in *K. brevis* blooms. The right nutrients are required at the right time to create conditions conducive for these blooms to form [7]. Changes in flow, such as that due to hurricanes or intensive wet weather, bring new nutrients that can help to support blooms. The statistical analysis by Maze et al. [48] indicates that there are significant differences in the Peace and Caloosahatchee River flows between periods of large blooms and periods without blooms. It is important to note that flows from the Caloosahatchee are actively managed and regulated—to reduce potential flooding in the Lake Okeechobee region—whereas those of the Peace are not. The machine learning algorithms used here illustrated strong relationships between river flow and blooms. Strong river flow that occurs following hurricanes, regardless of whether those flows are natural or enhanced by active management, especially when these flows follow extended droughts, deliver substantial nutrient loads.

Air temperatures over Eastern North America (including Florida) are expected to increase ~ 1.5 °C by 2050 and 3–4 °C by 2100 (relative to 2000), according to recent climate projections [96]. Additionally, rainfall over Florida is projected to decrease by 20–30% during the summer but to increase by 10–20% during the fall–winter, which is the season during which *K. brevis* blooms typically occur. This work underscores the important interactive roles of nutrient pollution and river flow in the increased frequency of *K. brevis* blooms in Florida. Due to climate change and the predicted increase in extreme precipitation events in a warming climate [97–99], it is expected that HABs will occur more frequently in the future, in Florida and elsewhere, unless substantial reductions in TN and TP land-based use and loading in the major rivers are achieved.

5. Conclusions

In conclusion, four new machine learning models were developed for the WFS and explored with regard to wind direction, temperature, river flow, nutrient load, and sea surface height as explanatory variables in predicting *K. brevis* blooms. The models had different strengths due to the differing degrees of complexity of the models, and they responded differently to the cross-validation procedures used. Overall, the RVM and NB models performed the best in predicting past events. By manipulating the range of explanatory variables, insight into the strength of their impact on blooms was obtained. These findings highlight that not only are reductions in both N and P necessary to reduce blooms, but reductions from multiple rivers are more effective than reductions from a single river. These models can be helpful in exploring the most effective combinations of nutrient reductions. Because river drainage basins are large, a 10–20% increase in fall–winter rainfall will translate into increases in discharges of multiple rivers with their combined higher nutrient loads during the *K. brevis* bloom period. This implies that to control blooms through nutrient reductions, greater reductions will be required than under present day flow conditions.

Author Contributions: Conceptualization, M.F.L.; methodology, M.F.L. and V.L.; validation, M.F.L.; resources, M.F.L. and P.M.G.; writing—original draft preparation, M.F.L.; writing—review and editing, P.M.G. and V.L.; supervision, P.M.G.; project administration, P.M.G.; funding acquisition, P.M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Oceanic and Atmospheric Administration (NOAA) National Centers for Coastal Ocean Science Competitive Research program under awards No. NA17NOS4780180 and NA19NOS4780183.

Data Availability Statement: All the data and code are publicly available and accessible online. The data and code can be found at: https://github.com/mfli-ml/WFS_ML/.

Acknowledgments: We thank T. Kana for helpful comments on this manuscript. M.F.L. also thanks James Bennett High School, Salisbury MD. This is contribution number 6038 of the University of Maryland Center for Environmental Science and contribution number ECO997 from the NOAA ECOHAB program.

