**Machine Learning Classification Algorithms Predict**

***Karenia brevis* Blooms on the West Florida Shelf**

**Marvin F. Li1**

**Patricia M. Glibert2\***

1James M. Bennett High School, 300 E College Ave, Salisbury, MD 21804 USA  
2University of Maryland Center for Environmental Science, Horn Point Laboratory, PO Box 775, Cambridge, MD 21613 USA;

\*Corresponding author: glibert@umces.edu. ORCHID: 0000-0001-5690-1674

**Abstract**

Harmful Algal Blooms (HABs), events that cause fish kills and create human health problems by poisoning seafood and contaminating water supplies, have increased in frequency, magnitude, and impacts around the world. From 2017 to early 2019, blooms of the toxic dinoflagellate *Karenia brevis* swept over the West Florida coast, resulting in thousands of tons of dead fish, deaths to many other marine organisms, numerous respiratory-related hospitalizations, and hundreds of millions of dollars in economic damage. Machine learning algorithms, including Support Vector Machine (SVM), Relevance Vector Machine (RVM), Naïve Bayes classifier (NB) and Artificial Neural Network (ANN), were developed to predict *K. brevis* blooms by using wind, temperature, streamflow, nutrient, and satellite altimetry data. Comparing the 20-year monitoring data set of abundance of this dinoflagellate using all algorithms, SVM was found to have the highest accuracy in bloom prediction, 62%. Using the SVM model, we showed that upwelling-favorable northerly winds increase *K. brevis* probability a and on-shore westerly winds prevent blooms from dispersing offshore. It was also found that large river flows supply the nutrients that fuelblooms. 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.

Highlights

* Machine learning algorithms had high accuracy in predicting *Karenia brevis* blooms
* Algorithms accounted for wind temperature, streamflow, and nutrient conditions
* Northerly winds increase bloom probability; westerly winds support blooms inshore
* Riverine nutrients from multiple rivers will be required to reduce blooms

**Graphical abstract**



**1. Introduction**

Harmful algal blooms (HABs) have been increasing globally, with more HABs, more often in new and different places, often lasting longer and having a wide range of environmental impacts and toxicities (e.g., Anderson, 1989; Hallegraeff, 1993; Glibert and Burkholder, 2018). Both nutrient pollution and climate change are now recognized to play important roles in this expansion (Anderson, 2002; Heisler et al., 2008; Fu et al., 2012; Wells et al., 2015; Glibert and Burford, 2017; Glibert, 2019a).

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 (Steidinger, 2009). However, recent analyses suggest that bloom events have increased 15-fold from the 1950s to 1990s (Brand and Compton, 2007). From 2017-2019, southwest Florida experienced an unusually prolonged (18 months) *K. brevis* bloom. At its maximum, this bloom covered a region with more than 250 km of coastline, encompassing recreational beaches and numerous commercial and recreational shellfish beds (Fig. 1; Glibert, 2019b). With Florida’s continuing population growth, more people are exposed to *K. brevis* and its toxins than in earlier years and the prolonged duration of recent blooms is increasing the period of exposure (Heil et al., 2014).

While *K. brevis* is typically thought of as a coastal bloom species, blooms are actually initiated offshore and then transported to coastal waters where they flourish and persist for months in nutrient-rich waters (Steidinger, 2009). Upwelling transports *K. brevis* cells to the coast (Weisberg and He, 2003; Liu and Weisberg, 2012; Mayer et al., 2017), but strong upwelling over the shelf break may actually suppress *K. brevis* blooms or favor competing taxa such as diatoms (Weisberg et al., 2014; Liu et al., 2016). The nutrient sources, pathways, and processes supporting and maintaining *K. brevis* blooms include not only upwelling, but also riverine nutrient inputs that bring wastewater effluent and agricultural runoff. Other nutrient sources include benthic nutrient fluxes, atmospheric deposition, nutrients released by other phytoplankton and decaying fish from fish kills, submarine groundwater discharge, and mixotrophic grazing - suggesting complex environmental interactions of this important driver (Hu et al., 2006; Vargo et al., 2008; Vargo, 2009; Lenes et al., 2008; Glibert et al., 2009; Heil et al., 2014; O’Neil and Heil, 2014).

The massive bloom of 2017-2019, as well as the large-scale bloom observed in 2005, were likely propelled by unusual events. Hu et al. (2006) 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), Michael (2018) and Tropical Storm Gordon (2018) are suspected of contributing to the severity of the 2017-2019 *K. brevis* bloom (Glibert, 2019b). 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.

There is a strong need to advance predictions of *K. brevis*, and other HABs more generally, but there are many challenges in modeling discrete HAB species (Glibert et al., 2010; McGillicuddy et al., 2010; Anderson, 2014; Franks, 2018; Flynn and McGillicuddy, 2018). There are several types of models in operational use for *K. brevis* (Weisberg and He, 2003; Walsh et al., 2003; Stump et al., 2009). An operational forecasting system, maintained by the National Oceanic and Atmospheric Administration, provides 3-5 day outlooks of blooms, using satellite remote sensing of chlorophyll *a*, in-situ sampling, and wind buoy data (Stump et al., 2003). The main goal of these forecasts is to inform managers and the public in coastal areas where public health may be compromised (Stump et al., 2009). However, modeling longer-term trends has been limited.

