Journal of Environmental Management 364 (2024) 121463

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Research article

Integrating temporal decomposition and data-driven approaches for predicting coastal harmful algal blooms

Zhengxiao Yan, Nasrin Alamdari *

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

ARTICLE INFO

Handling Editor: Lixiao Zhang

Keywords:
Harmful algal bloom prediction
Hybrid model
Machine learning
Temporal decomposition

ABSTRACT

Frequent coastal harmful algal blooms (HABs) threaten the ecological environment and human health. Biscayne Bay in southeastern Florida also faces algal bloom issues; however, the mechanisms driving these blooms are not fully understood, emphasizing the importance of HAB prediction for effective environmental management. The overarching goal of this study is to offer a robust HAB predictive framework and try to enhance the understanding of HAB dynamics. This study established three scenarios to predict chlorophyll-a concentrations, a recognized representative of HABs: Scenario 1 (S1) using single nonlinear machine learning (ML) algorithms, hybrid Scenario 2 (S2) combining linear models and nonlinear ML algorithms, and hybrid Scenario 3 (S3) combining temporal decomposition and ML (TD-ML) algorithms. The novel-developed S3 TD-ML hybrid models demonstrated superior predictive capabilities, achieving all R^2 values above 0.9 and MAPE under 30% in tests, significantly outperforming the S1 with an average R^2 of 0.16 and the S2 with an R^2 of −0.06. S3 models effectively captured the algal dynamics, successfully predicting complex time series with extremes and noise. In addition, we unveiled the relationship between environmental variables and chlorophyll-a through correlation analysis and found that climate change might intensify the HABs in Biscayne Bay. This research developed a precise predictive framework for early warning and proactive management of HABs, offering potential global applicability and improved prediction accuracy to address HAB challenges.

1. Introduction

Coastal harmful algal bloom (HAB) is a prominent environmental phenomenon that affects human activities and ecosystems in coastal areas worldwide (Hallegraeff et al., 2021). HABs result from the rapid overgrowth of certain microscopic algae called phytoplankton, thriving in favorable environmental conditions, often leading to ocean surface hypoxia, discoloration, and the production of toxins harmful to marine life and human health (Anderson et al., 2021). The formation of coastal HABs is affected by the comprehensive influence of various environmental factors (Yan et al., 2024b). For example, excess nutrients by human activities, such as nitrogen and phosphorus, can cause ecosystem imbalances and create proper conditions for algal blooms (Dai et al., 2023). Coastal HABs have become a growing concern due to their ecological impacts on the food web and biodiversity and economic impacts on fisheries, tourism, and human health, especially under climate change (Wells et al., 2015). Climate change impacts, sea surface temperature rise, elevated pCO_2, and ocean acidification, leading to potential changes in nutrient dynamics and physical environmental conditions, have created favorable conditions for the proliferation of certain harmful algal species (Glibert, 2020). As these conditions increasingly favor the formation of specific HABs, the future frequency, intensity, and geographic range of these blooms may continue to expand, intensifying the challenges to coastal ecosystems and human societies (Gobler, 2020). Mitigating blooms and their detrimental influences requires a comprehensive understanding of their causes and effective management strategies, including timely HAB prediction (Deng et al., 2021). Therefore, ensuring the accuracy of HAB prediction is necessary to better respond to this global environmental challenge and to protect the ecological environment and human well-being in coastal areas.

Various prediction methods have been developed to address the HAB issues (Yan et al., 2024b). Process-based models require detailed comprehension but face challenges in accurately expressing life processes due to complexities in unknown algae dynamics (Flynn and McGillicuddy, 2018). Traditional empirical-statistical models, which use empirical formulas or statistical fits to correlate environmental variables with HAB indicators, sometimes struggle to capture complex

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

https://doi.org/10.1016/j.jenvman.2024.121463
Received 6 March 2024; Received in revised form 23 April 2024; Accepted 9 June 2024
Available online 14 June 2024
0301-4797/Published by Elsevier Ltd.

