Science of the Total Environment 912 (2024) 169253

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Predicting coastal harmful algal blooms using integrated data-driven analysis of environmental factors

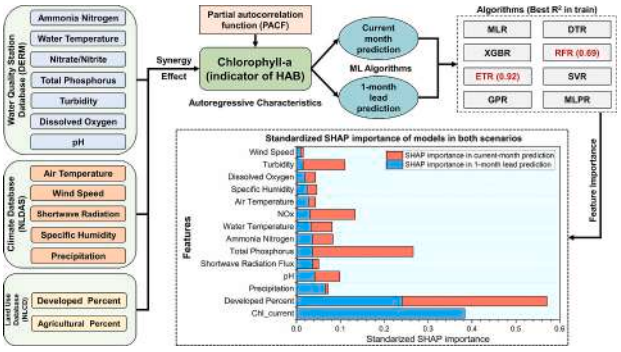
Zhengxiao Yan, Sara Kamanmalek, Nasrin Alamdari *

[a] Department of Civil and Environmental Engineering, FAMU-FSU College of Engineering, Florida State University, Tallahassee, FL 32310, USA

HIGHLIGHTS

- We predicted Chlorophyll-a levels, harmful algal bloom indicator, by ML algorithms.
- We built models by multi-source physical, geochemical, climate, and land use data.
- One-month lead predictions are more accurate than current-month predictions.
- Tree-based machine learning algorithms predicted Chl-a levels more accurately.
- HAB in Biscayne Bay was impacted most by land use of upstream Miami and nutrients.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: José Virgílio Cruz

Keywords:
Harmful algal blooms
HAB prediction
Machine learning
Land use
Nutrients
SHAP value

ABSTRACT

Coastal harmful algal blooms (HABs) have become one of the challenging environmental problems in the world's thriving coastal cities due to the interference of multiple stressors from human activities and climate change. Past HAB predictions primarily relied on single-source data, overlooked upstream land use, and typically used a single prediction algorithm. To address these limitations, this study aims to develop predictive models to establish the relationship between the HAB indicator – chlorophyll-a (Chl-a) and various environmental stressors, under appropriate lagging predictive scenarios. To achieve this, we first applied the partial autocorrelation function (PACF) to Chl-a to precisely identify two prediction scenarios. We then combined multi-source data and several machine learning algorithms to predict harmful algae, using SHapley Additive exPlanations (SHAP) to extract key features influencing output from the prediction models. Our findings reveal an apparent 1-month autoregressive characteristic in Chl-a, leading us to create two scenarios: 1-month lead prediction and current-month prediction. The Extra Tree Regressor (ETR), with an R^2 of 0.92, excelled in 1-month lead predictions, while the Random Forest Regressor (RFR) was most effective for current-month predictions with an R^2 of 0.69. Additionally, we identified current month Chl-a, developed land use, total phosphorus, and nitrogen oxides (NOx) as critical features for accurate predictions. Our predictive framework, which can be applied to coastal regions worldwide, provides decision-makers with crucial tools for effectively predicting and mitigating HAB threats in major coastal cities.

* Corresponding author.
E-mail address: nalamdari@eng.famu.fsu.edu (N. Alamdari).

https://doi.org/10.1016/j.scitotenv.2023.169253
Received 11 October 2023; Received in revised form 20 November 2023; Accepted 7 December 2023
Available online 14 December 2023
0048-9697/Published by Elsevier B.V.

1. Introduction

Large-scale algal blooms have evolved into a significant environmental challenge in many major water systems worldwide over recent decades (Paerl et al., 2014; Wells et al., 2015; Griffith and Gobler, 2020; Xia et al., 2020; Anderson et al., 2021). These coastal algal blooms are influenced by hydrological and geochemical factors, such as flow rate and nutrient concentrations, which are in turn affected by human activities and climate change (Hinder et al., 2012; Zhou et al., 2021; Medina et al., 2022). Certain harmful algae species, primarily diatoms and dinoflagellates, can proliferate rapidly under specific conditions like warm temperatures and excessive nutrients, leading to harmful algal blooms (HABs) (Glibert et al., 2018). HABs in coastal waters have been documented in numerous coastal zones within the United States (U.S.). For example, a coast-wide bloom of the toxic diatom *Pseudo-nitzschia* along the west coast of North America led to the largest recorded neurotoxin outbreak in spring of 2015, resulting in extensive closures of various economic seafood industries for prolonged periods (McCabe et al., 2016). Additionally, blooms of the dinoflagellate *Karenia brevis* are observed almost annually in Florida's Gulf of Mexico (Anderson et al., 2021). These blooms sometimes drift from southwest Florida to its east coast causing fish and shellfish mortality, respiratory irritation in humans, and seawater discoloration (Weisberg et al., 2019). Given the associated risks HABs pose to human health, the economy, and the ecological environment, it is essential to predict HAB occurrences accurately. Identifying key influencing factors is crucial for implementing preventive measures well ahead to mitigate potential losses. (Fleming et al., 2011; Kouakou and Poder, 2019).

The formation of HAB is usually the result of a combination of multi-factors reaching a suitable growth environment (Wells et al., 2015). The factors contributing to the development and persistence of many HABs include eutrophication and deterioration in water quality, especially excess nitrogen and phosphorus, and climate change (Glibert, 2020; Zhou et al., 2022). However, many studies focused on investigating the effects of nutrients and climate drivers, such as temperature and pH, on HABs individually (i.e., examining one factor at a time) or considered only a limited number of factors (Glibert, 2020). For example, Wang et al. (2021) applied Global Nutrient Model to simulate nutrient discharge and predict HAB persistence in Chinese coastal areas but overlooked various climate factors. Using a meta-analysis, Brandenburg et al. (2019) demonstrated that higher temperatures and elevated CO_2 would enhance marine harmful algae growth in temperate regions. However, these experiments typically used control variate approaches to alter a single factor like CO2 or water temperature - an overly simplistic assumption that may not hold true in complex estuarine environments. Vilas et al. (2014) incorporated several physical and climate factors, including temperature, salinity, and upwelling index, but excluded chemical variables. Maze et al. (2015) analyzed a few physical factors like wind speed, flow rate, and loop current without considering the impacts of nutrients or meteorological elements on *Karenia brevis*. These studies did not consider the synergistic effects of multiple interacting factors and or relied solely on single data sources such as onsite data or remote sensing data for predicting HABs, which could lead to increased prediction errors and inapplicability due to the oversimplification of complex environmental interactions (Yajima and Derot, 2017; Hill et al., 2020; Izadi et al., 2021; Wells et al., 2015).

The main prediction methods for HABs include biophysical process-based models and empirical-statistical models. Biophysical models quantify the physical, chemical, and biological processes that drive HABs (Flynn and McGillicuddy, 2018). For example, Walsh et al. (2016) developed a 2D dynamic ecological model including several HAB species with strict parameterization of the boundary conditions. Baek et al. (2021) employed the Environmental Fluid Dynamic Code (EFDC) model to produce ocean properties such as retention time to aid in predicting *Alexandrium catenella* blooms. Despite their potential, biophysical models are not readily applicable for predicting HABs due to the complex nature of these processes, many of which have yet to be accurately identified and characterized. Additionally, these models demand high data accuracy and intricate parameterization alongside specific initial and boundary conditions; therefore their development and application at this stage are relatively challenging and cannot account for every essential feature (Roiha et al., 2010). Biophysical models for predicting HABs have evolved over time by incorporating a wider range of variables, improving parameterization techniques, and leveraging advancements in computational power and technology to better simulate the complex biological, chemical, and physical processes that drive HABs. On the other hand, traditional empirical-statistical models quantify relationships between observations by predicting a variable's value based on measurements of other variables (Franks, 2018). Over time they have evolved with variations such as generalized linear models (GLM), which offer an advancement over traditional empirical-statistical approaches (Franks, 2018). In this case, Singh et al. (2014) applied a logistic GLM to categorize *Dinophysis* concentration in the coastal Arabian Sea as above or below a threshold concentration. Feki et al. (2013) utilized a mixed-effect GLM to characterize the *Karenia selliformis* blooms in the southwestern Mediterranean Sea. However, a significant limitation of these traditional empirical-statistical models is their lack of time dependence. While they demonstrate relationships between variables, they often fail to capture the underlying dynamics of the system (Flynn and McGillicuddy, 2018). Since these models were primarily designed to infer relationships between variables rather than making predictions directly, their accuracy varies, and sometimes their results contradict theoretical expectations (Díaz et al., 2016). Recently, new empirical-statistical models, particularly machine learning models, have emerged as powerful tools for predicting HABs.

HAB prediction based on traditional statistical or biophysical models has not yet yielded convincing conclusions (Xia et al., 2020). As a result, data-driven machine learning models have emerged as ideal alternatives and have recently gained popularity due to their significant advantages such as the ability to capture temporal dynamics, handle complex interactions among variables effectively, process large datasets, and deal with non-linear relationships between predictors (Bergen et al., 2019). Numerous researchers have utilized machine learning algorithms for regression problems such as predicting HAB time series in various coastal regions including Tolo Habour (Muttil and Chau, 2006; Deng et al., 2021), Galician coast (Vilas et al., 2014), Calabash Bay (Coad et al., 2014), Gulf of Mexico (Gokaraju et al., 2011; Li et al., 2021), Genoa area (Asnaghi et al., 2017), East China Sea (Xu et al., 2014). These models, using data-driven algorithms, establish 'black-box' models with high accuracy, eliminating the need for representing unknown physical, chemical, and biological processes mathematically (Lary et al., 2016). However, it is crucial to acknowledge the inherent challenges with machine learning models. They can be prone to overfitting if not properly tuned, which may lead to poor generalization of new data. In addition, the quality of training data is vital; noisy or incomplete data can introduce biases or inaccuracies, undermining model performance. Furthermore, selecting an optimal algorithm from a wide array of choices can be complex due to their differing assumptions and strengths. These challenges underline the importance of careful model development and validation in using machine learning for HAB prediction. There are two types of machine learning models for predicting algal blooms: (1) the classification problem of judging whether algal bloom occurs and (2) the regression problem of predicting Chl-a or cell concentrations. In general, classification problems are generally straightforward to solve but require the appropriate assignment of a "bloom or no bloom" threshold (Izadi et al., 2021; Valbi et al., 2019). On the other hand, while regression models are more complex, they provide more practical numerical results as defining thresholds can be challenging, particularly when multiple harmful algae types are present at a specific location (Asnaghi et al., 2017; Yu et al., 2021). Multiple studies, such as Deng et al. (2021) and Li et al. (2014), have employed lagging strategies for predicting HAB time series in advance. In the context of HAB

prediction, lagging strategies refer to the use of past data - or 'lags' - to predict future events. However, few studies have provided comprehensive explanations for the choice of time lag, potentially diminishing the credibility of setting time lags based solely on experiential grounds without rigorous statistical testing. Despite the effectiveness of machine learning models in prediction, their nonlinear nature often complicates interpretability (Jordan and Mitchell, 2015). Therefore, understanding the factors influencing harmful algae growth remains pivotal for accurate prediction and prevention efforts. Certain applications failed to clarify the significance individual features hold or used absolute mean value feature importance which could potentially mislead decision-makers developing strategies to mitigate or prevent HABs in the absence of knowledge about the "crucial features" (Deng et al., 2021; Xia et al., 2020; Yajima and Derot, 2017). Hence, extracting the degree of positive or negative influences exerted by each factor on HAB growth is imperative, enhancing the practical applicability of predictive models.

Biscayne Bay, our case study site, is located within Miami-Dade County in South Florida, and is essential to the quality of life for Miami-Dade County residents. Biscayne Bay supports the local economy, promotes tourism, and offers countless recreational options. Over the decades, rapid population growth in Miami-Dade County has led to intense urbanization, agricultural activity, and adjustments to water management systems, which are already showing signs of stress due to anthropogenic influences (Millette et al., 2019). A recent study found that chlorophyll-a (Chl-a) concentration, an accepted indicator of eutrophication and HABs, had dramatically increased in the Bay over 20 years, exceeding the Florida Department of Environmental Protection (FDEP) nutrient criteria (Millette et al., 2019; Papenfus et al., 2020). Recent algal blooms in August 2020 led to a significant fish kill event in Biscayne Bay, resulting in thousands of dead fish. The seagrass coverage also was reported to drop considerably due to the excessive growth of algae (Santos et al., 2020). However, a noticeable research gap exists in predicting chlorophyll-a (Chl-a) concentrations in Biscayne Bay, with particularly few studies considering the impacts of upstream land use. This is a significant oversight since it is well-known that land use strongly influences physicochemical and hydrological factors which can exacerbate downstream algal blooms. Urbanization and agriculture can lead to increased nutrient runoff like nitrogen and phosphorus into water bodies like Biscayne Bay, nutrients essential for algal growth, triggering eutrophication leading to HABs. Moreover, alterations due to land use changes can modify hydrological flow patterns, potentially increasing water retention time and providing more opportunities for algae proliferation. While this correlation has been observed and studied extensively within freshwater systems (Kim et al., 2021; Norton et al., 2012), it has been rarely considered in coastal HAB studies. Addressing this research gap would not only contribute valuable insights into how land use changes influence coastal eutrophication patterns but could also inform better resource management strategies aimed at mitigating HABs, ultimately protecting both the ecological health of Biscayne Bay and its value to the local community.