Machine learning approaches have also been used in predicting HABs in the Gulf of Mexico, but with different objectives. Liu and Weisberg (2012) used such approaches to demonstrate the role of deep-ocean forcing on the West Florida Shelf in major bloom occurrences. Weisberg et al. (2014) 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 outcompetes any nascent *Karenia brevis* bloom. Liu et al. (2016) 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 during 1998, 2002, 2009, 2010, 2013. This 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. Hill et al. (2019) used satellite remote sensing of chlorophyll from 2003 to 2018, as well as sea surface temperature and bathymetry, as inputs to a convolutional neural network (designed for spatial data) to predict the presence of a *K. brevis* event in the near future (2-8 days). They also used the technique of long short-term memory to process the sequential data.

Due to their powerful nonlinear modeling capability, machine learning methods are proving to be very helpful in predicting blooms and in understanding how various factors may modulate bloom strength. The ANN model approach was used to predict algal blooms in Hong Kong coastal waters (Lee et al., 2003) and to predict outbreaks of the dinoflagellate *Dinophysis acuminata* in southern Spain (Velo-Suarez and Gutierrez-Estrada, 2007). More recently, a neural network approach was used to predict presence/absence and abundance of the dinoflagellate *Karlodinium* and the diatom *Pseudo-nitzschia* in Alfacs Bay in the northwest Mediterranean Sea (Guallar et al., 2016), and SVM models were used to predict blooms in freshwater reservoirs (Xie et al., 2012).

In this research, we use machine learning algorithms to predict *K. brevis* on the West Florida Shelf over a twenty-year period using discharge, nutrient, weather, and sea surface data. Specifically, we examined if we could assess (1) how wind direction and strength affect the frequency of *K. brevis* blooms on the West Florida Shelf, and (2) how discharge from different rivers, with differing nutrient loads, fuels *K. brevis*.

**2. Methods**

***2.1 The data set and preparation of explanatory and dependent variables***

To develop the machine learning models, *in-situ* data of *K. brevis* cell densities (number of cells per L) over a twenty-year period (1998-2018) on the WFS were obtained from the Florida Fish and Wildlife Conservation Commission (<https://myfwc.com/>). These data represent water samples collected during regular monitoring along the Florida coast and during suspected or confirmed *K. brevis* events. The data used herein were limited to samples collected between latitudes of 25.85 degrees (Marco Island) and 29.14 degrees (Mouth of Suwanee River) and at most 9 km from the coast since most of the *K. brevis* blooms occurred within this area on WFS and setting a fixed area for the data analysis ensued consistency. In order to overcome the spatial and temporal inconsistency in the data, the 5 highest cell counts across the spatial gradient were averaged for each week to produce a weekly mean. Since the *K. Brevis* measurements were not made on a continuous basis and likely under-sampled under low cell density conditions, we discretized weekly averages into a binary variable: with cell counts greater than 105 cells/L counting as *K. brevis* events.

Streamflow data were obtained from United States Geological Survey (USGS) stations in the major rivers that discharge onto the West Florida Shelf (<https://waterdata.usgs.gov/nwis>**).** 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), as indicated on Fig.1. Nutrient data from the major rivers were downloaded from the Tampa Bay and Charlotte Harbor Water Atlas ([http://www.wateratlas.usf.edu/](https://waterdata.usgs.gov/nwis;http:/www.wateratlas.usf.edu/)). Weekly averaged nutrient concentrations were multiplied by weekly averaged streamflow to estimate total nitrogen (TN) and total phosphorus (TP) loads.

Hourly wind and temperature data were obtained from the National Data Buoy Center (NDBC) stations (<https://www.ndbc.noaa.gov/>;Fig. 1) across the WFS. The hourly (?) wind speeds were used to calculate weekly averages using a simple vector average. Satellite altimetry, obtained from the E.U. Copernicus Marine Service Information (<http://marine.copernicus.eu/>), 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. (2015).

Data were aggregated into the following form; each row of the dataset is , where are the explanatory variables of river discharge, nutrient concentration, wind speed and direction, temperature, and sea surface height difference, and is the dependent variable of discretized *K. brevis* cell densities. Machine learning algorithms aim to map to. Open source R packages were used (Stone, 1974; Geisser, 1975; Burman et al., 1994; Cawley and Talbot, 2004; Anguita et al., 2009; Bergmeir and Benitz, 2012; R Core Team, 2017).

The number of HAB events (318) was significantly smaller than the number of events without HABs (755). This created an imbalanced classification problem (Sun et al., 2009). Several data-level approaches have been developed to address this issue. We explored two approaches: (1) the minority class of the training data was randomly oversampled such that the sample size of events with and without HABs are roughly equal in the synthetic training dataset (Japkowize and Holte, 2000); (2) the minority class was oversampled by generating new synthetic data using SMOTE preprocessing algorithm (Chawla et al., 2002; Fernandez et al., 2018).

***2.2 Machine Learning Algorithms***

To predict *K. brevis* cell density and test the strength of various explanatory variables, four machine learning algorithms were used: a) Support Vector Machine (SVM), b) Relevance Vector Machine (RVM); c) Naive Bayes (NB), and d) Artificial Neural Network (ANN). They represent a range of machine learning algorithms with different methodology and varying complexity. ANN is based on a system of connected nodes to simulate neurons in a biological brain. SVM belongs to a class of algorithms called kernel methods. NB classifiers are a family of simple probabilistic classifiers. RVM has an identical function to SVM but provides probabilistic classification.