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

References

- Anderson, D.M. Toxic Algal Blooms and Red Tides: A Global Perspective. In *Red Tides: Biology, Environmental Science, and Toxicology*; Okaichi, T., Anderson, D., Nemoto, T., Eds.; Elsevier Science Publishing Company: New York, NY, USA, 1989; pp. 11–16.
- Hallegraeff, G.M. A review of harmful algal blooms and their apparent global increase. *Phycologia* **1993**, *32*, 79–99. [[CrossRef](#)]
- Glibert, P.M.; Burkholder, J.M. Causes of harmful algal blooms. In *Harmful Algal Blooms: A Compendium Desk Reference*; Shumway, S., Burkholder, J.M., Morton, S.L., Eds.; Wiley Blackwell: Singapore, 2018; pp. 1–38.
- Heisler, J.; Glibert, P.M.; Burkholder, J.; Anderson, D.; Cochlan, W.; Dennison, W.; Dortch, Q.; Heil, C.; Humphries, E.; Lewitus, A.; et al. Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae* **2008**, *8*, 3–13. [[CrossRef](#)]
- Fu, F.X.; Tatters, A.O.; Hutchins, D.A. Global change and the future of harmful algal blooms in the ocean. *Mar. Ecol. Progr. Ser.* **2012**, *470*, 207–233. [[CrossRef](#)]
- Wells, M.L.; Trainer, V.L.; Smayda, T.J.; Karlson, B.S.; Trick, C.G.; Kudela, R.M.; Ishikawa, A.; Bernard, S.; Wulff, A.; Anderson, D.A.; et al. Harmful algal blooms and climate change: Learning from the past and present to forecast the future. *Harmful Algae* **2015**, *49*, 68–93. [[CrossRef](#)]
- Glibert, P.M.; Burford, M.A. Globally changing nutrient loads and harmful algal blooms: Recent advances, new paradigms and continuing challenges. *Oceanography* **2017**, *30*, 44–55. [[CrossRef](#)]
- Glibert, P.M. Harmful algal at the complex nexus of eutrophication and climate change. *Harmful Algae* **2020**, *9*. [[CrossRef](#)]
- Steidinger, K.A. Historical perspective on *Karenia brevis* red tide research in the Gulf of Mexico. *Harmful Algae* **2009**, *8*, 549–561. [[CrossRef](#)]
- Brand, K.; Compton, A. Long-term increase in *Karenia brevis* abundance along the southwest Florida coast. *Harmful Algae* **2007**, *6*, 232–252. [[CrossRef](#)] [[PubMed](#)]
- Glibert, P.M. Why were the water and beaches in west Florida so gross in summer 2018? Red tides! *Front. Young Minds* **2019**, *7*. [[CrossRef](#)]
- Fears, D.; Rozsa, L. Florida's Unusually Long Red Tide Is Killing Wildlife, Tourism and Businesses. The Washington Post. Available online: https://www.washingtonpost.com/national/health-science/floridas-unusually-long-red-tide-is-killing-wildlife-tourism-and-businesses/2018/08/28/245fc8da-aad5-11e8-8a0c-70b618c98d3c_story.html (accessed on 12 December 2018).
- Monuz, C.R. Red tide episode kills record number of sea turtles. *Herald Tribune*, 15 January 2019.
- Heil, C.A.; Bronk, D.A.; Dixon, L.K.; Hitchcock, G.L.; Kirkpatrick, G.J.; Mulholland, M.R.; O'Neil, J.M.; Walsh, J.J.; Weisberg, R.H.; Garrett, M. The Gulf of Mexico ECOHAB: *Karenia* program 2006–2012. *Harmful Algae* **2014**, *38*, 3–7. [[CrossRef](#)]
- Weisberg, R.H.; He, R. Local and deep-ocean forcing contributions to anomalous water properties on the West Florida Shelf. *J. Geophys. Res.* **2003**, *108*, 3184. [[CrossRef](#)]
- Liu, Y.; Weisberg, R.H. Seasonal variability on the West Florida Shelf. *Progr. Oceanogr.* **2012**, *104*, 80–98. [[CrossRef](#)]
- Mayer, D.A.; Weisberg, R.H.; Zheng, L.; Liu, Y. Winds on the West Florida Shelf: Regional comparisons between observations and model estimates. *J. Geophys. Res.* **2017**, *122*, 834–846. [[CrossRef](#)]
- Weisberg, L.; Zheng, L.; Liu, Y.; Lembke, C.; Lenes, J.M.; Walsh, J.J. Why a red tide was not observed on the west Florida continental shelf in 2010. *Harmful Algae* **2014**, *38*, 119–126. [[CrossRef](#)]
- Liu, Y.; Weisberg, R.H.; Lenes, J.M.; Zheng, L.; Hubbard, K.; Walsh, J.J. Offshore forcing on the “pressure point” of the West Florida Shelf: Anomalous upwelling and its influence on harmful algal blooms. *J. Geophys. Res.* **2016**, *121*, 5501–5515. [[CrossRef](#)]
- Hu, C.; Muller-Karger, F.E.; Swarzenski, P.W. Hurricanes, submarine groundwater discharge, and Florida's red tides. *Geophys. Res. Lett.* **2006**, *33*, L11601. [[CrossRef](#)]