Despite human and ecological health risks associated with HABs, there is a lack of holistic understanding of factors influencing HABs. Previous studies have often overlooked the crucial role of upstream land use in contributing to downstream coastal HABs. Additionally, research on coastal HABs has primarily relied on either on-site water quality data or remote sensing data, with limited integration of multiple data sources. Furthermore, there is a lack of studies that have conducted pre-analysis for autocorrelation and comparative testing of machine learning model performance, specifically regarding time-series data and in-depth feature importance analysis. To address these gaps, we developed a framework to predict coastal HABs considering multiple influencing factors including physiochemical parameters, upstream land use, and meteorological variables. In an innovative approach, our study specifically incorporated upstream Miami-Dade County land use as an essential feature impacting the adjacent downstream HABs in Biscayne Bay. Our framework integrated data from multiple sources including water quality stations, remote sensing assimilation, and upstream land use. This comprehensive data integration enhanced the accuracy of Chl-a predictions in Biscayne Bay. Water quality station data include the features/factors detectable on-site influencing HAB growth, and remote sensing assimilation data can provide additional information about climate, adding depth to the dataset (Wells et al., 2015). To determine the appropriate time lag for our model and avoid relying on uncertain experiences that may introduce errors, we applied a statistically quantified method called partial autocorrelation function (PACF) during the model establishment phase. Furthermore, this study utilized eight machine learning algorithms to identify the best algorithm for the data structure in our case study and explainable tools, SHAP values, to explore the important features influencing HABs. Given the gaps identified in previous research, the objectives of this research are (1) determining the autoregressive characteristics of chlorophyll-a and the proper time lag for prediction; (2) identifying the best performance machine learning algorithm for predicting HAB in Biscayne Bay while considering multiple data sources and land use; and (3) discovering the important features of harmful algae growth and bloom formation. The developed HAB prediction framework with a clear and distinct flow chart can be easily applied to any other coastal area worldwide. We expect that the HAB prediction framework established in Biscayne Bay can help scientists and stakeholders understand the formation mechanisms of HABs and provide an applied reference, and theoretical guidance, or even a prediction standard framework, for the protection of coastal ecological systems around the world.

2. Materials and methods

This study predicts Chl-a concentrations in Biscayne Bay, an accepted indicator of HABs, by employing data-driven methods and integrating multiple data sources. Our approach is based on two prediction scenarios. In doing so, we employed eight machine learning algorithms incorporating 24 years of monthly field observations, climate data, and land use data to predict chlorophyll-a from 1997 to 2020. Since HAB may exhibit autoregressive characteristics, where past values influence future ones, we constructed two distinct prediction scenarios. The first scenario involved predicting Chl-a concentrations for the current month, while the second scenario aimed at predicting Chl-a levels one month in advance. These scenarios were designed after determining the autoregressive properties of HABs. Finally, we analyzed the key drivers of the best-performing models in both scenarios separately.

2.1. Study area, modeling variables, and data collection

Our study focuses on Biscayne Bay, the largest oligotrophic estuary on the southeast Florida coast bordering Miami-Dade County along the Atlantic Ocean, covering about 700 km^2 (see Fig. 1). We divided the northern, central, and southern watersheds based on a combination of sub-watersheds provided by the South Florida Water Management District (SFWMD). Based on available data on Biscayne Bay, we selected hydrogeological and geochemical, climate, and land use drivers as representative features and chlorophyll-a as the target for machine learning modeling (Table 1). Chlorophyll-a (Chl-a) is one of the essential components of algal cells used in oxygenic photosynthesis, which is a predominant surrogate to reflect the algal abundance in HAB investigations (Deng et al., 2021; Ly et al., 2021). Physiochemical variables include ammonia nitrogen, nitrate/nitrite, dissolved oxygen, pH, water temperature, total phosphorus, and turbidity. Climate drivers include air temperature, wind speed, shortwave radiation, specific humidity, and precipitation. Land use drivers include percent developed land use and percent agricultural land use (Table 1). We addressed multicollinearity, a statistical phenomenon that distorts the interpretation of the model, inflates the standard errors of the coefficients, and makes estimates extremely sensitive to minor changes in the model. We carefully evaluated the correlation among our independent variables to

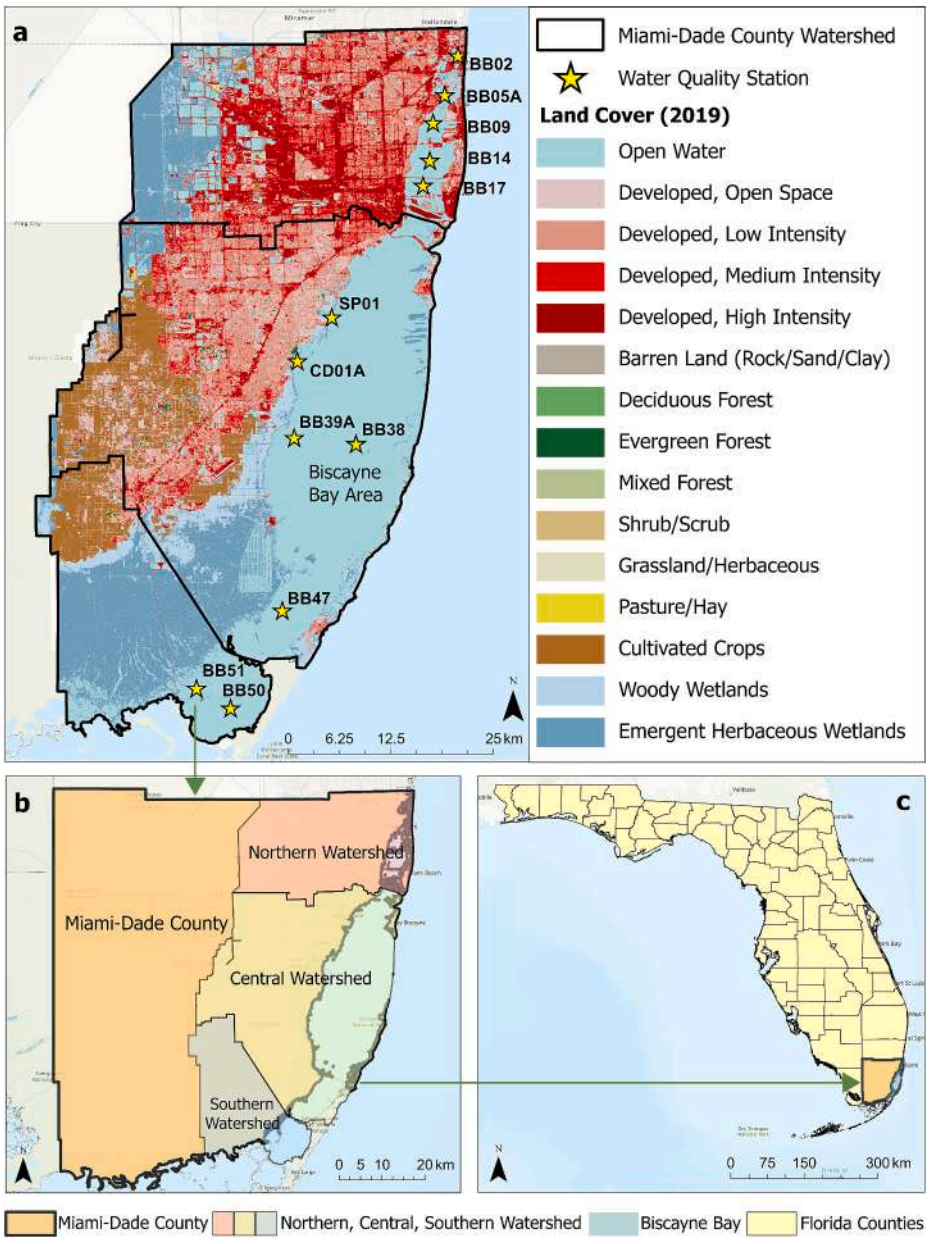


Fig. 1. The study area of Miami-Dade County and Biscayne Bay, including the locations of 12 water quality stations and various land use areas within Miami-Dade. Please note that the northern, central, and southern watersheds were divided by the combination of sub-watersheds provided by the South Florida Water Management District (SFWMD).

mitigate these risks. As illustrated in Supplementary information Fig. S2, we found a perfect negative correlation (correlation coefficient $r = -1.00$, *SI* Fig. S2) between the percent of developed land use and the percent of agricultural land use. Considering the more severe HABs in the northern Biscayne Bay (Fig. 3) and the highly urbanized nature of its upstream areas (Fig. 1), we excluded the agricultural percent from our analysis to mitigate multicollinearity issues. Regarding other instances of high correlation ($r > 0.8$, *SI* Fig. S2), these are primarily observed among climatic variables. However, considering that our dataset has a significantly larger number of samples compared to features, we were cautious about reducing features further to avoid losing valuable information. This approach allows us to maintain a balance between capturing a comprehensive set of influencing factors and ensuring the robustness of our model. It ensures that our model, including the interpretation of SHAP values, accurately reflects the contribution of each feature, thus providing more reliable insights into the dynamics influencing HABs.

All collected data have been subjected to thorough Quality Assurance and Quality Control (QA/QC) processes by their respective organizations. Additionally, we assessed extreme outliers, defined as data points exceeding 3* the interquartile range (IQR), and set the extreme outliers as missing values to ensure the highest data integrity. We received 12 station water quality data provided by the Miami-Dade Division of Environmental Resources Management (DERM). The data includes chlorophyll-a (Chl-a), ammonia nitrogen, nitrate/nitrite (NOx), dissolved oxygen (DO), pH, water temperature, total phosphorus (TP), and turbidity, provided by the Miami-Dade Division of Environmental Resources Management (DERM). Limited missing values were primarily found in the even-numbered months of the data provided by DERM from 1997 to 1999. We implemented the autoencoder (a neural network algorithm) to fill in missing data within this dataset. The average Chl-a concentrations of 12 water quality stations have a boosting trend in

Table 1
Summary of the target (dependent variable) and features (independent variables) in the prediction of HAB.

Variable and unit	Definition	Data source
Chlorophyll-a (ug/L)	Representative of HABs	DERM
Ammonia nitrogen (mg/L)	Concentration of ammonia nitrogen at gage location	DERM
Nitrate/nitrite (mg/L)	Concentration of NOx at gage location	DERM
Dissolved oxygen (mg/L)	Concentration of dissolved oxygen at gage location	DERM
pH	pH at gage location	DERM
Water temperature (°C)	Water temperature at gage location	DERM
Total phosphorus (mg/L)	Concentration of the sum of all phosphorus compounds that occur in various forms at gage location	DERM
Turbidity or water clarity (NTU)	Cloudiness of a fluid caused by suspended solids	DERM
Air temperature (°C)	2-m above ground air temperature for an area	NLDAS
Wind speed (m/s)	10-m above ground wind speed for an area	NLDAS
Shortwave radiation (W/m^2)	Shortwave radiation flux downwards for an area	NLDAS
Specific humidity (kg/kg)	2-m above ground specific humidity for an area	NLDAS
Precipitation (kg/m^2)	Precipitation monthly total for an area	NLDAS
Percent developed land use (%)	Ratio of developed area to total land area	NLCD
Percent agricultural land use (%)	Ratio of agricultural area to total land area	NLCD

the bay during our study period (*SI* Fig. S1). We acquired climate data, including air temperature, wind speed, shortwave radiation, specific humidity, and precipitation, from North American Land Data Assimilation System (NLDAS) Primary Forcing Data L4 Monthly 0.125 * 0.125 degree V002, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: February 2022 (Xia et al., 2012). We obtained land use data from the National Land Cover Database (NLCD) in 2001, 2004, 2006, 2008, 2011, 2013, 2016, and 2019 (Dewitz and U.S. Geological Survey, 2021). We calculated the developed percent and agricultural percent by the sub-watershed data from the South Florida Water Management District and the NLCD from the U.S. Geological Survey for each of the three regions. To bridge the gaps between these years with available land use data, we employed methods of linear interpolation and polynomial extrapolation. Developed/agricultural percents are the developed/agricultural area divided by the total area except for open water and wetlands, which account for a large portion of the regions. If they were not removed, the data would not be comparable across watersheds. After addressing missing values, we assigned the calculated yearly developed and agricultural percentages to each month within the same year for the respective sub-watersheds.

2.2. Autoregressive analysis of chlorophyll-a time series data in predictive modeling

Autoregressive characteristics refer to the regression relationship between a current value in a time series and its immediate predecessor. Given that many time series datasets possess autoregressive traits, we aimed to determine the optimal scenario for predicting Chl-a through autoregressive analysis. Identifying these autoregressive characteristics in Chl-a time series data is crucial for our study as it helps us capture temporal dependencies that can improve prediction accuracy. The Partial Autocorrelation Function (PACF) provides partial correlation, which is different from simple autocorrelation, by defining the relationship between time series observations and preceding observations while eliminating interference from other lags. This is particularly beneficial when analyzing time series data with autoregressive characteristics as it allows us to isolate and understand each lag's unique contribution rather than their combined effect as captured by simple correlation. We used Chl-a time series data to verify if they exhibited autoregressive traits using PACF. To illustrate the results of the autoregression analysis, we applied the 'graphics.tsaplots.plot_pacf' function from Python's 'statsmodels' package. This enabled us to describe the PCCs of Chl-a's 1–12 month time steps across 12 stations (Seabold and Perktold, 2010). When PCC is close to $+1/-1$, it indicates a strong positive/negative correlation signifying pronounced autoregressive characteristics. Conversely, when PCC approaches zero, it suggests an absence or minimal presence of autoregressive characteristics. Significant autocorrelation at specific lags was identified where $p < 0.05$ or where confidence intervals did not cross zero at the 95th percentile.