*a. Support Vector Machine*

The Support Vector Machine (SVM) model is a supervised machine learning algorithm that seeks the hyperplane that best separates two labeled classes from each other (Figure 2a):

 (1)

where  represents the vector for the explanatory variables,  is the hyperplane that separates the two classes,  is the slope of the hyperplane, *b* is the intercept, and f()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 or minimizing the cost function CF

 (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 (Cortes and Vapnik, 1995) according to Eq. 3 as follows,

 (3)

where the slack variables , is either 1 or -1, and 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 (Eq. 4):

 (4)

where are the Lagrangian multipliers and is rewritten as a linear combination of the training patterns (Nello and John, 2000). The constant *b* can be found with the Karush-Kuhn-Tucker Conditions (Basak et al., 2007; Vapnik, 1995).

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 (Boser et al., 1992; Scholkopf and Smola, 2002). The Gaussian radial basis function

was chosen as the kernel function because of its computational efficiency:

(5)

where is the kernel parameter controlling the sensitivity of the kernel function.

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

*b. Relevance Vector Machine*

Relevance Vector Machine (RVM) has an identical functional form to SVM but uses Bayesian probabilistic framework to estimate the parameters (Tipping 2001; Camps-Valls et al. 2006). To obtain maximum likelihood estimate of and and avoid overfitting, the Baysian 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 of N+1 hyperparameters to enforce a preference for smoother. Then the posterior estimate of the unknown parameters given the data is obtained using Bayes’ rule. Since the posterior probability can be evaluated exactly, RVM seeks to maximize the marginal likelihood with respect to the hyperparameters and . RVM typically uses much fewer basis functions than SVM. RVM was applied herein using the radial basis function as the kernel function.

*c. Naïve Bayes*

The Na**ï**ve Bayes (NB) classifier is a simple probabilistic classifier based on the Bayes’ Rule and requires strong “naïve” independence between the features (Maron 1961, Hand and Yu 2001). Given a new observation **x**, it finds the class that maximizes the conditional probability , the likelihood of a class given the observation. Uses Bayes’ theorem, the conditional probability can be calculated using

(6)

where is the prior probability of observing a class , is the likelihood of observing **x** given , and is the probability of observing **x**.

ssuming strong na**ï**ve independence, the probabilistic chain rule can be used to transform the likelihood of **x** into the probabilities of each of the features of **x** given a class:

(7)

This study used the Gaussian NB in which the Gaussian distribution

(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.

*d. Artificial Neural Network*

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 5 (Hassoun, 1995; Schmidhuber, 2015). ANN can be turned into a classifier by discretizing the network’s output. The basic substructure of ANN is perceptron (Fig. 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 for node j in layer l is shown below (Eq. 9):

 (9)

where **x** are the input variables,  is the output at layer (l-1), wji are the synaptic weights, and  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 neural network: 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 (Werbos, 1990). 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 reaches below a desired threshold. Herein, two hidden layers with 20 and 10 neurons were used in the ANN model.

***2.3 Model Evaluation***

The predictive skill of the machine learning algorithms was evaluated using the *k*-fold cross-validation approach that has been widely used in machine learning classification problems (Anguita et al., 2009; Cawley and Talbot, 2004). In this approach, the data are randomly subdivided into *k* disjointed subsets of equal size. Then, for each different 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 (Stone, 1974; Geisser, 1975). The mean of the test statistics over all *k* models is called the cross-validation estimate of the test statistic. We chose *k*=10 in our study such that each subset contained 2-years of data.

The data are assumed to be independent during the k-fold cross-validation. This assumption might be inappropriate for the time series which might be linearly related to a lagged version of itself. Thus, the data herein were further validated by block cross-validation (Bergmeir and Benitex, 2012; Roberts 2017; Burman et al., 1994; Racine 2000). The data were divided by chronological order into 10 subsets of 2-years each: 1998-1999, 2000-2001… 2017-2018 (Bergmeir and Benitex, 2012; Roberts 2017). 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* bloom (Sun et al., 2009; Powers, 2011). The accuracy measures the overall accuracy of the prediction:

(10)

where TP is the number of HABs weeks predicted correctly, FN is the number of HABs weeks predicted to be non-HABs, TN is the number of non-HABs weeks predicted correctly, and FP is the number of non-HABs weeks predicted to be HABs. The recall is the ratio of the correctly-predicted HABs weeks to the total number of the observed HABs weeks:

. (11)

The precision is the ratio of the correctly-predicted HABs weeks to the total number of the predicted HABs weeks:

. (12)

The F1 measures the balance between precision and recall:

. (13)

The testing metrics were averaged over all the ten models. To further test the models’ predictions, a time series of the cross-validation predictions was created.