21. Vargo, G.A.; Heil, C.A.; Fanning, K.A.; Dixon, K.L.; Neely, M.B.; Lester, K.; Ault, D.; Murasko, S.; Havens, J.; Walsh, J.; et al. Nutrient availability in support of *Karenia brevis* blooms on the central West Florida Shelf: What keeps *Karenia* blooming? *Cont. Shelf Res.* **2008**, *28*, 73–98. [[CrossRef](#)]
22. Vargo, G.A. A brief summary of the physiology and ecology of *Karenia brevis* Davis (G. Hansen and Moestrup comb. nov.) red tides on the West Florida Shelf and of hypotheses posed for their initiation, growth, maintenance, and termination. *Harmful Algae* **2009**, *8*, 573–584. [[CrossRef](#)]
23. Lenos, J.M.; Darrow, B.A.; Walsh, J.J.; Prospero, J.M.; He, R.; Weisberg, R.H.; Vargo, G.A.; Heil, C.A. Saharan dust and phosphatic fidelity: A three-dimensional biogeochemical model of *Trichodesmium* as a nutrient source for red tides on the West Florida Shelf. *Cont. Shelf Res.* **2008**, *28*, 1091–1115. [[CrossRef](#)]
24. Glibert, P.M.; Burkholder, J.M.; Kana, T.M.; Alexander, J.A.; Schiller, C.; Skelton, H. Grazing by *Karenia brevis* on *Synechococcus* enhances their growth rate and may help to sustain blooms. *Aquat. Microb. Ecol.* **2009**, *55*, 17–30. [[CrossRef](#)]
25. O’Neil, J.M.; Heil, C.A. Preface to ECOHAB: *Karenia* Special Edition of Harmful Algae. *Harmful Algae* **2014**, *38*, 1–2. [[CrossRef](#)]
26. Glibert, P.M.; Allen, J.I.; Bouwman, L.; Brown, C.; Flynn, K.J.; Lewitus, A.; Madden, C. Modeling of HABs and eutrophication: Status, advances, challenges. *J. Mar. Syst.* **2010**, *83*, 262–275. [[CrossRef](#)]
27. Glibert, P.M.; Beusen, A.H.W.; Bouwman, A.F.; Burkholder, J.M.; Flynn, K.J.; Heil, C.A.; Li, M.; Lin, C.H.; Madden, C.J.; Mitra, A.; et al. Multifaceted climatic and nutrient effects on harmful algae require multifaceted model. In *Climate Change and Marine and Freshwater Toxins*, 2nd ed.; Botana, L.M., Louzao, C., Vilarinho, N., Eds.; DeGruyter Publishers: Berlin, Germany, 2021; pp. 473–518.
28. McGillicuddy, D.J., Jr.; de Young, B.; Doney, S.; Glibert, P.M.; Stammer, D.; Werner, F.E. Models: Tools for synthesis in international oceanographic research programs. *Oceanography* **2010**, *23*, 126–139. [[CrossRef](#)] [[PubMed](#)]
29. Anderson, D.M. HABs in a changing world: A perspective on harmful algal blooms, their impacts, and research and management in a dynamic era of climatic and environmental change. In *Harmful Algae 2012, Proceedings of the 15th International Conference on Harmful Algae: 29 October—2 November 2012*; Kim, H.-G., Reguera, B., Hallegraeff, G.M., Lee, C.K., Han, M.S., Choi, J.K., Eds.; CECSO: Changwon, Gyeongnam, 2014; pp. 3–17.
30. Franks, P.J.S. Recent advances in modeling of harmful algal blooms. In *Global Ecology and Oceanography of Harmful Algal Blooms*; Glibert, P.M., Berdalet, E., Burford, M., Pitcher, G., Zhou, M.J., Eds.; Springer: Cham, Switzerland, 2018; pp. 359–380.
31. Flynn, K.J.; McGillicuddy, D.J. Modeling marine harmful algal blooms: Current status and future prospects. In *Harmful Algal Blooms: A Compendium Desk Reference*; Shumway, S., Burkholder, J.M., Morton, S.L., Eds.; Wiley Blackwell: Singapore, 2018; pp. 115–134.