2.3. Application and performance evaluation of machine learning models

We analyzed the monthly PCC results for each station. Only the PCCs that either passed the significance test ($p < 0.05$) or fell outside the established confidence interval were deemed to exhibit autoregressive characteristics (Coad et al., 2014; Deng et al., 2021; Li et al., 2014; Xia et al., 2020). Upon identifying these significant lags using PACF, we carefully integrated this information into our next step, developing ML models. This was achieved by using the lags as inputs for the models, allowing them to capture the temporal dependencies and patterns in the time-series Chl-a data. By incorporating the lags, the ML models are better equipped to understand and predict the temporal dynamics of HABs in Biscayne Bay. Based on this, we choose to implement two prediction scenarios for the HAB of Biscayne Bay: current month Chl-a prediction and specific month lead Chl-a prediction. Given the incomplete understanding of complex physiochemical processes in HAB species, we utilized data-driven machine-learning algorithms for HAB prediction. Limited studies have focused on the monthly prediction of Chl-a concentrations, which presents unique challenges due to the complex nature of environmental data, including non-linearity, high dimensionality, and variable interactions. To address these challenges, we aimed to identify the best-performing ML algorithms for predicting Chl-a concentrations in Biscayne Bay. Common supervised machine learning algorithms for regression encompass linear models like Linear Regression, tree-based methods such as Decision Trees (fundamental) and Random Forests (ensemble learning), Support Vector Machines adept at complex function mapping with kernel function selection, Neural Networks, known for their adaptability and ability to handle noisy data, and Gaussian Process Models, valued for their probabilistic approach and capability to estimate uncertainty in predictions (Gramacy, 2020; Ray, 2019). Our selection process involved five representative ML algorithm categories, encompassing eight algorithms in total, each chosen for its distinct advantages in environmental analysis. (1) A linear model (Multivariable Linear Regression – MLR). This model, assuming a linear relationship between input and output variables, is straightforward and provides a baseline for comparison. The scenario of predicting HABs may be too complex for the MLR. However, it still serves as an excellent benchmark to determine the extent of improvement offered by other non-linear ML algorithms compared to a linear model. (2) Tree-based models including Decision Tree Regression (DTR), Extreme Gradient Boosting Regression (XGBR), Random Forest Regression (RFR), and Extra-Trees Regression (ETR). These models excel in handling complex environmental datasets due to their high interpretability and ability to manage numerical and categorical data. They are particularly effective in capturing non-linear relationships and interactions among variables and common in environmental data (Yajima and Derot, 2017). (3) Support Vector Machine in the form of Support Vector Regression (SVR), often chosen for its effectiveness in high-dimensional spaces. SVR is adept at handling the intricate patterns found in environmental datasets, making it a robust choice for complex data analysis (Vilas et al., 2014). (4) Gaussian Process Model through Gaussian Process Regression (GPR). This non-parametric method is

valuable for its provision of uncertainty measures along with predictions. GPR offers advantages for environmental data with considerable uncertainty (Gramacy, 2020). (5) A neural network approach via Multi-layer Perceptron Regression (MLPR). MLPR is suitable for capturing the intricate and often non-linear patterns present in large environmental datasets (Huang et al., 2019). By encompassing a diverse range of machine learning approaches, our study provides a comprehensive comparison in Chl-a concentration prediction across different types of models.

Each algorithm was employed to develop two models: a) predicting the current month's Chl-a levels and b) predicting a specific month lead's Chl-a levels. All models were developed under the framework of 'sklearn', 'numpy', and 'pandas' packages and used grid search to find the best hyperparameters for each model (Pedregosa et al., 2011). In this process, grid search was employed to systematically explore a wide range of hyperparameter values, enhancing the likelihood of identifying the optimal global solution for all critical parameters. This method proved effective in our context, allowing for a thorough yet efficient tuning given the small size of our dataset. The XGBR additionally employed the 'xgboost' package (Chen and Guestrin, 2016). The training set consisted of data from 1997 to 2015, accounting for approximately 79 % of the total dataset, while the test set comprised data from 2016 to 2020, representing approximately 21 % of the dataset. The 79 %–21 % split between the training and test sets was chosen to ensure a robust learning process while still maintaining an adequate set for validation. A larger training set (79 %) provides ample data for the models to learn effectively, while the test set (21 %) allows us to assess how well these models perform on unseen data, ensuring they can generalize beyond their training period. To determine the effectiveness of predictive machine learning models, we selected five metrics designed specifically for regression problems, including R^2 (coefficient of determination), MAE (mean absolute error), MSE (mean squared error), MAPE (mean absolute percentage error), and MedAE (median absolute error).

2.4. Relative importance of factors influencing chlorophyll-a concentrations

Machine learning models often pose challenges in terms of explainability, with a general trend of decreased interpretability as accuracy increases (Gilpin et al., 2018; Lipton, 2018). To bridge this gap, we employed SHAP values from the game theory-based 'shap' package to assess the relative importance of each feature influencing HAB (Lundberg et al., 2020; Pedregosa et al., 2011). The 'shap.explainer' function enabled us to plot feature importance for global explainability, facilitating a better understanding of decision-making within our models.

3. Results

3.1. Spatiotemporal variations of chlorophyll-a levels in Biscayne Bay

This study aimed to develop predictive models for HABs using multi-source data and machine learning algorithms, focusing on the relationship between the HAB indicator - chlorophyll-a (Chl-a) and various environmental stressors. We identified key influencing features and established accurate prediction scenarios, thereby aiding decision-makers in the early detection of potential HAB threats. We analyzed Chl-a concentrations measured at 12 Biscayne Bay stations from 1997 to 2020 on yearly (Fig. 2a) and monthly scales (Fig. 2b). As shown in Fig. 2a, there was a noticeable increase in Chl-a concentrations during 2005–2007, with 2006 experiencing the highest average concentration at 3.1 μg/L, indicating a substantial outbreak of HABs during this period. Notable peaks also occurred in 2014 (1.7 μg/L), 2017 (1.8 μg/L), and 2020 (1.9 μg/L). In contrast, lower averages were recorded in years like 1998 (0.9 μg/L) and 2009 (1.0 μg/L). In addition, *SI* Fig. S1a displays the average Chl-a concentrations and their linear fit collected from the 12 stations in Biscayne Bay between 1997 and 2020, illustrating an upward trend in average Chl-a concentration throughout the research period. Substantial outbreaks were noted in 1999, followed by an extended bloom from 2005 to 2007. Further outbreaks were observed in subsequent years: specifically, in 2010, 2014, 2017, and 2020. Overall, an increasing trend of Chl-a concentrations was observed throughout the study period, particularly after 2005, with the majority of average concentrations above 1.0 μg/L (see Fig. 2a and *SI* Fig. S1a). Similarly, after 2005, outliers became more pronounced and were characterized by more extreme values.

Analysis of the monthly averaged Chl-a concentration, as shown in Fig. 2b, indicates that the period from late spring to early fall may be particularly susceptible to high occurrences of HABs and ecosystem vulnerability. Specifically, from May to November, Chl-a concentrations were substantially elevated, exhibiting an increase of 31 % compared to other months within the year. The elevated Chl-a concentrations from late spring to early fall, as shown in Fig. 2b, are likely due to several seasonal environmental factors that favor the growth of HABs. During these months, warmer temperatures can stimulate algal growth by accelerating their metabolic and reproductive rates. This period often coincides with increased rainfall, which can lead to higher nutrient

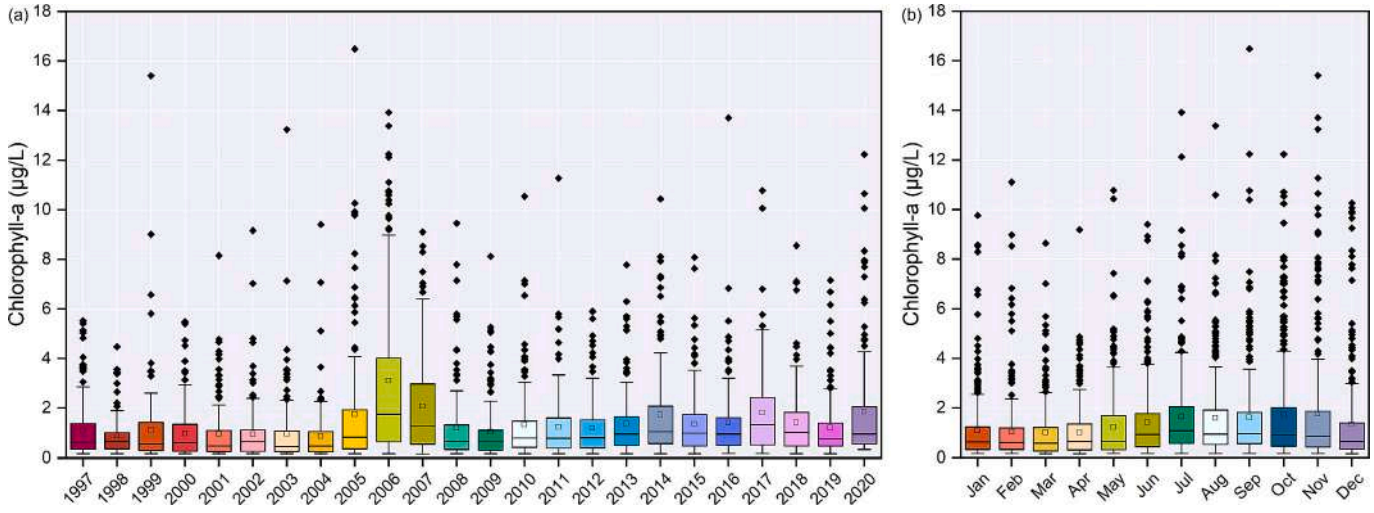


Fig. 2. Box and whisker plot of average chlorophyll-a concentrations in Biscayne Bay from 1997 to 2020. (a) Yearly scale, (b) monthly scale. The box portion of the box plot includes the 25th to 75th percentile data, and the horizontal line is the median. The square symbol stands for the mean value. The rhombus symbol represents the outlier. Any value that is 1.5 * IQR greater than the third quartile is designated as an outlier.

runoff from surrounding lands into the bay, further fueling algal blooms. Spatial variation in Chl-a concentration over two twelve-year periods, 1997–2008 (Fig. 3a) and 2009–2020 (Fig. 3b), is detailed in Fig. 3. The highest Chl-a concentrations were observed in northern Biscayne Bay (2.1 μg/L), while the southern and central regions recorded relatively lower concentrations at 1.1 μg/L and 0.7 μg/L respectively. Over time, an increase in Chl-a concentrations was noted in the northern and central areas of Biscayne Bay, whereas a decrease was observed in the south (see Fig. 3 and Table S1). This spatial-temporal trend underscores potential HAB issues during late spring to early fall, especially in northern and central regions where upstream areas are highly urbanized. Therefore, these findings highlight both temporal vulnerability (late spring to early fall) due to seasonality factors like temperature increases and nutrient availability; as well as spatial vulnerability with more urbanized northern and central regions showing higher Chl-a concentrations over time.

3.2. Autoregressive characteristics of chlorophyll-a concentrations

We investigated the presence of autoregressive characteristics, which refer to the regression relationship between a current value in a time series and its immediate predecessor, in Chl-a concentrations from water quality stations in Biscayne Bay. In our analysis of Chl-a concentrations in Biscayne Bay, we relied on the Partial Autocorrelation Function (PACF) to determine the optimal lead time for HAB predictions. The partial autocorrelation coefficients (PCCs) from each station demonstrated significant autoregressive characteristics for Chl-a concentrations at a one-month interval ($n = 11$ out of 12) instead of more than it,

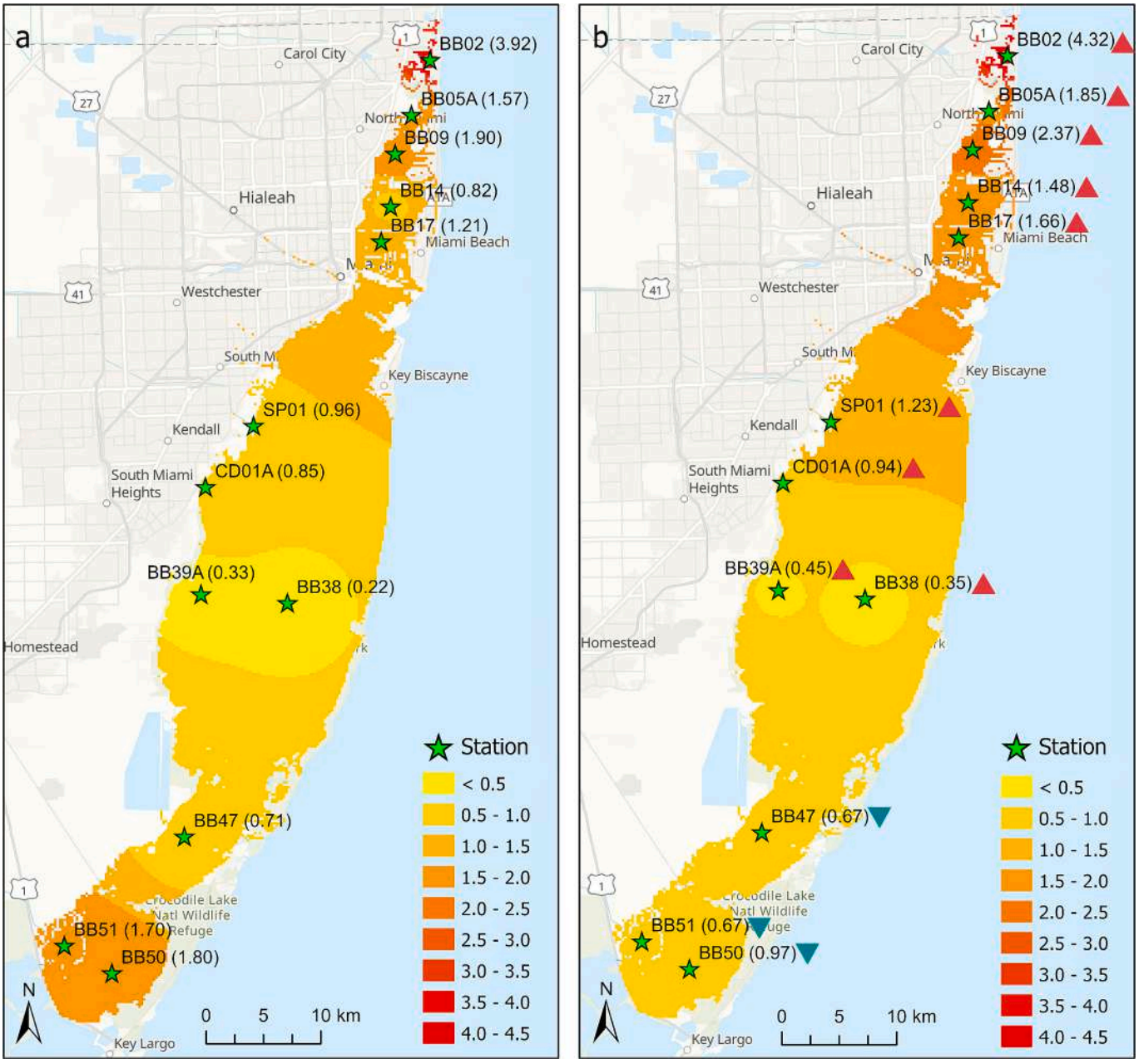


Fig. 3. Spatial distribution of average chlorophyll-a concentration (μg/L) at water quality stations in Biscayne Bay. (a) From 1997 to 2008, (b) from 2009 to 2020. The inverse distance weighting (IDW) was employed as the spatial interpolation method to visualize the spatial variation. Average Chl-a concentrations for each site during their respective study periods are provided in parentheses. An upward-pointing red triangle indicates an increase in average concentration, while a downward-pointing green triangle denotes a decrease.