***2.4 Sensitivity Analysis***

Machine learning classifiers were used to determine the factors that affect *K. brevis* blooms and each of their significance. To do this, we applied the Platt scaling (e.g., Platt 1999) which uses a logistic transformation to convert classifier predictions into probability distributions over the classes. First, the machine learning algorithm was trained on the entire dataset. Platt scaling was then used calculate the probability of *K. brevis* bloom (Eq. 14):

(14)

where is a sample, is one of the classes, is the classifier output, and A and B are scalar constants (Lin et al., 2007). Each explanatory variable was varied by 1-2 standard deviations around its mean while the other variables were set to their respective annual mean values. Line plots and contour diagrams of HABs probability as a function of explanatory variables were created by varying one or two explanatory variables at a time. Previous studies suggested that wind-driven upwelling advects *K. brevis* onshore and riverine nutrient loading fuels the *K. brevis* blooms. To quantify these effects, we calculated *K. brevis* probability as a function of wind speed and direction, riverine discharge, and nutrient loading.

**4. Results**

***4.1 Overall model performance***

The four different machine learning approaches, SVM, RVM, NB and ANN were applied and their predictability of the 20-year time series (1998-2018) were tested relative to the observed *K. brevis* cell concentrations along the West Florida Shelf. Using the blocked cross validation procedure, the SVM approach performed the best (Fig. 3a; Table 1). It was 38% accurate in predicting weeks with blooms, 76% accurate for weeks without blooms, and 62% accurate overall. The RVM had a near-identical accuracy overall, 61% (60%, 60% and 61%, respectively). The NB approach had the second-highest accuracy (61%; 47%; 52%, respectively), and the ANN approach performed the weakest of the 3 models, but nevertheless still performed reasonably well in predicting bloom occurrences (29%; 74%; 60%, respectively). The comparison of the time series with the Relevance Vector Machine (RVM) illustrates that both prolonged blooms with high cell counts–and periods of only a short duration with relatively low cell counts–are captured well by the model (Fig. 3a). Given the irregularity of blooms both temporally and spatially (Fig. 3b), and associated sampling that is dictated by the events and not by prescribed times and stations, the model is robust in capturing such a diverse range of conditions.

***4.2 Role of wind speed***

Having established that SVM was the most robust model approach, it was used to examine 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). Bloom probability was much higher under northerly winds than under southerly winds (Fig. 4a). Bloom probability reached a maximum of 38% under northerly wind, while strong southerly wind reduced bloom probability to <20%. Northerly winds drive coastal upwelling, thereby transporting *K. brevis* from the offshore waters to coastal waters. Additionally, coastal upwelling delivers inorganic nutrients from the ocean that can help fuel the blooms. Strong, compared to weak, westerly winds increased bloom probability by 10%, from a low of 35% to a high of 45% (Fig. 4b). 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.

***4.3 Role of river flow and associated nutrients***

The probability of *K. brevis* outbreaks as a function of discharge from the Suwanee, Hillsborough, Myakka, Peace and Caloosahatchee Rivers, all of which discharge into the Western Florida Shelf, was analyzed using SVM (Figs. 1, 5a-e). For the Caloosahatchee River across all discharge levels, the probability of *K. brevis* blooms was consistently high (39-42%) and increased linearly as river discharge increased. The Caloosahatchee River had the highest discharge of the rivers examined, and it transported the highest amount of nutrients. The slope in bloom probability with change in discharge was highest with the Hillsborough River, with low discharge yielding a 10% probability in blooms, increasing to 50% with high discharge. Increases in discharge from the Peace and Suwanee Rivers also increased bloom probability substantially, from 23-42% and 17-41% respectively, across the range of typical flows. Changes in discharge from the Myakka River yielded probabilities that changed from 27-39%. In addition, the shape of the relationship varied among the rivers. For the Peace and Myakka Rivers, the *K. brevis* blooms and discharge are tightly coupled only at low discharge rates. For the Suwanee and Hillsborough river, *K. brevis* probability as a function of the riverine discharge resembles a sigmoidal distribution. High discharge rates likely provide nutrient amounts that exceed the nutrient demand of *K. brevis* and thus further increases have little effect.

The composition of the nutrients discharged by the different rivers also varied. The probability of blooms for each river was calculated as a function of their TN and TP loads, and these increases had varied effects (Fig. 5 f-m). With increasing TN, the largest increases in bloom probability were found for the Hillsborough and Peace Rivers, whereas for the Myakka River, no significant increase in probability was seen as TN increased, and probability decreased in the Caloosahatchee River. For TP, however, increases in probability were seen for the Hillsborough and Caloosahatchee Rivers, but a parabolic relationship was noted for the Myakka and Peace Rivers. 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.

By comparing TN and TP discharge from different rivers, it can be seen that large reductions in both nutrients are needed to have a substantial impact on reducing the frequency of *K. brevis* blooms (Fig. 6). 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 >60% to <20%.

**5 Discussion**

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 is thought to have caused the deaths of hundreds of tons of fish, hundreds of manatees, dolphin, and sea turtles, as well as many reports of hospitalization visits due to respiratory distress (e.g., Munoz, 2019). Fisheries closures, as well as revenue lost by local businesses, also had massive economic impacts (Fears and Rozsa, 2018). 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 heath as well as human health.

Herein, the overall performance of the machine learning algorithms was not significantly affected by the sea level height difference that was used to represent the effective Loop Current. However, for 1998, 2002, 2009, 2010, and 2013, the SVM ? model had a much higher false positive rate (38.1%, 51.4%, 22.0%, 51.9%, 55.2%) versus 37.8% for all years. This suggests other factors not considered in the explanatory variables may be needed to improve bloom prediction for those years.