32. Stumpf, R.P.; Culver, M.E.; Tester, P.A.; Tomlinson, M.; Kirkpatrick, G.J.; Pederson, B.A.; Truby, E.; Ransibrahmanakul, V.; Soracco, M. Monitoring *Karenia brevis* blooms in the Gulf of Mexico using satellite ocean color imagery and other data. *Harmful Algae* **2003**, *2*, 147–160. [[CrossRef](#)]
33. Stumpf, R.P.; Tomlinson, M.C.; Calkins, J.A.; Kirkpatrick, B.; Fisher, K.; Nierenberg, K.; Currier, R.; Wynne, T.T. Skill assessment for an operational algal bloom forecast system. *J. Mar. Syst.* **2009**, *76*, 151–161. [[CrossRef](#)] [[PubMed](#)]
34. Weisberg, R.H.; Barth, A.; Alvera-Azcárate, A.; Zheng, L. A coordinated coastal ocean observing and modeling system for the West Florida Shelf. *Harmful Algae* **2009**, *8*, 585–598. [[CrossRef](#)]
35. Walsh, J.J.; Weisberg, R.H.; Dieterle, D.A.; He, R.; Darrow, B.P.; Jolliff, J.K.; Lester, K.M.; Vargo, G.A.; Kirkpatrick, G.J.; Fanning, K.A.; et al. Phytoplankton response to intrusions of slope water on the West Florida Shelf: Models and observations. *J. Geophys. Res.* **2003**, *108*, 15. [[CrossRef](#)]
36. Walsh, J.J.; Jolliff, J.K.; Darrow, B.P.; Lenos, J.M.; Milroy, S.P.; Remsen, A.; Dieterle, D.A.; Carder, K.L.; Chen, F.R.; Vargo, G.A.; et al. Red tides in the Gulf of Mexico: Where, when, and why. *J. Geophys. Res.* **2006**, *111*, 1–46. [[CrossRef](#)]
37. Milroy, S.P.; Dieterle, D.A.; He, R.; Kirkpatrick, G.J.; Lester, K.M.; Steidinger, K.A.; Vargo, G.A.; Walsh, J.J.; Weisberg, R.H. A three-dimensional biophysical model of *Karenia brevis* dynamics on the west Florida shelf: A look at physical transport and potential zooplankton grazing controls. *Cont. Shelf Res.* **2008**, *28*, 112–136. [[CrossRef](#)]
38. Lenos, J.M.; Darrow, B.P.; Walsh, J.J.; Jolliff, J.K.; Chen, F.L.; Weisberg, R.W.; Zheng, L. A 1-D simulation analysis of the development and maintenance of the 2001 red tide of the ichthyotoxic dinoflagellate *Karenia brevis* on the West Florida shelf. *Cont. Shelf Res.* **2012**, *41*, 92–110. [[CrossRef](#)]
39. Cruz, R.C.; Reis Costa, P.; Vinga, S.; Krippahl, L.; Lopes, M.B. A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. *J. Mar. Sci. Eng.* **2021**, *9*, 283. [[CrossRef](#)]
40. Lee, J.H.W.; Huang, Y.; Dickman, M.; Jayawardena, A.W. Neural network modelling of coastal algal blooms. *Ecol. Model.* **2003**, *159*, 179–201. [[CrossRef](#)]
41. Velo-Suarez, L.; Gutierrez-Estrada, J.C. Artificial neural network approaches to one-step weekly prediction of *Dinophysis acuminata* blooms in Huelva (Western Andalusia, Spain). *Harmful Algae* **2007**, *6*, 361–371. [[CrossRef](#)]
42. Guallar, C.; Delgado, M.; Diogène, J.; Fernández-Tejedo, M. Artificial neural network approach to population dynamics of harmful algal blooms in Alfacs Bay (NW Mediterranean): Case studies of *Karlodinium* and *Pseudo-nitzschia*. *Ecol. Model.* **2016**, *338*, 37–50. [[CrossRef](#)]
43. Xie, Z.; Lou, I.; Ung, W.K.; Mok, K.M. Freshwater algal bloom prediction by support vector machine in Macau storage reservoirs. *Math. Prob. Eng.* **2012**, *2012*, 397473. [[CrossRef](#)]
44. Shen, J.; Qin, Q.; Wang, Y.; Sisson, M. A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to nutrient loading. *Ecol. Model.* **2019**, *398*, 44–54. [[CrossRef](#)]