primarily because the autoregressive features for intervals beyond one month did not consistently pass the significance tests (Fig. 4). The autoregressive relationship's predictive strength and statistical significance diminished for longer intervals (Fig. 4). This trend indicated that while Chl-a levels of a given month are significantly influenced by the preceding month, this influence does not extend as consistently to further past months. Therefore, focusing on a one-month lead time for HAB predictions is statistically justified for capturing the dynamics of Chl-a concentrations in Biscayne Bay. This pattern was evidenced by a high correlation between current and subsequent month Chl-a levels across nearly all stations ($p < 0.05$), as depicted in Fig. 4. Therefore, we designed a 1-month lead prediction scenario where we predict the next month's Chl-a concentrations using the historical data. While certain time steps have passed significance tests at specific sites, such as two months for BB14, three months for BB51, and four months for BB38, choosing a 1-month time step that most sites passed the significance test is the most rational approach for the entire bay. However, at station CD01A, no significant autoregressive pattern was found. This means that there was not a strong correlation between the current month's Chl-a levels and the subsequent month's levels at this station. This may be because many studies have found that the life cycle of coastal harmful algae is less than or inconsistent with one month (see details in Sections 4.1 and 4.2). As a result, relying solely on past data to predict future values may not be effective. In light of this, we introduced a second scenario where we predict current-month Chl-a concentrations. This allowed us to develop more accurate predictions by leveraging

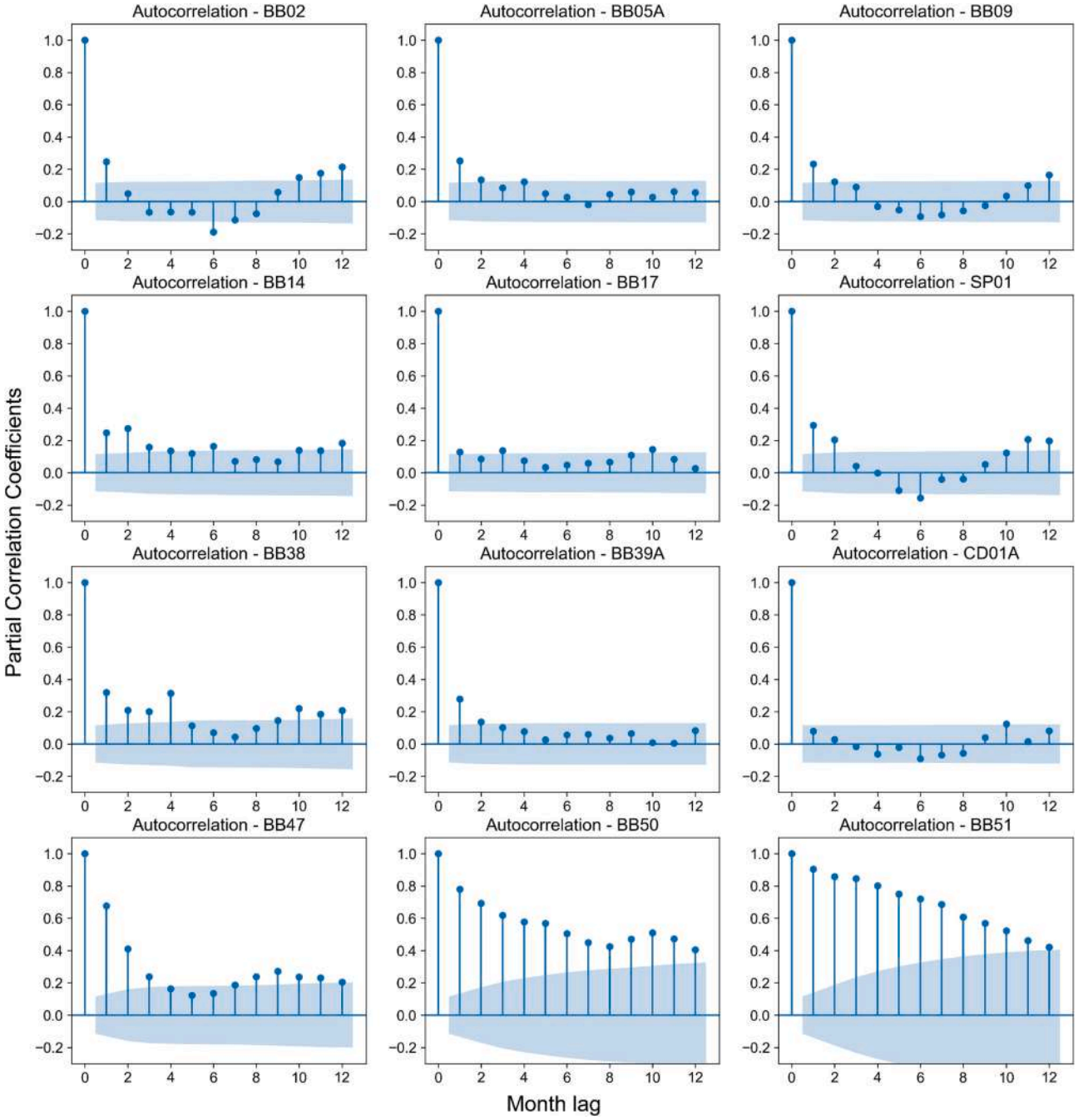


Fig. 4. Partial Autocorrelation Function (PACF) results of chlorophyll-a concentrations across all stations. The dark blue shade represents the 95 % confidence interval. If partial correlation coefficients lie outside this confidence interval, it indicates a strong correlation between the current value in a time series and its immediate predecessor.

autoregressive characteristics where they were present while also accommodating situations where such patterns were absent or less pronounced.

3.3. Predictive modeling of coastal algal blooms by multiple machine learning algorithms

We predicted Chl-a concentrations in Biscayne Bay using Environmental Stressors integrated into eight machine-learning algorithms. Building on the autoregressive characteristics of Chl-a levels outlined in Section 3.2, we hypothesized that predictive models incorporating data on Chl-a concentrations from one month prior could enhance prediction accuracy. To validate this hypothesis and identify the most effective machine learning algorithms, we deployed these eight algorithms to predict Chl-a concentrations in Biscayne Bay under two scenarios: current-month prediction and one-month lead prediction, which resulted in a total of 16 models for evaluation.

The prediction results for each algorithm under both scenarios, the current-month prediction and 1-month lead prediction, are presented in Tables 2 and 3, respectively. Our findings indicated that RFR performed best for current-month predictions (Table 2), while ETR excelled at one-month lead predictions (Table 3). Fig. 5 summarizes the performance of each algorithm (based on R^2 for the test) under both scenarios and includes an efficiency frontier plot to visually represent their relative performances. According to Caro et al. (2010), entities operating on the efficiency frontier are considered optimal within a given system; thus indicating RFR and ETR as top-performing algorithms in our study. In addition, the ETR model for one-month lead prediction outperforms all other models across both scenarios, indicating that 1-month lead prediction is achievable in our case study.

Fig. 6 depicts the performance of the two most effective models (i.e., ETR for 1-month lead prediction and RFR for current-month prediction) alongside differences between predicted and observed values across the entire dataset. Scatter density plots (Fig. 6a for ETR and Fig. 6c for RFR) indicate the concentration of data points and display the overall linear fits for both algorithms, while scatter training/test plots (Fig. 6b for ETR and Fig. 6d for RFR) provide separate visualizations of the training set and test set alongside their respective linear fits, facilitating a better understanding of the data distribution and model performance in the context of test versus train scenarios. As shown in Fig. 6, the ETR in the 1-month lead prediction slightly outperformed RFR in the current-month prediction, indicated by higher R2 and lower scores in other evaluation metrics. Generally, the predicted values are found to be lower than the observed ones. In addition, both models appear to struggle with accurately predicting extreme values. Overall, The prediction results are generally satisfactory, with R^2 values surpassing 0.4 for both test datasets using ETR and RFR under both scenarios. We presented the time series plots of the two best-performing models for each scenario for all stations, where most data points fall within the 50 % error range (see *SI* Fig. S3 and *SI* Fig. S4). In light of our findings, 1-month lead prediction with the ETR model provides a reliable approach for predicting Chl-a concentrations, despite the inherent challenges in predicting extreme values.

3.4. Identification of important features for coastal algal blooms in Biscayne Bay

Machine learning models can generate accurate predictions. However, understanding the underlying signals that these models rely on for decision-making can often be challenging. To better understand the key inputs influencing Chl-a predictions in our models, we evaluated feature importance using the 'shap' package that can explain the importance based on game theory. Fig. 7 visualizes the relative importance of input features on the output of the two best-performing models (ETR for 1-month lead prediction and RFR for current-month prediction). Fig. 7a indicates that the developed percent, total phosphorous, and NOx are the three most influential features in both models. Fig. 7b reveals that for the ETR model predicting 1-month ahead, the current month's Chl-a concentration is the most significant feature, followed by developed percent and precipitation. Fig. 7c presents that in the RFR model predicting for the current month, developed percent, total phosphorus, and NOx hold maximum importance. In both models, wind speed appears to be the least important feature among all considered inputs.

4. Discussion

4.1. The spatiotemporal distribution and autoregressive characteristics of chlorophyll-a

Our study's primary objective was to examine the spatiotemporal distribution and autoregressive properties of Chl-a in Biscayne Bay, aiming to understand potential temporal and spatial scales of HABs and improve their prediction. Our analysis revealed distinct periods with higher median and extreme Chl-a concentrations, specifically 2005–2007, 2010, 2014, 2017, and 2020, as compared to other times (Fig. 2a). The increase in Chl-a concentrations has been associated with seagrass degradation and macroalgae takeover as reported by Rudnick et al. (2006), Collado-Vides et al. (2013), and Santos et al. (2020). Fig. 2b indicates elevated levels of Chl-a concentrations from June to November (i.e., the wet season), implying an increased likelihood of HAB occurrences in Biscayne Bay during summer and autumn. However, this does not rule out the possibility of HABs occurring in South Florida's dry season from December to May. For example, Santos et al. (2020) applied satellite data to suggest that *Anadyomene spp* blooms peaked during February (dry season) and November, which indicates that blooms can occur even in the dry season. However, their study did not investigate blooms throughout all months but focused only on the dry season to allow for image selection with minimal cloudiness and sunlight reflection effects. In addition, their findings revealed that a post-bloom stage was reached in November 2014 and 2015, suggesting that the blooms likely peak before November, the end of the wet season. Therefore, based on the collective evidence from both studies, it can be concluded that the wet season continues to be the most probable timeframe for the occurrence of large-scale algal blooms. This knowledge is

Table 2
Current-month prediction results for each algorithm with metrics.

Algorithm	R^2		MAE		MSE		MAPE		MedAE	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
MLR	0.26	0.21	0.90	1.00	2.22	2.22	1.37	1.23	0.56	0.74
DTR	0.47	0.40	0.67	0.78	1.57	1.69	0.89	0.63	0.31	0.44
SVR	0.41	0.37	0.64	0.78	1.77	1.78	0.64	0.66	0.29	0.47
GPR	0.68	0.31	0.56	0.84	0.95	1.95	0.74	0.77	0.29	0.51
XGBR	0.62	0.42	0.60	0.78	1.12	1.65	0.83	0.67	0.33	0.45
RFR	0.69	**0.44**	**0.45**	**0.74**	0.92	**1.59**	**0.48**	**0.58**	**0.18**	0.38
ETR	**0.75**	0.41	**0.45**	0.76	**0.75**	1.66	0.50	0.65	0.20	0.43
MLPR	0.62	0.37	0.59	0.76	1.13	1.79	0.76	0.63	0.31	**0.36**[a]

[a] Note: the metrics of the best-performing algorithm are bolded.

Table 3
1-month LEAD prediction results for each algorithm with metrics.

Algorithm	R^2		MAE		MSE		MAPE		MedAE	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
MLR	0.48	0.35	0.70	0.77	1.56	1.88	0.90	0.65	0.39	0.42
DTR	0.57	0.39	0.60	0.75	1.27	1.76	0.67	**0.56**	0.30	**0.35**
SVR	0.55	0.42	0.73	0.84	1.34	1.69	1.19	0.88	0.54	0.58
GPR	0.56	0.42	0.64	0.75	1.32	1.68	0.79	0.63	0.34	0.40
XGBR	0.72	0.42	0.53	0.76	0.85	1.68	0.65	0.66	0.28	0.39
RFR	0.73	0.44	0.47	0.75	0.80	1.63	0.54	0.63	0.23	0.39
ETR	**0.92**	**0.47**	**0.21**	**0.72**	**0.28**	**1.54**	**0.20**	0.58	**0.08**	0.37
MLPR	0.55	0.42	0.67	0.79	1.32	1.67	0.94	0.74	0.39	0.46

Note: the metrics of the best-performing algorithm are bolded.

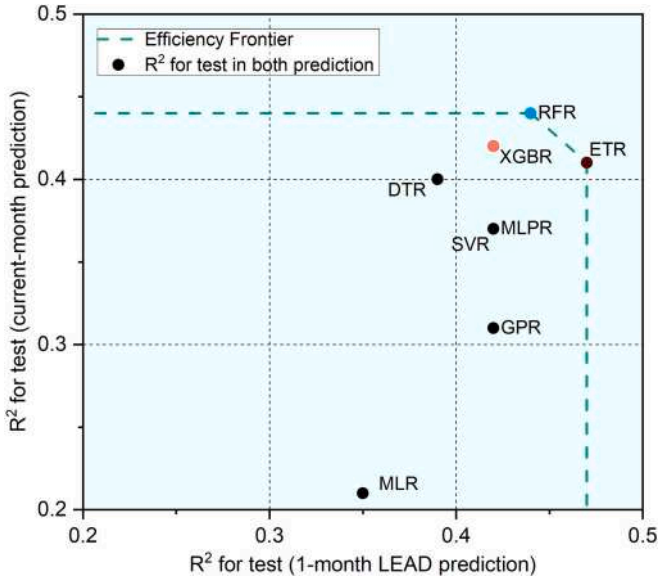


Fig. 5. Performance of algorithms (R^2 for test) in current-month prediction and 1-month lead prediction and the efficiency frontier.

crucial as it helps us identify potential high-risk periods for HABs, thus informing timely preventive measures.