There are several differences between the methodology applied herein and the Hill et al. (2019) analysis. First, different explanatory variables were used. The Hill et al. (2019) study used satellite remote sensing chlorophyll as a proxy for *K. brevis*, whereas direct cell counts were used here. Second, they did not consider wind speed, river flow or nutrient loads. These approaches are all complementary and show the promise of machine learning approaches not only in modeling various aspects of *K. brevis* blooms, but HAB events more generally.

Although there have been debates about the extent to which anthropogenic nutrients fuel *K. brevis* blooms (e.g., Brand and Compton, 2007; Heil et al., 2014 and references therein), there is no doubt that Florida’s continuing population growth has accelerated eutrophication. The nutritional pathways and sources of nutrients supporting *K. brevis* blooms are complex (e.g., Vargo et al., 2008; Glibert et al., 2009; Heil et al., 2014; O’Neil and Heil, 2014), the fact that nutrient loads have increased is, in itself, an insufficient explanation for the expansion in *K. brevis* blooms. It takes the right nutrients at the right time to create conditions conducive for these blooms to form (Glibert and Burford, 2017). 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. (2015) indicates that there are significant differences Peace and Caloosahatchee River flows between periods of large blooms and periods without blooms. The SVM machine learning algorithm used here illustrated strong relationships between river flow and blooms.

Florida, among many states and environmental protection agencies around the world, has established, or is working to establish, nutrient reduction targets to mitigate water quality problems in their water bodies (Zhao et al., 2016; Herrero et al., 2019). 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. Since 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.

Air temperature over the Eastern North America (including Florida) is expected to increase ~1.5 oC by 2050 and 3-4 oC by 2100 (relative to 2000), according to recent climate projections (IPCC, 2014). Additionally, rainfall over Florida is projected to decrease by 20-30% during the summer but will 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. With climate change and the predicted increase in extreme precipitation events in a warming climate (Sillman et al., 2013a,b; Russo et al., 2014), it is expected that will likely be more frequent HABs in the future, in Florida and elsewhere, unless substantial reductions in TN and TP land-based use and loading in the major rivers is accomplished.

**Acknowledgements**

PMG received support from the National Oceanic and Atmospheric Administration National Centers for Coastal Ocean Science Competitive Research program under awards No. NA17NOS4780180 and NA19NOS4780183. We thank T. Kana for helpful comments on this manuscript. This is contribution number xxxx of the University of Maryland Center for Environmental Science and contribution number ECOyyy from the NOAA ECOHAB program.

**References**

Anderson, D. M. 1989. Toxic algal blooms and red tides: A global perspective. *In:* Okaichi, T., Anderson, D., Nemoto, T (eds.). Red Tides: Biology, Environmental Science, and Toxicology. (Elsevier Science Publishing Company, New York) pp. 11-16.

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: Kim, H.-G., Reguera, B., Hallegraeff, G.M., et al. (Eds), Harmful Algae 2012: Proceedings of the 15th International Conference on Harmful Algae: October 29 - November 2, 2012 (CECO, Changwon, Gyeongnam, 2014) pp. 3–17.

Anderson, D.A., Glibert, P.M., Burkholder, J.M. 2002. Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. Estuaries 25, 562-584.

Anguita, D., Ghio, A., Ridella, S., Sterpi, D. 2009. K-Fold cross validation for error rate estimate in support vector machines. In: Proceedings of The 2009 International Conference on Data Mining, DMIN 2009, July 13-16, 2009, Las Vegas, USA, pp. 1-7

Bergmeir, C., Benitez, J. M. 2012. On the use of cross-validation for time series predictor evaluation. Inform. Sci*.* 191, 192-213. Doi:10.1016/j.ins.2011.12.028.

Boser, B., Guyon, I., Vapnik, V.1992. A training algorithm for optimal margin classifiers. In: COLT ’92 Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, pp. 144-152.

Brand, K., Compton, A. 2007. Long-term increase in *Karenia brevis* abundance along the southwest Florida coast. Harmful Algae 6, 232-252. Doi: 10.1016/j.hal.2006.08.005.

Burman, P.R., Chow, E., Nolan, D. 1994. A cross-validatory method for dependent data. Biometrika **81**, 351-358. Doi: 10.2307/2336965.

Calaway, R., Microsoft Corporation, Weston, S., Tenenbaum, D. 2017. doParallel: Foreach parallel adaptor for the 'parallel' Package. R package version 1.0.11. <https://CRAN.R-project.org/package=doParallel>.

Cawley, G. C., Tablot, N. L.C. 2004. Fast exact leave-one-out cross validation of sparse least-squared support vector machines. Neural Networks 17, 1467-1475. Doi: 10.1016/j.neunet.2004.07.002.

Fears, D., Rozsa, L. Aug. 28, 2018. Florida’s unusually long red tide is killing wildlife, tourism and businesses. The Washington Post. 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

Fernandez, A., Garcia, S., Herrera, F., Chawla, N. V. 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Art. Intel. Res. 61, 863–905. Doi: 10.1613/jair.1.11192.