45. Gokaraju, B.; Durbha, S.S.; King, R.L.; 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. Selected Topics Appl. Earth Observ. Rem. Sens.* **2011**, *4*, 710–720. [CrossRef]
46. Hill, P.R.; Kumar, A.; Temini, M.; Bull, D.R. HABNet: Machine learning, remote sensing based detection and prediction of harmful algal blooms. *IEEE J. Selected Topics Appl. Earth Observ. Rem. Sens.* **2019**, arXiv:1912.02305.
47. Florida Fish and Wildlife Conservation Commission. Available online: <https://myfwc.com/research/redtide/> (accessed on 24 February 2020).
48. Maze, G.; Olascoaga, M.J.; Brand, L. Historical analysis of environmental conditions during Florida red tide. *Harmful Algae* **2015**, *50*, 1–7. [CrossRef]
49. US Water Data for the Nation. Available online: <https://waterdata.usgs.gov/nwis> (accessed on 24 October 2020).
50. University of South Florida Water Institute. Available online: <http://www.wateratlas.usf.edu> (accessed on 6 February 2020).
51. National Data Buoy Center. Available online: <https://www.ndbc.noaa.gov/> (accessed on 10 October 2020).
52. E.U. Copernicus Marine Service Monitoring Service (CMEMS). Available online: <http://marine.copernicus.eu/> (accessed on 6 February 2020).
53. Sun, Y.; Wong, A.K.C.; Kamel, M.S. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [CrossRef]
54. Japkowicz, N.; Holte, R. Workshop report: AAAI2000 workshop on learning from imbalanced data-sets. *AI Mag.* **2000**, *22*, 127–136.
55. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
56. Fernandez, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Art. Intel. Res.* **2018**, *61*, 863–905. [CrossRef]
57. Haibo, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data. Engin.* **2009**, *21*, 1263–1284. [CrossRef]
58. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
59. Nello, C.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
60. Basak, D.; Pal, S. Patranabis. Support vector regression. *Neural Info. Process Letts Rev.* **2007**, *11*, 203–224.
61. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
62. Boser, B.; Guyon, I.; Vapnik, V. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
63. Schölkopf, B.; Smola, A. Support Vector Machines and Kernel Algorithms. In *Encyclopedia of Biostatistics*; Armitage, P., Colton, T., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 2002; pp. 5328–5335.
64. Tipping, M.E. Sparse Bayesian Learning and the Relevance Vector Machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244. [CrossRef]
65. Camps-Valls, G.; Gómez-Chova, L.; Muñoz-Marí, J.; Vila-Francés, J.; Amorós-López, J.; Calpe-Maravilla, J. Retrieval of oceanic chlorophyll concentration with relevance vector machines. *Remote Sens. Environ.* **2006**, *105*, 23–33. [CrossRef]
66. Maron, M.E. Automatic indexing: An experimental inquiry. *J. Assoc. Comp. Mach.* **1961**, *8*, 404–417. [CrossRef]
67. Hand, D.J.; Yu, K. Idiots Bayes—not so stupid after all? *Int. Stat. Rev.* **2001**, *69*, 385–398.
68. Hassoun, M.H. *Fundamentals of Artificial Neural Networks*; The MIT Press: Massachusetts, MA, USA, 1995; 511p.
69. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
70. Werbos, P. Backpropagation through time: What it does and how to do it. *Proc. IEEE* **1990**, *78*, 1550–1560. [CrossRef]
71. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learn*; MIT Press: Cambridge, MA, USA, 2016.
72. Hijmans, R. Raster: Geographic Data Analysis and Modeling. R Package Version 3.0–7. 2017. Available online: <https://CRAN.R-project.org/package=raster> (accessed on 12 March 2019).
73. Calaway, R.; Microsoft Corporation; Weston, S.; Tenenbaum, D. doParallel: Foreach Parallel Adaptor for the 'Parallel' Package. 2017. R Package Version 1.0.15. Available online: <https://CRAN.R-project.org/package=doParallel> (accessed on 12 March 2019).
74. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. Kernlab. An S4 package for kernel methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20. [CrossRef]
75. Torgo, L. *Data Mining Using R: Learning with Case Studies*; CRC Press: Boca Raton, FL, USA, 2010; pp. 209–211.
76. Schnute, J.; Boers, M.; Haigh, R.; Couture-Beil, A.; Chabot, D.; Grandin, C.; Johnson, A.; Wessel, P.; Antonio, F.; Lewin-Koh, N.J.; et al. PBSmapping: Mapping Fisheries Data and Spatial Analysis Tools. R Package Version 2.70.4. 2017. Available online: <https://CRAN.R-project.org/package=PBSmapping> (accessed on 13 March 2019).
77. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F.; Chang, C.C.; Lin, C.C. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R Package Version 1.7–2. 2021. Available online: <https://CRAN.R-project.org/package=e1071> (accessed on 2 May 2021).
78. Fritsch, S.; Guenther, F.; Wright, M.N.; Suling, M.; Mueller, S.M. Training of Neural Networks. R Package Version 1.44.2. 2019. Available online: <https://CRAN.R-project.org/package=neuralnet> (accessed on 3 October 2019).
79. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. R Package Version 3.3.5. 2020. Available online: <https://cran.r-project.org/package=ggplot2> (accessed on 2 January 2021).