Recent changes to Biscayne Bay's condition have raised concerns about ongoing ecological deterioration (Collado-Vides et al., 2011). Eutrophication is widely believed to be a factor in causing HAB occurrence, and developed areas are prone to produce more pollutants that cause eutrophication (Anderson et al., 2002; Heisler et al., 2008; Glibert, 2020). As shown in Fig. 3, Chl-a concentrations are higher in the northern bay than in other regions. This pattern likely results from excessive pollution from Miami-Dade County, which is located upstream of northern Biscayne Bay and is characterized by high urbanization and population density (Fig. 1 and *SI* Fig. S1b). Carey et al. (2011) further suggested a connection between land use in Miami-Dade County and nutrient levels in the bay and canals. Evidence supporting this includes reported fish kills in the northern bay and signs of seagrass cover loss over recent years (Santos et al., 2020). Millette et al. (2019) reported a significant increase in Chl-a concentrations throughout Biscayne Bay from 1995 to 2014. However, our study found varied degrees of increase across all regions except for South Bay, comprising Barnes Sound and Card Sound, during different periods compared to findings from Millette et al. (2019). While development in the northern watershed reached its peak with minimal increase during our study period, development within central and southern watersheds increased substantially (*SI* Fig. S1b). In this case, there was a considerable decrease in Chl-a concentration in the south bay, which is likely due to a substantial decline in agricultural land area. During the period from 1997 to 2008, many fertilizers were washed into South Bay by rainfall, potentially leading to algal proliferation. However, between 2009 and 2020, as developed areas increased and agricultural land decreased, the severity of eutrophication might have lessened due to a reduction in total pollutants. This change could explain the observed decrease in Chl-a levels during this period. This finding underscores the need for targeted environmental management strategies in these areas to mitigate eutrophication and prevent HAB occurrence.

Investigating the autoregressive characteristics of algal concentrations is crucial for predicting HABs. Previous studies have shown that different algal species may reach peak concentrations at varying time intervals. For instance, experiments conducted by Wang et al. (2017) indicated that algal density typically peaks around 21 days. Similarly, Hasegawa et al. (2001) found that *Closterium aciculare* enters the stationary phase after 2 weeks. In addition, Xia et al. (2020), Deng et al. (2021), and Izadi et al. (2021) applied a time-lag strategy to improve prediction results, which found 10 days, 7 days, and 8 days ahead models yielding the best accuracy, respectively. These findings suggest that the period for algae to reach peak concentrations following nutrient intake can range from 7 to 21 days. This finding is significant as it underscores the importance of considering species-specific growth dynamics when predicting HABs. However, many studies have overlooked examining the autoregressive characteristics of the specific algae of interest, relying heavily on previous research to improve predictive accuracy. Different types of algae exhibit varying growth rates in various environments (Brandenburg et al., 2019). Given Biscayne Bay's complex mix of chlorophytes, cyanobacteria, and diatoms, the life cycle of its HABs remains unclear (Wachnicka et al., 2020). In addition, previous studies have reported significant variations in residence times across different regions of Biscayne Bay (Wang et al., 2003; Wang and Kreeke, 1986). Specifically. Barnes Sound has been reported to have residence times that extend over several months. In the western sections of South Biscayne Bay, the typical residence time is around one month. Conversely, in the northern bay areas, residence times are generally shorter, often lasting less than half a month. Consequently, we propose that HABs in the bay exhibit autoregressive characteristics useful for prediction purposes. Our PACF results confirmed the autoregressive characteristics of a 1-month lag for Chl-a concentrations at most water quality stations (Fig. 4), suggesting a correlation between Chl-a concentrations one month apart. It was worth noting that the autoregressive characteristics were particularly evident in the results for three water quality stations—BB47, BB50, and BB51—with BB51 achieving a PCC of approximately 0.9 (Fig. 4). These three stations are located in the southernmost part of Biscayne Bay, within Barnes Sound and Card Sound, which have limited water exchange with the ocean and Florida Bay, resulting in longer residence times than other stations (Sengupta et al., 1978). This phenomenon contributes to the higher PCC values observed.

Our study offers a comprehensive examination of the spatiotemporal distribution and autoregressive properties of Chl-a in Biscayne Bay. The insights gained not only advance our understanding of HAB dynamics but also contribute directly towards improving their prediction; a key step towards better management strategies for preserving marine

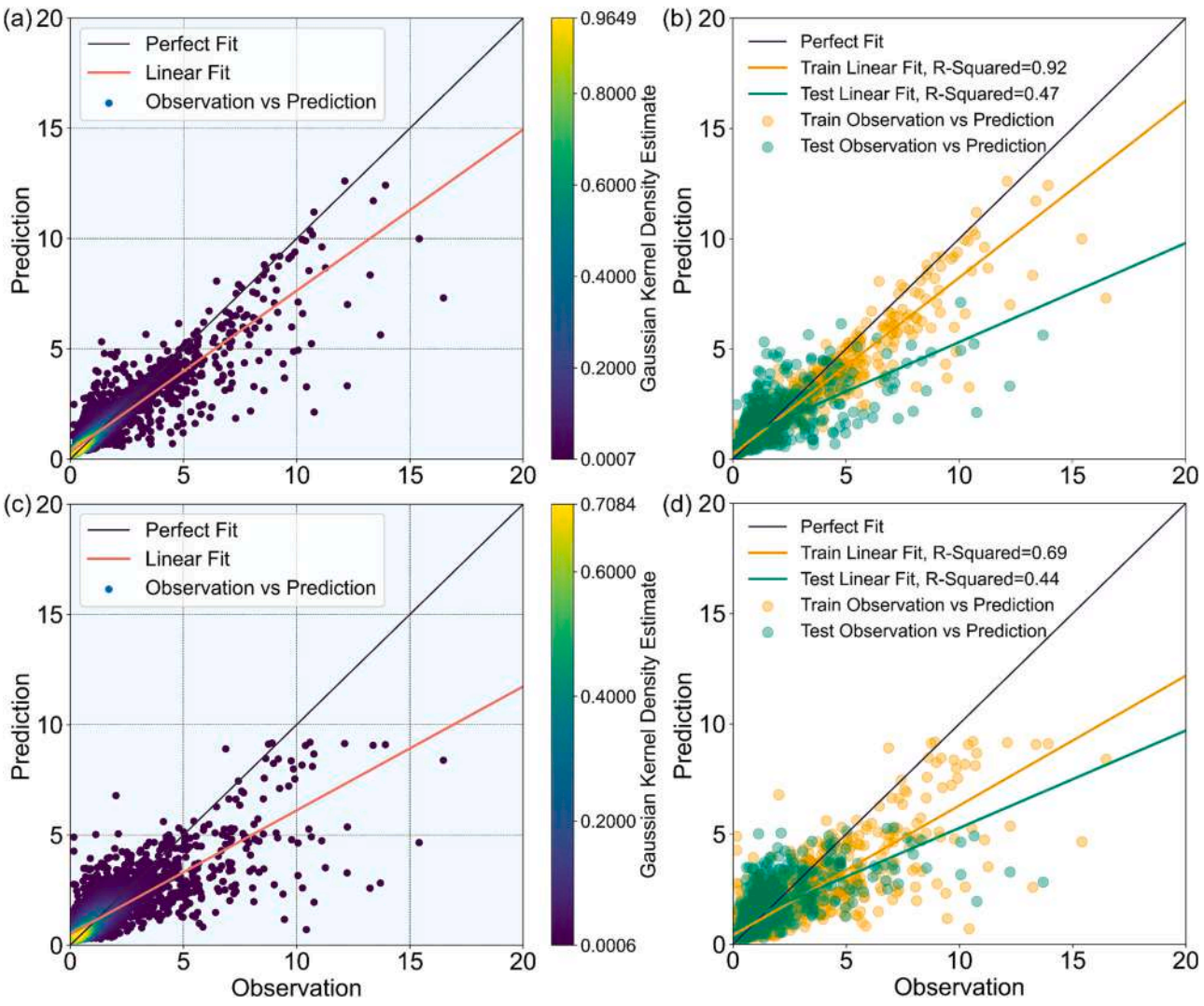


Fig. 6. Scatter density plot (left) and scatter training/test plot (right) for 1-month lead prediction in Extra-Trees Regressor (top) and current-month prediction in Random Forest Regressor (bottom). (a) Scatter density plot, ETR, 1-month lead prediction; (b) Scatter test/train plot, ETR, 1-month lead prediction; (c) Scatter density plot, RFR, current-month prediction; (d) Scatter test/train plot, RFR, current-month prediction. Note that: scatter density plots are composed of all data points, while the scatter plots are composed of training and test sets separately.

resources and protecting public health. Our findings enable us to predict when and where HABs are most likely to occur. With this information, authorities can implement timely preventive measures such as increased water quality monitoring during high-risk periods or targeted mitigation efforts in high-risk regions. Furthermore, the recognition of a one-month autoregressive pattern at most stations enables us to predict potential algal blooms one month ahead based on the current month's data. This provides valuable lead time for implementing appropriate response strategies before a bloom event occurs. In conclusion, our study's findings provide critical insights that can inform proactive management strategies aimed at preventing or mitigating the impacts of HABs in Biscayne Bay.

4.2. *Prediction of coastal algal blooms by machine learning methods*

Given the diversity, interaction, and intricacy of the multiple environmental factors that control algal blooms, developing a robust model that provides precise predictive ability can be a formidable challenge (Xia et al., 2020). Machine learning methods, emerging as viable alternatives to traditional biophysical process-based and statistical-empirical models, are increasingly applied in environmental studies, particularly in predicting HABs (Friedel et al., 2020; Izadi et al., 2021; Cruz et al., 2021). Machine learning models offer a feasible strategy for simulating HABs in complex coastal environments. These data-driven approaches provide higher prediction accuracy and can handle multi-dimensional datasets and nonlinear problems, avoiding the need for complicated and unresolved mathematical formulas f (Bergen et al., 2019; Jordan and Mitchell, 2015). For instance, *SI* Fig. S2 indicated that the Chl-a concentration was not correlated with any feature with a relatively high coefficient except for the 1-month lag Chl-a, stating the importance of including Chl-a's autoregressive characteristics in prediction and highlighting the inherent challenges in predicting monthly HABs. Nevertheless, our development of the ETR and RFR nonlinear predictive models has been successful. These models surpass the accuracy of linear counterparts, addressing the regression issue associated with algal blooms in Biscayne Bay effectively. This suggests that machine learning can serve as a practical means for nonlinear prediction, even when predicting the occurrence of algal blooms in coastal areas. Tables 2 and 3, along with Fig. 5, indicate that the 1-month lead prediction is more accurate than the current month prediction, with higher R^2 and lower error metrics. The optimal models for these scenarios are RFR for current-month predictions and ETR for 1-month lead predictions (Caro et al., 2010). In this study, ETR and RFR, both tree-based machine learning algorithms, excelled in predicting Chl-a concentrations in

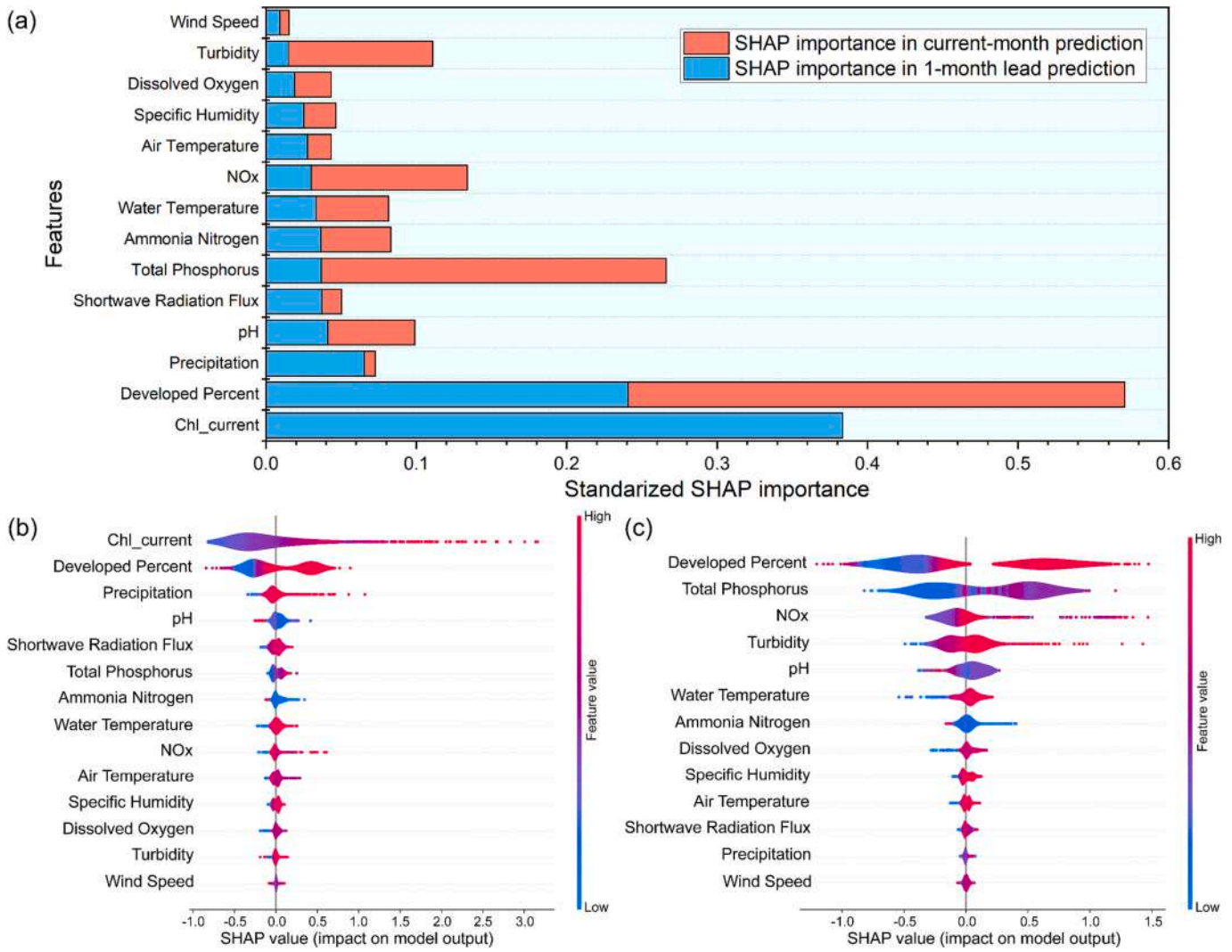


Fig. 7. Relative feature importance from ETR and RFR models. (a) Standardized SHAP importance of models in both scenarios. (b) SHAP value of features and the impact on the model output in the 1-month lead prediction of ETR. (c) SHAP value of features and the impact on the model output in the current-month prediction of RFR.