Flynn, K.J., McGillicuddy, D.J. 2018. Modeling marine harmful algal blooms: current status and future prospects. In: Shumway, S.E., J.M. Burkholder, S.L. Morton (Eds), Harmful algal blooms: A compendium desk reference (Wiley Blackwell, Noida, India), pp. 115-134.

Franks, P.J.S. 2018. Recent advances in modeling of harmful algal blooms. In: Glibert P.M., Berdalet, E., Burford, M. Pitcher, G. and Zhou, M.J. (eds.), Global Ecology and Oceanography of Harmful Algal Blooms (Springer, Cham, Switzerland), pp. 359-380.

Fu, F.X., Tatters, A.O., Hutchins, D.A. 2012. Global change and the future of harmful algal blooms in the ocean. Mar. Ecol. Progr. Ser. 470, 207-233. Doi: 10.3354/meps10047.

Geisser, S. 1975. The predictive sample reuse method with applications. J. Amer. Stat. Assoc. 70, 320-328. Doi: 10.2307/2285815.

Glibert P.M*.* 2019a. Harmful algal at the complex nexus of eutrophication and climate change. Harmful Algae*.* Doi:10.1016/j.hal.2019.03.001.

Glibert, P.M. 2019b. Why were the water and beaches in west Florida so gross in summer 2018? Red tides! Front. Young Minds. Doi: 10.3389/frym.2019.00010.

Glibert, P.M., Burford, M.A. 2017. Globally changing nutrient loads and harmful algal blooms: Recent advances, new paradigms and continuing challenges. Oceanography 30(1), 44-55. Doi: 10.5670/oceanog.2017.110.

Glibert, P.M. and J.M. Burkholder. 2018. Causes of harmful algal blooms. In: Shumway, S., J.M. Burkholder and S.L. Morton (eds.), Harmful Algal Blooms: A Compendium Desk Reference. (Wiley Blackwell, Singapore), pp. 1-38.

Glibert, P.M., Burkholder, J.M., Kana, T.M., Alexander, J.A., Schiller, C., Skelton, H. 2009. Grazing by *Karenia brevis* on *Synechococcus* enhances their growth rate and may help to sustain blooms. Aquat. Microb. Ecol.55**,** 17-30. Doi: 10.3354/ame1279.

Glibert, P.M., Allen, J.I., Bouwman, L., Brown, C., Flynn, K.J., Lewitus, A., Madden, C. 2010. Modeling of HABs and eutrophication: status, advances, challenges. J. Mar. Syst. 83, 262–275. Doi: 10.1016/jmarsys.2010.05.004.

Guallar, C., Delgado, M., Diogène, J., Fernández-Tejedo, M. 2016. Artificial neural network approach to population dynamics of harmful algal blooms in Alfacs Bay (NW Mediterranean): Case studies of *Karlodinium* and *Pseudo-nitzschia*. Ecol. Mod. 338, 37-50.

Hallegraeff, G.M. 1993. A review of harmful algal blooms and their apparent global increase. Phycologia 32, 79-99.

Hand, D.J., Yu, K. 2001. Idiots Bayes—not so stupid after all?. Int. Stat. Rev. 69, 385–398.

Heil, C.A., Bronk, D. A., Dixon, L. K., Hitchcock, G. L., Kirkpatrick, G. J., et al. 2014. The Gulf of Mexico ECOHAB: *Karenia* program 2006–2012. Harmful Algae 38, 3-7. Doi: 10.1016/j.hal.2014.07.015.

Heisler, J., Glibert, P.M., Burkholder, J., Anderson, D., Cochlan, W., Dennison, W., Dortch, Q. et al*.* 2008. Eutrophication and harmful algal blooms: A scientific consensus. Harmful Algae 8, 3-13. Doi: 10.1016/j.hal.2008.08.006.

Herrero, F.S., Teixeira, H., Poikane, S. 2019. A novel approach for deriving nutrient criteria to support good ecological status: Application to coastal and transitional waters and indications for use. Front. Mar. Sci. Doi: 10.3389/fmars.2019.00255.

Hijmans, R. 2017. raster: Geographic data analysis and modeling. R package version 2.6-7. <https://CRAN.R-project.org/package=raster>.

Hill, P.R., Kumar, A., Temini, M., Bull, D.R. 2019. HABNet: Machine learning, remote sensing based detection and prediction of harmful algal blooms. IEEE J Selected Topics Appl. Earth Observ. Rem. Sens. arXiv:1912.02305.

Hu, C., Muller-Karger, F.E., Swarzenski, P.W. 2006. Hurricanes, submarine groundwater discharge, and Florida’s red tides. Geophys. Res. Lett. 33, L11601. Doi: 10.1029/2005GL0254449.

IPCC, Summary for policymakers, in Climate Change. 2014*:* Impacts, adaptation, and vulnerability. Part A: global and sectoral aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. C. B. Field, et al. (Eds.) (Cambridge Univ. Press Cambridge, United Kingdom and New York, NY, USA), pp. 1-32.

Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A. 2004. Kernlab - An S4 package for kernel methods in R. J. Stat. Software 11(9), 1-20. <http://www.jstatsoft.org/v11/i09/>.

Lee, J.H.W., Y. Huang, M. Dickman, Jayawardena, A.W. 2003. Neural network modelling of coastal algal blooms. Ecol. Model. 159, 179-201. Doi: 10.1016/S0304-3800(02)00281-8.