80. R Core Team. A Language and Environment for Statistical Computing. 2017. R: R Foundation for Statistical Computing, Vienna, Austria. Available online: <https://www.R-project.org/> (accessed on 13 March 2019).
81. Anguita, D.; Ghio, A.; Ridella, S.; Sterpi, D. K-Fold cross validation for error rate estimate in support vector machines. In Proceedings of the 2009 International Conference on Data Mining, Miami, FL, USA, 13–16 July 2009; DMIN: Las Vegas, NV, USA, 2009; pp. 1–7.
82. Cawley, G.C.; Tablot, N.L.C. Fast exact leave-one-out cross validation of sparse least-squared support vector machines. *Neural Netw.* **2004**, *17*, 1467–1475. [[CrossRef](#)] [[PubMed](#)]
83. Stone, M. Cross-validated choice and assessment of statistical predictions. *J. Roy. Stat. Soc. Ser. B* **1974**, *36*, 111–133. [[CrossRef](#)]
84. Geisser, S. The predictive sample reuse method with applications. *J. Amer. Stat. Assoc.* **1975**, *70*, 320–328. [[CrossRef](#)]
85. Bergmeir, C.; Benitez, J.M. On the use of cross-validation for time series predictor evaluation. *Inform. Sci.* **2012**, *191*, 192–213. [[CrossRef](#)]
86. Roberts, D.R. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [[CrossRef](#)]
87. Burman, P.R.; Chow, E.; Nolan, D. A cross-validated method for dependent data. *Biometrika* **1994**, *81*, 351–358. [[CrossRef](#)]
88. Racine, J. Consistent cross-validated model-selection for dependent data: Hv-block cross-validation. *J. Economet.* **2000**, *99*, 39–61. [[CrossRef](#)]
89. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
90. Platt, J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*; Smola, A.J., Bartlett, P., Schölkopf, S., Eds.; MIT Press: Cambridge, MA, USA, 1999; pp. 61–74.
91. Lin, H.; Lin, C.; Weng, R.C. A note on Platt’s probabilistic outputs for support vector machines. *Mach. Learn.* **2007**, *68*, 267–276. [[CrossRef](#)]
92. He, R.; Weisberg, R.H. A Loop Current intrusion case study on the West Florida Shelf. *J. Phys. Oceanogr.* **2003**, *33*, 465–477. [[CrossRef](#)]
93. Hadjisolomou, E.; Stefanidis, K.; Herodotou, H.; Michaelides, M.; Papatheodorou, G.; Papastergiadou, E. Modelling freshwater eutrophication with limited limnological data using artificial neural networks. *Water* **2021**, *13*, 1590. [[CrossRef](#)]
94. Deng, T.; Chau, K.W.; Duan, H.-F. Machine learning based marine water quality prediction for coastal hydro-environment management. *J. Envir. Manag.* **2021**, *284*, 112051. [[CrossRef](#)] [[PubMed](#)]
95. Zhou, Y. Real-time probabilistic forecasting of river water quality under data missing situation: Deep learning plus post-processing techniques. *J. Hydrol.* **2020**, *589*, 125164. [[CrossRef](#)]
96. IPCC (Intergovernmental Panel on Climate Change). Impacts, adaptation, and vulnerability, Summary for Policymakers. In *Climate Change*; Field, C.B., Barros, V.R., Dokken, D.J., Mach, K.J., Mastrandrea, M.D., Bilir, T.E., Chatterjee, M., Ebi, K.L., Estrada, Y.O., Genova, R.C., et al., Eds.; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2014; pp. 1–32.
97. Sillmann, J.; Kharin, V.V.; Zhang, X.; Zwiers, F.W.; Bronaugh, D. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *J. Geophys. Res. Atmos.* **2013**, *118*, 1716–1733. [[CrossRef](#)]
98. Sillmann, J.; Kharin, V.V.; Zwiers, F.W.; Zhang, X.; Bronaugh, D. Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *J. Geophys. Res. Atmos.* **2013**, *118*, 2473–2493. [[CrossRef](#)]
99. Russo, S.; Dosio, A.; Graversen, R.G.; Sillmann, J.; Carrao, H.; Dunbar, M.B.; Singleton, A.; Montagna, P.; Barbola, P.; Vogt, J.V. Magnitude of extreme heat waves in present climate and their projection in a warming world. *J. Geophys. Res. Atmos.* **2014**, *119*, 12500–12512. [[CrossRef](#)]