Biscayne Bay, suggesting that these types of algorithms may be particularly effective for datasets like ours, which are characterized by thousands of samples and tens of features (Mienye et al., 2019).

The life cycle of algae, particularly harmful varieties, often does not align with a monthly duration, typically spanning only a few weeks or days (Liang et al., 2023; Liu et al., 2020; Pelusi et al., 2020). This discrepancy presents challenges when utilizing monthly data to accurately predict Chl-a concentrations for subsequent time steps in this study. Specifically, constraints related to labor costs and observational equipment often limit the availability of time series data to semi-monthly or monthly frequencies. Such limitations can lead to inevitable errors in predicting HABs due to significant gaps between data collection frequencies and the actual life cycles of algae. To address these issues, prior studies leveraged daily or weekly remote sensing data directly or employed techniques such as linear interpolation to convert monthly data into a daily format. For example, Deng et al. (2021) successfully utilized neural networks to obtain satisfactory results by interpolating monthly and biweekly data into daily data. Izadi et al. (2021) demonstrated that XGBoost outperformed random forests and support vector machines, achieving an overall accuracy of 96 % by applying daily satellite data from MODIS (Moderate Resolution Imaging Spectroradiometer) using remote sensing techniques. While the implementation of data interpolation to obtain more frequent data intervals is a feasible strategy (Lee et al., 2003; Li et al., 2014), there is a potential risk of increased errors associated with this approach, especially when the dataset is non-linear. An alternative strategy is the direct use of monthly data for predictions to minimize potential risks associated with interpolated datasets and possible overestimations in predictive outcomes. Although this approach might yield lower accuracy and result in longer prediction intervals that do not align with the life cycle of harmful algae, it benefits from being based on relatively "authentic and accurate" data. Furthermore, it offers the advantage of providing proactive warnings for HAB management and mitigation over extended periods while also providing broader trend assessments through its predictive outcomes. For instance, Yajima and Derot (2017) applied random forest to predict monthly Chl-a, with the R^2 around 0.4–0.5. In addition, Jackson-Blake et al. (2022) predicted monthly and seasonal Chl-a with an R^2 value close to 0.4 by the Gaussian Bayesian network. Considering the aforementioned points, although we maintain that the 1-month lead prediction in our study has performed well and within acceptable limits (see *SI* Fig. 3 and *SI* Fig. 4), a shorter lead prediction would have yielded better accuracy.

Our analysis reveals that ETR generally performed better in the 1-month lead prediction than RFR in the current-month prediction. Overall, the effectiveness of tree-based machine learning algorithms, such as ETR and RFR, over other algorithms such as Support Vector

Regression (SVR), Gaussian Process Regression (GPR), and Multi-layer Perceptron Regression (MLPR) can be attributed to several factors. ETR and RFR excel in handling complex non-linear relationships between predictors and the target variable, which is crucial in environmental studies where relationships are often non-linear. They naturally account for interactions between features, are less sensitive to outliers, offer good interpretability alongside prediction accuracy, and handle high dimensional spaces well without requiring feature scaling or transformation. While each algorithm has its strengths, ETR and RFR were particularly effective for our study due to these attributes.

Our predictions tended to be slightly lower than the observed Chl-a concentrations, and the model's performance decreased when predicting extreme values (see Fig. 6). Similar challenges associated with predicting extreme values were also observed in the study by Yajima and Derot's (2017). These difficulties can be due to various factors including nonlinear relationships between variables, poor data quality or inappropriate data structure, and overfitting of the model (Ying, 2019). In light of these observations, further refinement of our approach is warranted. A potential strategy involves developing a hybrid predictive framework that integrates time series decomposition with machine learning to capture the seasonality and trends of HABs, enhancing predictive accuracy. Despite applying grid search techniques for tuning hyperparameters and considering regularization parameters to prevent overfitting, there are still flaws evident in our prediction metrics. To address these issues more effectively in future work, we may explore several strategies including error correction (Kim et al., 2022), clustering (Du et al., 2017), decomposition-ML hybrid models (Zhu et al., 2023), ensemble ML models (Zhang and Mahadevan, 2019), and additional or alternative regularization techniques (Srivastava et al., 2014).

To prevent and manage the occurrence of more severe HAB events, it is necessary to make reliable early warning predictions of algal blooms by considering key environmental variables (Xia et al., 2020). However, a lack of understanding of complex physical-biological processes and limited data availability have hindered the development and application of process-based modeling for early warning of algal blooms (Park et al., 2015). Our study indicates that the 1-month lead prediction model outperformed the current month prediction model, highlighting the potential for providing early warnings for HAB occurrences. Therefore, this finding suggests that data-driven machine-learning models could be a promising alternative approach when process-based modeling for early warning of algal blooms is not feasible. Particularly in scenarios where the data structure is similar to our case study, we strongly recommend a tree-based lead prediction model, especially since they have demonstrated superior accuracy and practical predictability compared to other methods and thus they can significantly enhance early warning decision systems for managing coastal HAB risks.

4.3. Important features regulating coastal algal blooms in Biscayne Bay

The occurrence of coastal HABs is attributed to a variety of factors, such as chemical factors (nitrogen and phosphorus), physical factors (salinity, dissolved oxygen, and pH), and meteorological factors (temperature and precipitation) (Anderson et al., 2021; Deng et al., 2021; Glibert, 2020; Griffith and Gobler, 2020). These factors influence the synthesis of proteins, nucleic acid, chlorophyll, and numerous processes, including metabolism, respiration, etc. (Glibert, 2020; Wells et al., 2015). In addition, phytoplankton predators, such as zooplankton, and physical factors, such as residence time or flow rate, have an impact on HAB formation (Glibert, 2020). While our current study in Biscayne Bay primarily focused on chemical, physical, land use, and meteorological factors influencing HABs, we recognize the potential impact of additional ecological and hydrological variables, such as predator populations and flow rates, which were not integrated into our ML-based predictive model because of the absence of long-term measurements for these biological and physical variables. Given this limitation in our study's scope due to data availability constraints, future research could focus on incorporating these variables. Specifically, including these available region-specific data is essential when adapting our predictive framework for HAB prediction in other regions. Integrating such comprehensive data in the ML model and feature importance analysis not only enhances predictive accuracy but also aids in developing a more nuanced understanding of HAB occurrences. Encompassing a wide array of relevant factors is crucial for our predictive framework. It ensures a robust and effective model to form region-specific prevention strategies and measures, thereby addressing the unique challenges posed by diverse environmental conditions.

Utilizing available data, we conducted a quantitative analysis with SHAP values. This analysis assessed the importance of each feature in the RFR and ETR models. Our findings indicated that in Biscayne Bay, Chl-a concentration is primarily influenced by the developed percentage of upstream Miami-Dade County, the current month's Chl-a concentration, total phosphorus, and NOx. Previous studies have demonstrated that Chl-a concentrations in the next period are strongly associated with Chl-a in the last period (Deng et al., 2021; Kim et al., 2014; Li et al., 2014). Consistent with these findings, our ETR model indicated that the current month Chl-a has a significant impact on the prediction for next month's levels, indicating a trend where an increase in the current month's Chl-a is likely to lead to a corresponding increase in the following month (Fig. 7a). In addition, total phosphorus and NOx are key indicators of eutrophication and crucial nutrient sources for algae proliferation, as reported by several studies highlighting their significance in HAB occurrence (Anderson et al., 2002; Deng et al., 2021; Glibert, 2020; Li et al., 2014). Similarly, our results suggested that increases in total phosphorus and NOx correspondingly lead to increases in Chl-a concentrations (i.e., positive correlation) (Fig. 7a and b). These findings align with the fact that eutrophication leads to HAB occurrence (Anderson et al., 2002), a phenomenon we have effectively confirmed for Biscayne Bay. Biscayne Bay is typically considered oligotrophic, and phosphorus is crucial in maintaining seagrass (Alexandre et al., 2021). Our findings indicate that phytoplankton tend to thrive better in conditions of low ammonia nitrogen and high phosphorus by SHAP values (Fig. 7). A study indicated that the seagrass *Thalassia testudinum*, as opposed to macroalgae like *Anadyomene* sp., is more resilient in high ammonium concentration environments (Alexandre et al., 2021). Additionally, both seagrass and macroalgae exhibit strong competitiveness for phosphorus. Interestingly, the northern bay has noted a slight downward trend in ammonium levels (Millette et al., 2019). Given that all micro and macro algae grow better in low ammonia nitrogen conditions and are competitive for phosphorus sources, and considering the observed decrease in ammonium in the northern bay, it can be inferred that algae could become formidable competitors to seagrass. This competition tends to lead to algal blooms in the northern and even expand to the central bay. While these blooms might be non-toxic, their proliferation could gradually replace the seagrass, as has already been observed, significantly disrupting the ecological balance of the system (Santos et al., 2020).

Climate change is projected to result in fluctuations in water and air temperature, as well as pH levels, with a specific projection of increased temperature and decreased pH in future scenarios for Biscayne Bay (Hinder et al., 2012). Although not all harmful algae species will exhibit increased activity due to climate change, a majority of species, particularly dinoflagellates, are likely to experience higher reproduction rates due to factors such as increasing temperatures, ocean acidification, etc. (Fu et al., 2012; Glibert, 2020; Wells et al., 2015). Our findings suggest that high temperatures and low pH levels positively influence the growth of HABs in Biscayne Bay (Fig. 7b and c). The projected increase in temperature and decrease in pH induced by climate change in the bay are expected to create more favorable conditions for the growth and proliferation of HABs. Consequently, HABs will likely become more prosperous in the future climate scenario within Biscayne Bay.

Upstream land use (i.e., developed percent) was the critical feature for HAB prediction in our study. Since water quality in Biscayne Bay is

closely linked to land use patterns in the upstream Miami-Dade County watersheds (Caccia and Boyer, 2005; Carey et al., 2011), and the northern bay corresponding to the most urbanized northern watershed is also the area with the most severe HAB (Fig. 3), we speculated that upstream land use may be related to downstream HAB. Although the "developed percent" is a relatively static dataset with minor short-term variations, it has long-term impacts on the downstream aquatic environment and ecosystems (Wang et al., 2022). While previous studies have focused on various aspects of coastal HAB prediction, the specific consideration of upstream land use as an independent factor has been less emphasized or explored in detail. In response to this gap, our study introduced an approach by incorporating the developed percentage of each of the three sections of upstream Miami-Dade County into the analysis of chlorophyll-a concentrations in the respective downstream portion of the bay. Surprisingly, the percentage of development emerged as one of the top two significant factors in both the current month and 1-month lead predictions. This suggests that incorporating the developed percentage adds valuable information to the prediction model, potentially providing insight into nutrient data. This is because areas with a higher proportion of impervious surfaces typically indicate an increase in pollutant levels, particularly during periods of heavy rainfall (Liu et al., 2014). Based on our findings, we strongly recommend that future studies and predictive models of coastal HABs consider upstream land use as a significant factor. Incorporating this into the prediction model may enhance its accuracy. While some models may exhibit overfitting, this does not necessarily equate to a loss of interpretability or predictability; such models can be beneficial in guiding further data collection or in developing more robust models. Additionally, It is important to note that overfitting is a relative concern rather than an absolute one (Hawkins, 2004). In this context, the other simpler models created in this study failed to achieve the same level of fitting quality as our developed RFR and ETR, thereby justifying their acceptance despite potential overfitting concerns.

4.4. Implications for prevention of coastal algal blooms

The diversity of HAB species coupled with the varying impacts, pose significant challenges for the authorities and stakeholders involved in coastal resource management. (Anderson, 2009). The factors impacting the formation mechanism of HABs are still unknown in many coastal areas, which presents a significant challenge in regulating or controlling these factors. Thus, it is essential to conduct more in-depth research on all aspects of HABs, especially in identifying the key factors affecting algal growth (Anderson, 2009). This study sheds light on the crucial factors influencing HABs and provides a predictive model for Chl-a concentrations in the following month, which can be used for the control implementation strategies and provides practical information for decision-makers. Our findings indicate that precipitation, total phosphorus, NOx, pH, water temperature, and notably upstream land use, specifically the degree of development, are the prominent features impacting Chl-a concentrations. This novel observation about upstream land use opens new avenues for both future research and policy-making. Future studies could focus on how specific aspects of urban development contribute to nutrient runoff and subsequent HABs. This interdisciplinary approach could merge urban planning with ecology and water management to fully comprehend these dynamics. From a policy perspective, this understanding can inform more effective coastal resource management strategies. Urban planning policies could be reevaluated with an emphasis on reducing impervious surfaces and enhancing green spaces to mitigate nutrient runoff. Given the complexity of managing climatic factors associated with HABs, we recommend directing regulatory efforts towards other manageable aspects such as nutrient levels linked with urbanization processes. In Miami-Dade County, the implementation of watershed management strategies must be prioritized to reduce impervious surfaces and lower the concentration of pollutants discharged into the bay from this highly urbanized watershed. This can be effectively achieved by consistently monitoring effluent from wastewater treatment plants and promoting an increase in green infrastructure implementation. Although we focused on Biscayne Bay, the framework developed in this study can be easily adapted for other coastal areas that are proximate to large cities. Our study not only deepens our understanding of Chl-a dynamics in coastal areas but also presents a robust predictive tool for future studies and management strategies. Given that the feature importance identified through machine learning and SHAP analyses does not indicate causal effects, we recommend conducting sensitivity analyses after developing a predictive model using this framework (PACF+ML + SHAP). Such analyses will help ensure that stakeholders and decision-makers can confidently rely on and act upon the model's predictions. For instance, one could adjust the most important features identified by SHAP values by ±20 %, observing how changes in input impact positive or negative shifts in Chl-a concentration. This approach would validate whether the real-world impacts align with the interpretations provided by SHAP analyses before the decision-making implementation.