Lenes, J.M., Darrow, B.A., Walsh,J .J., Prospero, J.M., He, R., Weisberg, R.H., et al. 2008. 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. 28, 1091-1115. Doi : 10.1016/j.csr.2008.02.009.

Liu, Y.,Weisberg, R.H. 2012. Seasonal variability on the West Florida Shelf. Progr. Oceanogr. 104, 80-98. Doi: 10.1016/j.pocean.2012.06.001.

Liu, Y., Weisberg, R.H., Lenes, J.M., Zheng, L. et al. 2016. [Offshore forcing on the "pressure point" of the West Florida Shelf: Anomalous upwelling and its influence on harmful algal blooms](http://dx.doi.org/10.1002/2016JC011938), J. Geophys. Res.121, 5501-5515. Doi: 10.1002/2016JC011938.

Maron, M.E. 1961. Automatic indexing: an experimental inquiry. J. Assoc. Comp. Mach. 8, 404-417. Doi: 10.11145/321075.321084.

Mayer, D.A., Weisberg, R.H., Zheng, L, Liu, Y. 2017. [Winds on the West Florida Shelf: Regional comparisons between observations and model estimates](http://dx.doi.org/10.1002/2016JC012112). J. Geophys. Res. Oceans 122, 834-846. Doi: 10.1002/2016JC012112.

Maze, G., Olascoaga, M.J., Brand, L. 2015. Historical analysis of environmental conditions during Florida red tide. Harmful Algae 50, 1-7. Doi: 10.1016/j.hal.2015.10.003.

McGillicuddy, D.J., Jr., de Young, B., Doney, S., Glibert, P.M., Stammer, D., Werner, F.E. 2010. Models: Tools for synthesis in international oceanographic research programs. Oceanography 23, 126-139. Doi: 10.5670/oceanog.2010.28.

Monuz, C.R. Jan 15, 2019. Red tide episode kills record number of sea turtles. Herald Tribune. https://www.heraldtribune.com/news/20190115/red-tide-episode-kills-record-number-of-sea-turtles

O’Neil, J.M., Heil, C.A. 2014. Preface to ECOHAB: *Karenia* Special Edition of Harmful Algae. Harmful Algae 38, 1-2.

Pebesma, E., Bivand, R. 2005. Classes and methods for spatial data in R. RNews 5 (2), <https://cran.r-project.org/doc/Rnews/>.

Platt, J. C. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, A. et al. (ed.), Advances in Large Margin Classifiers. (MIT Press, Cambridge MA), pp. 61-74.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Racine, J. 2000. Consistent cross-validatory model-selection for dependent data: *hv*-block cross-validation. J. Economet. 99, 39-61. Doi: 10.1016/s0304-4076(00)00030-0.

Roberts, D. R. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography40, 913-929. Doi: 10.1111/ecog.02881.

Russo, S., Dosio, A., Graversen, R.G., Sillmann, J., Carrao, H., Dunbar, M.B. et al. 2014. Magnitude of extreme heat waves in present climate and their projection in a warming world. J. Geophys. Res. Atmos. 119, 12,500–12,512. Doi:10.1002/2014JD022098.

Schnute, J., Boers, M., Haigh, R. 2017. PBSmapping: Mapping fisheries data and spatial analysis tools. R package version 2.70.4. <https://CRAN.R-project.org/package=PBSmapping>.

Sillmann, J., Kharin, V.V., Zhang, X., Zwiers, F.W., Bronaugh, D. 2013a. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. J. Geophys. Res. Atmos. 118, 1716–1733. Doi:10.1002/jgrd.50203.

Sillmann, J., Kharin, V.V., Zwiers, F.W., Zhang, X., Bronaugh, D. 2013b. Climate extremes indices in the CMIP5 multimodle ensemble: Part 2. Future climate projections. J. Geophys. Res. Atmos.118, 2473-2493.

Steidinger, K.A. 2009. Historical perspective on *Karenia brevis* red tide research in the Gulf of Mexico. Harmful Algae8, 549-561. Doi: 10.101/j.hal.2008.11.009.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. J. Roy. Stat. Soc. Series B (Methodological) 36, 111–133. Doi: 10.1111/j.2517-6161.1974.tb00994.x.

Stump R.P., Culver, M.E., Tester, P.A., Tomlinson, M., Kirkpatrick, G.J. et al. 2003. Monitoring *Karenia brevis* blooms in the Gulf of Mexico using satellite ocean color imagery and other data. Harmful Algae 2, 147-160. Doi: 10.1016/S1568-9883(02)00083-5.

Stump, R.P., Tomlinson, M.C., Calkins, J.A., Kirkpatrick, B., Fisher, K. et al. 2009. Skill assessment for an operational algal bloom forecast system. J. Mar. Syst. 76(1-2), 151-161. Doi: 10.1016/j.marsys.2008.05.016.

Tipping, M.E. 2001. Sparse Bayesian Learning and the Relevance Vector Machine. J. Mach. Learn. Res. 1, 211-244. Doi: 10.1162/15324430152748236

Vapnik, V. 1995. The Nature of Statistical Learning Theory. Springer NY.

Vargo, G.A. 2009. 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 8, 573-584. Doi: 10.1016/j.hal.2008.11.002.