5. Conclusions

Predicting coastal algae patterns is critical for governmental decision-making due to the significant risks and consequences associated with HABs including a decrease in biodiversity, disruption of food chains, seafood contamination, fish kill, food poisoning, and respiratory issues. Thus, accurate prediction of coastal algae patterns allows governments to take proactive measures against these potential hazards. Biscayne Bay, our case study site bordering Miami-Dade County, is already showing signs of stress due to anthropogenic influence from rapid urbanization and the growing population in the region. In this study, we developed a framework to predict Chl-a concentrations (an indicator of HABs) in Biscayne Bay, Florida, for 24 years (1997–2020) using eight machine learning algorithms, including two prediction scenarios. Our findings suggest that the Extra Trees Regressor (ETR) algorithm performed the best in predicting Chl-a concentrations one month in advance. Therefore, we strongly recommend using tree-based models for future studies aiming to predict Chl-a concentrations. Our analysis revealed that several factors significantly influenced the Chl-a concentrations in the downstream Biscayne Bay. The most influential factors identified in our study were the current month's Chl-a concentration, the percentage of developed area in the upstream Miami-Dade County watershed, phosphorus levels, and NOx. Our analysis suggests a correlation between increases in these factors and rising Chl-a concentrations, highlighting a potential link that warrants further investigation. These findings emphasize the need for careful management and control of these factors to potentially mitigate algal blooms in the downstream bay.

While the overall predictive results provided valuable insights, it is important to acknowledge certain limitations within our study. We observed that predicted values were frequently lower than observed values, indicating potential room for model improvement, and achieving accurate prediction models poses a significant challenge when relying on monthly data due to variable proliferation periods of algal blooms not consistently aligning with monthly data intervals. Future improvements could include incorporating extreme value prediction methods into our model which can help better predict unusual but highly impactful events; selecting additional relevant features such as flow rate or predators of phytoplankton might improve model performance; exploring other machine learning algorithms, ensemble methods, or hybrid models as different algorithms might offer improved accuracy; utilizing novel data processing methods such as anomaly detection and decomposition techniques can help refine predictions by identifying outliers that deviate from expected behavior based on previous data trends. This study has significant implications for environmental management and policy; the identified influential factors, such as developed percent, phosphorus levels, and NOx, can be managed through strategic policy decisions. For instance, urban planning policies

could be revised to limit the percentage of developed lands in sensitive watershed areas. Similarly, stricter regulations could be imposed on industries to control NOx and practices that contribute to increased phosphorus levels. By doing so, we can better manage the health of our coastal ecosystems and mitigate the impact of harmful algal blooms. Our study establishes a crucial framework for predicting coastal algal blooms by integrating various data sources and multiple features. This framework includes time lag determination, machine learning regressors for HAB prediction, and the identification of key influencing factors using SHAP value. The procedures used in this framework are standardized and extendable to other coastal regions due to the use of generalizable environmental and climate factors as model inputs. The successful application of this framework in other coastal areas will lead to the identification of crucial factors influencing HABs, thereby establishing an early warning system for HABs across similar coastal regions.

CRediT authorship contribution statement

Zhengxiao Yan: Conceptualization, Data curation, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. **Sara Kamanmalek:** Supervision, Validation, Visualization, Writing – review & editing. **Nasrin Alamdari:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nasrin Alamdari reports financial support was provided by Florida State University. Nasrin Alamdari reports a relationship with Florida State University that includes: employment and funding grants.

Data availability

The model codes used in this study are accessible at https://github.com/EvanYan666/ML_HAB.

Acknowledgments

We would like to express our gratitude to the anonymous peer reviewers for their valuable feedback, which greatly improved this paper. This study is supported by the United States Environmental Protection Agency under grant number 02D21822 and National Science Foundation under grant number 2200384.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2023.169253.

References

Alexandre, A., Collado-Vides, L., Santos, R., 2021. The takeover of Thalassia testudinum by Anadyomene sp. at Biscayne Bay, USA, cannot be simply explained by competition for nitrogen and phosphorous. Mar. Pollut. Bull. 167, 112326 https://doi.org/10.1016/j.marpolbul.2021.112326.

Anderson, D.M., 2009. Approaches to monitoring, control and management of harmful algal blooms (HABs). Ocean Coast. Manag. 52, 342. https://doi.org/10.1016/j.ocecoaman.2009.04.006.

Anderson, D.M., Glibert, P.M., Burkholder, J.M., 2002. Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. Estuaries 25, 704–726. https://doi.org/10.1007/BF02804901.

Anderson, D.M., Fensin, E., Gobler, C.J., Hoeglund, A.E., Hubbard, K.A., Kulis, D.M., Landsberg, J.H., Lefebvre, K.A., Provoost, P., Richlen, M.L., Smith, J.L., Solow, A.R., Trainer, V.L., 2021. Marine harmful algal blooms (HABs) in the United States: history, current status and future trends. In: Harmful Algae, Global Harmful Algal Bloom Status Reporting, 102, p. 101975. https://doi.org/10.1016/j.hal.2021.101975.

Asnaghi, V., Pecorino, D., Ottaviani, E., Pedroncini, A., Bertolotto, R.M., Chiantore, M., 2017. A novel application of an adaptable modeling approach to the management of toxic microalgal bloom events in coastal areas. Harmful Algae 63, 184–192. https://doi.org/10.1016/j.hal.2017.02.003.

Baek, S.-S., Kwon, Y.S., Pyo, J., Choi, J., Kim, Y.O., Cho, K.H., 2021. Identification of influencing factors of A. catenella bloom using machine learning and numerical simulation. Harmful Algae 103, 102007. https://doi.org/10.1016/j.hal.2021.102007.

Bergen, K.J., Johnson, P.A., Hoop, M.V. de, Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science 363. https://doi.org/10.1126/science.aau0323.

Brandenburg, K.M., Velthuis, M., Van de Waal, D.B., 2019. Meta-analysis reveals enhanced growth of marine harmful algae from temperate regions with warming and elevated CO2 levels. Glob. Chang. Biol. 25, 2607–2618. https://doi.org/10.1111/gcb.14678.

Caccia, V.G., Boyer, J.N., 2005. Spatial patterning of water quality in Biscayne Bay, Florida as a function of land use and water management. Mar. Pollut. Bull. 50, 1416–1429. https://doi.org/10.1016/j.marpolbul.2005.08.002.

Carey, R.O., Migliaccio, K.W., Li, Y., Schaffer, B., Kiker, G.A., Brown, M.T., 2011. Land use disturbance indicators and water quality variability in the Biscayne Bay Watershed, Florida. Ecol. Indic. 11, 1093–1104. https://doi.org/10.1016/j.ecolind.2010.12.009.

Caro, J.J., Nord, E., Siebert, U., McGuire, A., McGregor, M., Henry, D., de Pouvourville, G., Atella, V., Kolominsky-Rabas, P., 2010. The efficiency frontier approach to economic evaluation of health-care interventions. Health Econ. 19, 1117–1127. https://doi.org/10.1002/hec.1629.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Coad, P., Cathers, B., Ball, J.E., Kadluczka, R., 2014. Proactive management of estuarine algal blooms using an automated monitoring buoy coupled with an artificial neural network. Environ. Model Softw. 61, 393–409. https://doi.org/10.1016/j.envsoft.2014.07.011.

Collado-Vides, L., Mazzei, V., Thyberg, T., Lirman, D., 2011. Spatio-temporal Patterns and Nutrient Status of Macroalgae in a Heavily Managed Region of Biscayne Bay, Florida, USA, 54, pp. 377–390. https://doi.org/10.1515/bot.2011.046.

Collado-Vides, L., Avila, C., Blair, S., Leliaert, F., Rodriguez, D., Thyberg, T., Schneider, S., Rojas, J., Sweeney, P., Drury, C., Lirman, D., 2013. A persistent bloom of Anadyomene J.V. Lamouroux (Anadyomenaceae, Chlorophyta) in Biscayne Bay, Florida. Aquat. Bot. 111, 95–103. https://doi.org/10.1016/j.aquabot.2013.06.010.

Cruz, R.C., Reis Costa, P., Vinga, S., Krippahl, L., Lopes, M.B., 2021. A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. J. Mar. Sci. Eng. 9, 283. https://doi.org/10.3390/jmse9030283.

Deng, T., Chau, K.-W., Duan, H.-F., 2021. Machine learning based marine water quality prediction for coastal hydro-environment management. J. Environ. Manag. 284, 112051 https://doi.org/10.1016/j.jenvman.2021.112051.

Dewitz, J., U.S. Geological Survey, 2021. National Land Cover Database (NLCD) 2019 Products. https://doi.org/10.5066/P9KZCM54.

Díaz, P.A., Ruiz-Villarreal, M., Pazos, Y., Moita, T., Reguera, B., 2016. Climate variability and Dinophysis acuta blooms in an upwelling system. In: Harmful Algae, Applied Simulations and Integrated Modelling for the Understanding of Toxic and Harmful Algal Blooms (ASIMUTH), 53, pp. 145–159. https://doi.org/10.1016/j.hal.2015.11.007.

Du, X., Shao, F., Wu, S., Zhang, H., Xu, S., 2017. Water quality assessment with hierarchical cluster analysis based on Mahalanobis distance. Environ. Monit. Assess. 189, 335. https://doi.org/10.1007/s10661-017-6035-y.

Feki, W., Hamza, A., Frossard, V., Abdennadher, M., Hannachi, I., Jacquot, M., Belhassen, M., Aleya, L., 2013. What are the potential drivers of blooms of the toxic dinoflagellate Karenia selliformis? A 10-year study in the Gulf of Gabes, Tunisia, southwestern Mediterranean Sea. Harmful Algae 23, 8–18.

Fleming, L.E., Kirkpatrick, B., Backer, L.C., Walsh, C.J., Nierenberg, K., Clark, J., Reich, A., Hollenbeck, J., Benson, J., Cheng, Y.S., Naar, J., Pierce, R., Bourdelais, A. J., Abraham, W.M., Kirkpatrick, G., Zaias, J., Wanner, A., Mendes, E., Shalat, S., Hoagland, P., Stephan, W., Bean, J., Watkins, S., Clarke, T., Byrne, M., Baden, D.G., 2011. Review of Florida red tide and human health effects. Harmful Algae 10, 224–233. https://doi.org/10.1016/j.hal.2010.08.006.

Flynn, K.J., McGillicuddy, D.J., 2018. Modeling marine harmful algal blooms: current status and future prospects. In: Harmful Algal Blooms. John Wiley & Sons, Ltd, pp. 115–134. https://doi.org/10.1002/9781118994672.ch3.

Franks, P.J., 2018. Recent advances in modelling of harmful algal blooms. In: Global Ecology and Oceanography of Harmful Algal Blooms, pp. 359–377.

Friedel, M.J., Wilson, S.R., Close, M.E., Buscema, M., Abraham, P., Banasiak, L., 2020. Comparison of four learning-based methods for predicting groundwater redox status. J. Hydrol. 580, 124200 https://doi.org/10.1016/j.jhydrol.2019.124200.

Fu, F.X., Tatters, A.O., Hutchins, D.A., 2012. Global change and the future of harmful algal blooms in the ocean. Mar. Ecol. Prog. Ser. 470, 207–233.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp. 80–89.

Glibert, P.M., 2020. Harmful algae at the complex nexus of eutrophication and climate change. In: Harmful Algae, Climate Change and Harmful Algal Blooms, 91, p. 101583. https://doi.org/10.1016/j.hal.2019.03.001.

Glibert, P.M., Berdalet, E., Burford, M.A., Pitcher, G.C., Zhou, M., 2018. Global Ecology and Oceanography of Harmful Algal Blooms. Springer.

Gokaraju, B., Durbha, S.S., King, R.L., Younan, N.H., 2011. A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the Gulf of Mexico. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 4, 710–720.
Gramacy, R.B., 2020. Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences. CRC Press.
Griffith, A.W., Gobler, C.J., 2020. Harmful algal blooms: a climate change co-stressor in marine and freshwater ecosystems. In: Harmful Algae, Climate Change and Harmful Algal Blooms, 91, p. 101590. https://doi.org/10.1016/j.hal.2019.03.008.
Hasegawa, H., Sohrin, Y., Seki, K., Sato, M., Norisuye, K., Naito, K., Matsui, M., 2001. Biosynthesis and release of methylarsenic compounds during the growth of freshwater algae. Chemosphere 43, 265–272. https://doi.org/10.1016/S0045-6535(00)00137-5.
Hawkins, D.M., 2004. The problem of overfitting. J. Chem. Inf. Comput. Sci. 44, 1–12. https://doi.org/10.1021/ci0342472.
Heisler, J., Glibert, P.M., Burkholder, J.M., Anderson, D.M., Cochlan, W., Dennison, W. C., Dortch, Q., Gobler, C.J., Heil, C.A., Humphries, E., Lewitus, A., Magnien, R., Marshall, H.G., Sellner, K., Stockwell, D.A., Stoecker, D.K., Suddleson, M., 2008. Eutrophication and harmful algal blooms: a scientific consensus. In: Harmful Algae, HABs and Eutrophication, 8, pp. 3–13. https://doi.org/10.1016/j.hal.2008.08.006.
Hill, P.R., Kumar, A., Temimi, M., Bull, D.R., 2020. HABNet: machine learning, remote sensing-based detection of harmful algal blooms. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 3229–3239. https://doi.org/10.1109/JSTARS.2020.3001445.
Hinder, S.L., Hays, G.C., Edwards, M., Roberts, E.C., Walne, A.W., Gravenor, M.B., 2012. Changes in marine dinoflagellate and diatom abundance under climate change. Nat. Clim. Chang. 2, 271–275. https://doi.org/10.1038/nclimate1388.
Huang, X., Gao, L., Crosbie, R.S., Zhang, N., Fu, G., Doble, R., 2019. Groundwater recharge prediction using linear regression, multi-layer perception network, and deep learning. Water 11, 1879. https://doi.org/10.3390/w11091879.
Izadi, M., Sultan, M., Kadiri, R.E., Ghannadi, A., Abdelmohsen, K., 2021. A remote sensing and machine learning-based approach to forecast the onset of harmful algal bloom. Remote Sens. 13, 3863. https://doi.org/10.3390/rs13193863.
Jackson-Blake, L.A., Clayer, F., Haande, S., Sample, J.E., Moe, S.J., 2022. Seasonal forecasting of lake water quality and algal bloom risk using a continuous Gaussian Bayesian network. Hydrol. Earth Syst. Sci. 26, 3103–3124. https://doi.org/10.5194/hess-26-3103-2022.
Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349, 255–260.
Kim, Y., Shin, H.S., Plummer, J.D., 2014. A wavelet-based autoregressive fuzzy model for forecasting algal blooms. Environ. Model Softw. 62, 1–10.
Kim, J., Jones, J.R., Seo, D., 2021. Factors affecting harmful algal bloom occurrence in a river with regulated hydrology. J. Hydrol. Reg. Stud. 33, 100769 https://doi.org/10.1016/j.ejrh.2020.100769.
Kim, J., Yu, J., Kang, C., Ryang, G., Wei, Y., Wang, X., 2022. A novel hybrid water quality forecast model based on real-time data decomposition and error correction. Process. Saf. Environ. Prot. 162, 553–565. https://doi.org/10.1016/j.psep.2022.04.020.
Kouakou, C.R.C., Poder, T.G., 2019. Economic impact of harmful algal blooms on human health: a systematic review. J. Water Health 17, 499–516. https://doi.org/10.2166/wh.2019.064.
Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. In: Geoscience Frontiers, Special Issue: Progress of Machine Learning in Geosciences, 7, pp. 3–10. https://doi.org/10.1016/j.gsf.2015.07.003.
Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modelling of coastal algal blooms. Ecol. Model. 159, 179–201. https://doi.org/10.1016/S0304-3800(02)00281-8.
Li, X., Yu, J., Jia, Z., Song, J., 2014. Harmful algal blooms prediction with machine learning models in Tolo Harbour. In: 2014 International Conference on Smart Computing. IEEE, pp. 245–250.
Li, M.F., Glibert, P.M., Lyubchich, V., 2021. Machine learning classification algorithms for predicting Karenia brevis blooms on the West Florida shelf. J. Mar. Sci. Eng. 9, 999. https://doi.org/10.3390/jmse9090999.
Liang, D., Xiang, H., Jin, P., Xia, J., 2023. Response mechanism of harmful algae Phaeocystis globosa to ocean warming and acidification. Environ. Pollut. 320, 121008 https://doi.org/10.1016/j.envpol.2023.121008.
Lipton, Z.C., 2018. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 16, 31–57.
Liu, J., Sample, D.J., Bell, C., Guan, Y., 2014. Review and research needs of bioretention used for the treatment of urban stormwater. Water 6, 1069–1099. https://doi.org/10.3390/w6041069.
Liu, Y., Hu, Z., Deng, Y., Tang, Y.Z., 2020. Evidence for production of sexual resting cysts by the toxic dinoflagellate Karenia mikimotoi in clonal cultures and marine sediments. J. Phycol. 56, 121–134. https://doi.org/10.1111/jpy.12925.
Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9.
Ly, Q.V., Nguyen, X.C., Lê, N.C., Truong, T.-D., Hoang, T.-H.T., Park, T.J., Maqbool, T., Pyo, J., Cho, K.H., Lee, K.-S., Hur, J., 2021. Application of machine learning for eutrophication analysis and algal bloom prediction in an urban river: a 10-year study of the Han River, South Korea. Sci. Total Environ. 797, 149040 https://doi.org/10.1016/j.scitotenv.2021.149040.
Maze, G., Olascoaga, M.J., Brand, L., 2015. Historical analysis of environmental conditions during Florida Red Tide. Harmful Algae 50, 1–7. https://doi.org/10.1016/j.hal.2015.10.003.
McCabe, R.M., Hickey, B.M., Kudela, R.M., Lefebvre, K.A., Adams, N.G., Bill, B.D., Gulland, F.M.D., Thomson, R.E., Cochlan, W.P., Trainer, V.L., 2016. An unprecedented coastwide toxic algal bloom linked to anomalous ocean conditions. Geophys. Res. Lett. 43, 10,366–10,376. https://doi.org/10.1002/2016GL070023.
Medina, M., Kaplan, D., Milbrandt, E.C., Tomasko, D., Huffaker, R., Angelini, C., 2022. Nitrogen-enriched discharges from a highly managed watershed intensify red tide (Karenia brevis) blooms in southwest Florida. Sci. Total Environ. 827, 154149 https://doi.org/10.1016/j.scitotenv.2022.154149.
Mienye, I.D., Sun, Y., Wang, Z., 2019. Prediction performance of improved decision tree-based algorithms: a review. In: Procedia Manufacturing, the 2nd International Conference on Sustainable Materials Processing and Manufacturing, SMPM 2019, 8–10 March 2019, Sun City, South Africa 35, pp. 698–703. https://doi.org/10.1016/j.promfg.2019.06.011.
Millette, N.C., Kelble, C., Linhoss, A., Ashby, S., Visser, L., 2019. Using spatial variability in the rate of change of chlorophyll a to improve water quality management in a subtropical oligotrophic estuary. Estuar. Coasts 42, 1792–1803. https://doi.org/10.1007/s12237-019-00610-5.
Muttil, N., Chau, K., 2006. Neural network and genetic programming for modelling coastal algal blooms. Int. J. Environ. Pollut. 28, 223.
Norton, L., Elliott, J.A., Maberly, S.C., May, L., 2012. Using models to bridge the gap between land use and algal blooms: an example from the Loweswater catchment, UK. In: Environmental Modelling & Software, Thematic Issue on Expert Opinion in Environmental Modelling and Management, 36, pp. 64–75. https://doi.org/10.1016/j.envsoft.2011.07.011.
Paerl, H.W., Gardner, W.S., McCarthy, M.J., Peierls, B.L., Wilhelm, S.W., 2014. Algal blooms: noteworthy nitrogen. Science 346, 175. https://doi.org/10.1126/science.346.6206.175-a.
Papenfus, M., Schaeffer, B., Pollard, A.I., Loftin, K., 2020. Exploring the potential value of satellite remote sensing to monitor chlorophyll-a for US lakes and reservoirs. Environ. Monit. Assess. 192, 808. https://doi.org/10.1007/s10661-020-08631-5.
Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci. Total Environ. 502, 31–41.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pelusi, A., Rotolo, F., Gallo, A., Ferrante, M.I., Montresor, M., 2020. Effects of elutriates from contaminated coastal sediments on different life cycle phases of planktonic diatoms. Mar. Environ. Res. 155, 104890 https://doi.org/10.1016/j.marenvres.2020.104890.
Ray, S., 2019. A quick review of machine learning algorithms. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). Presented at the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 35–39. https://doi.org/10.1109/COMITCon.2019.8862451.
Roiha, P., Westerlund, A., Nummelin, A., Stipa, T., 2010. Ensemble forecasting of harmful algal blooms in the Baltic Sea. In: Journal of Marine Systems, GEOHAB Modeling, 83, pp. 210–220. https://doi.org/10.1016/j.jmarsys.2010.02.015.
Rudnick, D., Madden, C.J., Kelly, S.P., Bennett, R., Cunniff, K., 2006. Algae blooms in eastern Florida Bay and southern Biscayne Bay. In: Coastal Ecosystems Division, South Florida Water Management District Technical Report.
Santos, R.O., Varona, G., Avila, C.L., Lirman, D., Collado-Vides, L., 2020. Implications of macroalgae blooms to the spatial structure of seagrass seascapes: the case of the Anadyomene spp.(Chlorophyta) bloom in Biscayne Bay, Florida. Mar. Pollut. Bull. 150, 110742.
Seabold, S., Perktold, J., 2010. Statsmodels: econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference. Austin, TX, p. 10.25080.
Sengupta, S., Lee, S.S., Miller, H.P., 1978. Three-dimensional Numerical Investigations of Tide and Wind Induced Transport Processes in Biscayne Bay.
Singh, A., Hårding, K., Reddy, H.R.V., Godhe, A., 2014. An assessment of Dinophysis blooms in the coastal Arabian Sea. Harmful Algae 34, 29–35.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Valbi, E., Ricci, F., Capellacci, S., Casabianca, S., Scardi, M., Penna, A., 2019. A model predicting the PSP toxic dinoflagellate Alexandrium minutum occurrence in the coastal waters of the NW Adriatic Sea. Sci. Rep. 9, 4166. https://doi.org/10.1038/s41598-019-40664-w.
Vilas, L.G., Spyrakos, E., Palenzuela, J.M.T., Pazos, Y., 2014. Support vector machine-based method for predicting Pseudo-nitzschia spp. blooms in coastal waters (Galician rias, NW Spain). Prog. Oceanogr. 124, 66–77.
Wachnicka, A., Browder, J., Jackson, T., Louda, W., Kelble, C., Abdelrahman, O., Stabenau, E., Avila, C., 2020. Hurricane Irma's impact on water quality and phytoplankton communities in Biscayne Bay (Florida, USA). Estuar. Coasts 43, 1217–1234. https://doi.org/10.1007/s12237-019-00592-4.
Walsh, J.J., Lenes, J.M., Darrow, B., Parks, A., Weisberg, R.H., 2016. Impacts of combined overfishing and oil spills on the plankton trophodynamics of the West Florida shelf over the last half century of 1965–2011: a two-dimensional simulation analysis of biotic state transitions, from a zooplankton-to a bacterioplankton-modulated ecosystem. Cont. Shelf Res. 116, 54–73.
Wang, J.D., Kreeke, J. van de, 1986. Tidal circulation in North Biscayne Bay. J. Waterw. Port Coast. Ocean Eng. 112, 615–631. https://doi.org/10.1061/(ASCE)0733-950X(1986)112:6(615).
Wang, J.D., Luo, J., Ault, J.S., 2003. Flows, salinity, and some implications for larval transport in south Biscayne Bay, Florida. Bull. Mar. Sci. 72, 695–723.
Wang, Li, Wang, X., Jin, X., Xu, J., Zhang, H., Yu, J., Sun, Q., Gao, C., Wang, Lingbin, 2017. Analysis of algae growth mechanism and water bloom prediction under the

effect of multi-affecting factor. In: Saudi Journal of Biological Sciences, Computational Intelligence Research & Approaches in Bioinformatics and Biocomputing, 24, pp. 556–562. https://doi.org/10.1016/j.sjbs.2017.01.026.

Wang, J., Bouwman, A.F., Liu, X., Beusen, A.H.W., Van Dingenen, R., Dentener, F., Yao, Y., Glibert, P.M., Ran, X., Yao, Q., Xu, B., Yu, R., Middelburg, J.J., Yu, Z., 2021. Harmful algal blooms in Chinese coastal waters will persist due to perturbed nutrient ratios. Environ. Sci. Technol. Lett. 8, 276–284. https://doi.org/10.1021/acs.estlett.1c00012.

Wang, Y., Chen, X., Gao, M., Dong, J., 2022. The use of random forest to identify climate and human interference on vegetation coverage changes in southwest China. Ecol. Indic. 144, 109463 https://doi.org/10.1016/j.ecolind.2022.109463.

Weisberg, R.H., Liu, Y., Lembke, C., Hu, C., Hubbard, K., Garrett, M., 2019. The coastal ocean circulation influence on the 2018 West Florida Shelf K. brevis red tide bloom. J. Geophys. Res. Oceans 124, 2501–2512. https://doi.org/10.1029/2018JC014887.

Wells, M.L., Trainer, V.L., Smayda, T.J., Karlson, B.S.O., Trick, C.G., Kudela, R.M., Ishikawa, A., Bernard, S., Wulff, A., Anderson, D.M., Cochlan, W.P., 2015. Harmful algal blooms and climate change: learning from the past and present to forecast the future. Harmful Algae 49, 68–93. https://doi.org/10.1016/j.hal.2015.07.009.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., Mocko, D., 2012. Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. J. Geophys. Res. Atmos. 117 https://doi.org/10.1029/2011JD016048.

Xia, R., Wang, G., Zhang, Y., Yang, P., Yang, Z., Ding, S., Jia, X., Yang, C., Liu, C., Ma, S., Lin, J., Wang, X., Hou, X., Zhang, K., Gao, X., Duan, P., Qian, C., 2020. River algal blooms are well predicted by antecedent environmental conditions. Water Res. 185, 116221 https://doi.org/10.1016/j.watres.2020.116221.

Xu, Y., Cheng, C., Zhang, Y., Zhang, D., 2014. Identification of algal blooms based on support vector machine classification in Haizhou Bay, East China Sea. Environ. Earth Sci. 71, 475–482. https://doi.org/10.1007/s12665-013-2455-3.

Yajima, H., Derot, J., 2017. Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. J. Hydroinf. 20, 206–220. https://doi.org/10.2166/hydro.2017.010.

Ying, X., 2019. An overview of overfitting and its solutions. J. Phys. Conf. Ser. 1168, 022022 https://doi.org/10.1088/1742-6596/1168/2/022022.

Yu, P., Gao, R., Zhang, D., Liu, Z.-P., 2021. Predicting coastal algal blooms with environmental factors by machine learning methods. Ecol. Indic. 123, 107334 https://doi.org/10.1016/j.ecolind.2020.107334.

Zhang, X., Mahadevan, S., 2019. Ensemble machine learning models for aviation incident risk prediction. Decis. Support. Syst. 116, 48–63. https://doi.org/10.1016/j.dss.2018.10.009.

Zhou, Y., Yan, W., Wei, W., 2021. Effect of sea surface temperature and precipitation on annual frequency of harmful algal blooms in the East China Sea over the past decades. Environ. Pollut. 270, 116224 https://doi.org/10.1016/j.envpol.2020.116224.

Zhou, Z.-X., Yu, R.-C., Zhou, M.-J., 2022. Evolution of harmful algal blooms in the East China Sea under eutrophication and warming scenarios. Water Res. 221, 118807 https://doi.org/10.1016/j.watres.2022.118807.

Zhu, X., Guo, H., Huang, J.J., Tian, S., Zhang, Z., 2023. A hybrid decomposition and machine learning model for forecasting chlorophyll-a and total nitrogen concentration in coastal waters. J. Hydrol. 619, 129207 https://doi.org/10.1016/j.jhydrol.2023.129207.