Vargo, G.A. Heil, C.A., Fanning,K.A., Dixon, K. L., Neely, M.B., Lester, K., A. et al. 2008. Nutrient availability in support of *Karenia brevis* blooms on the central West Florida Shelf: what keeps *Karenia* blooming?  Cont. Shelf Res. 28, 73-98. Doi: 10.1016/j.csr.2007.04.008.

Velo-Suarez, L., Gutierrez-Estrada, J.C. 2007. Artificial neural network approaches to one-step weekly prediction of *Dinophysis acuminata* blooms in Huelva (Western Andalucıa, Spain). Harmful Algae 6, 361-371. Doi: 10.1016/j.hal.2006.11.002.

Walsh, J.J., Weisberg, R.H., Dieterle, D.A., He, R., Darrow, B.P., Jolliff, J.K., et al. 2003. The phytoplankton response to intrusions of slope water on the West Florida Shelf: models and observations. J. Geophys. Res*.* 108, C6, 15. Doi: 10.1029/2002JC001406.

Weisberg, L. Zheng, L., Liu, Y., Lembke, C., Lenes, J.M.,Walsh, J.J., 2014. Why a red tide was not observed on the west Florida continental shelf in 2010. Harmful Algae 38, 119-126. Doi: 10.1016/j.hal.2014.04.010

Weisberg, R.H., He, R. 2003. Local and deep-ocean forcing contributions to anomalous water properties on the West Florida Shelf. J. Geophys. Res. 108(C6) 3184. Doi: 10.1029/2002JC001407.

Wells, M.L., Trainer V.L., Smayda, T.J., Karlson, B.S., Trick, C.G. et al. 2015. Harmful algal blooms and climate change: learning from the past and present to forecast the future. Harmful Algae 49, 68-93. Doi: 10.1016/j.hal.2015.07.009.

Xie, Z., Lou, I., Ung, W.K, Mok, K.M. 2012. Freshwater algal bloom prediction by support vector machine in Macau storage reservoirs. Math. Prob. Eng. Doi: 10.1155/2012/397473.

Zhao, X., Wang, H., Tang, Z., Qin, N., Li, H., Wu, F., Giesy, J.P. 2016. Amendment of water quality standards in China: viewpoint on strategic considerations. Envir. Sci. Pollut. Res. Int. Doi: 10.1007/s11356-016-7357-y.

**Author contributions**

MJL developed the models and wrote the paper. PMG advised the project and edited the manuscript.

**Competing Interests**

The authors have no competing interests.

**Data and code availability**

All the data and code are publicly available and accessible online. The data and code can be found at: <https://github.com/lim09749/WFS_ML/>.

**Table 1. Accuracy of the three machine learning approaches, as well as the RVM modification of SVM applied herein, as validated using** *k***-fold cross validation and block cross-validation.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | *K***-fold cross validation** | **Block cross-validation** |
| **SVM** | **HAB accuracy** | **0.63** | **0.38** |
| **Non-HAB acc.** | **0.85** | **0.76** |
| **Total accuracy** | **0.78** | **0.62** |
| **RVM** | **HAB accuracy** | **0.54** | **0.60** |
|  | **Non-HAB acc.** | **0.71** | **0.60** |
|  | **Total accuracy** | **0.66** | **0.61** |
| **Naïve Bayes (NB)** | **HAB accuracy** | **0.65** | **0.61** |
| **Non-HAB acc.** | **0.56** | **0.47** |
| **Total accuracy** | **0.59** | **0.52** |
| **ANN** | **HAB accuracy** | **0.42** | **0.29** |
| **Non-HAB acc.** | **0.76** | **0.74** |
| **Total accuracy** | **0.65** | **0.60** |

**Figure Legends**

**Fig. 1.** Map of Florida showing the region in red where *Karenia brevis* blooms were most intense in 2018-2019, and the rivers discharging into West Florida Shelf considered herein. The National Data Buoy Center stations from which wind and temperature data were acquired are also shown.

**Fig. 2.** Schematic diagram of *Karenia brevis* blooms on the West Florida Shelf. Northerly winds drive coastal upwelling that transports cells from offshore to coastal regions. Nutrients from agriculture and urban areas are carried by rivers and help fuel blooms. This nutrient delivery is accelerated by heavy precipitation events such as those from hurricanes.

**Fig. 3.** Comparison of Support Vector Machine output and observational data of *Karenia brevis.* (a) Time series of the observed (black line) and predicted (black dots) area-averaged *K. brevis* concentrations from 1998-2018. (b) Snapshots of the observed *K. brevis* distribution in selected months. The twenty-year timespan includes many years with blooms (2002, 2005, 2012, 2018) and without blooms (1998, 2010).

**Fig. 4.** Probability of *Karenia. brevis* as a function of wind speed and direction. 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, westerly wind keeps populations near the coast and prevents them from dispersing offshore.

**Fig. 5.** Probability of *Karenia brevis* as a function of riverine discharge (panels a-e), total nitrogen loading (TN; panels f-i) and total phosphorous loading (TP; panels j-m).

**Fig. 6.** Contour plots of *K. brevis* probability as a function of (a) Hillsborough and Peace River TN concentrations and of (b) Hillsborough and Caloosahatchee TP concentrations.