nonlinear relationships and lack predictive accuracy (Franks, 2018). Both approaches have their contribution to studying HAB dynamics but often encounter limitations in accurately capturing complex interactions and maintaining predictive accuracy under varied conditions. By contrast, machine learning algorithms provide a significant edge in HAB prediction by uncovering complex relationships without the need for explicit mathematical modeling of unknown processes. Recent studies have employed more granular daily and hourly data to predict HABs using sophisticated machine learning and deep learning techniques, demonstrating the potential for enhanced accuracy (Yan et al., 2024b). For instance, Barzegar et al. (2020) utilized a hybrid CNN-LSTM deep learning model to predict short-term water quality variables, showcasing the benefits of high-frequency data in capturing the rapid dynamics of environmental processes. Similarly, Mozo et al. (2022) developed a chlorophyll soft-sensor based on machine learning models for algal bloom predictions, further emphasizing the efficacy of fine-resolution data. Still, it is crucial to note that the effectiveness of machine learning in HAB prediction heavily relies on the completeness and quality of the data used (Asnaghi et al., 2017). For example, it has been observed that using monthly data for predicting HABs by a single ML algorithm often leads to poor results (Yajima and Derot, 2017). This can be attributed to harmful algae exhibiting rapid and dynamic changes over shorter intervals (Handy et al., 2008). Monthly data may not capture these variations effectively compared to finer data, resulting in less accurate predictions (Jackson-Blake et al., 2022). The limitation of unrepresentative features can hinder the model's ability to capture the complex and nuanced patterns essential for machine learning, which relies on these features to learn and generalize from the data. Given that geoscience data acquisition is often constrained in terms of time and space, exploring research on data processing methods, creating hybrid models, and conducting feature engineering are vital steps to enhance model performance within the bounds of limited data availability (Guo, 2017; Zhang, 2003).

Hybrid models in predictive analytics are an advanced approach that combines different modeling techniques to effectively address the limitations of single-model approaches, enabling more reliable predictions of complex data patterns (Zhu et al., 2023). Time series data can be predicted by combining linear and nonlinear hybrid models, as data comprising both linear, including linear trends and seasonal patterns, and nonlinear parts, like oscillatory patterns and complex interrelationships (Zhang, 2003). Similarly, when considering time series as a signal composed of multiple sub-signals with different frequencies, predictions can be made by summing each sub-signal exhibiting autoregressive characteristics (Zhu et al., 2023). This data decomposition approach to signal processing is an effective strategy for extracting dynamic features from time-series data, which weakens the noise and transforms the signal into a form that can be readily processed and analyzed for subsequent processing (Zuo et al., 2020). Combining temporal decomposition with machine learning (TD-ML) enables the hybrid model to extract underlying information and capture both local and global patterns to enhance prediction performance, and such hybrid models are beginning to be applied to some parameter prediction in geosciences (Chen et al., 2021; Tian et al., 2017). However, the potential of hybrid models for predicting coastal HABs is still underutilized, and a comparative study could reveal the best hybrid modeling techniques for accurate HAB prediction (Zhu et al., 2023).

To address the need for improved coastal water quality management, this study identified several vital gaps from previous studies: the absence of robust and accurate HAB prediction models using monthly data and a lack of comparative analysis among various hybrid models for HAB prediction (Alexandre et al., 2021). The overarching goal is to develop an improved monthly framework to predict HABs and enhance understanding of HAB dynamics in Biscayne Bay. To fill the gaps, we aim to design and compare three scenarios for predicting chlorophyll-a (Chl-a) concentration, an indicator of HABs (Lee et al., 2022): Scenario 1 (S1) single Support Vector Machine (SVM), Scenario 2 (S2) Seasonal Autoregressive Integrated Moving Average (SARIMA) combined with SVM, and Scenario 3 (S3) Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Variational Mode Decomposition (VMD) combined with SVM. Additionally, correlation coefficients will be applied to explore the physical mechanisms of the best-performing scenario. Biscayne Bay, adjacent to Miami-Dade County in South Florida, is chosen as our case study site due to its recent history of HAB-induced environmental issues, such as fish kills and seagrass die-off, which underscore the critical need for understanding and mitigating the consequences of anthropogenic activities on the bay's ecosystem (Alexandre et al., 2021; Santos et al., 2020). The HAB prediction framework developed in this study, which includes a comprehensive flow chart detailed in the 'Methods' section, is designed for Biscayne Bay but is also transferrable in other coastal regions. Its primary purpose is to provide early warning of HABs, offering practical tools for proactive management to protect diverse coastal ecosystems. Additionally, the correlation analysis conducted is expected to improve our understanding of HAB dynamics, assisting stakeholders in Biscayne Bay and potentially in similar environments globally.

2. Methods and materials

2.1. Study area

Biscayne Bay, located east of Miami-Dade County and situated along the southeastern coast of Florida, stretches approximately 97 km from north to south and encompasses an area of about 700 square kilometers (Fig. 1a). The bay is characterized by its intricate network of shallow waters, seagrass beds, and coral reefs. The extensive seagrass beds contribute to the bay's water quality by filtering pollutants and providing a habitat for countless organisms. The downstream Biscayne Bay holds significant importance for upstream Miami-Dade County. The bay has recreational opportunities, contributing to the vibrant tourism industry, and the economic value extends to fisheries, providing sustenance for local communities. However, the delicate balance between urbanization and the environment has challenged maintaining the bay's health. Urban runoff, pollution, and climate change threaten the water quality and marine ecosystems of Biscayne Bay. Recently, fish die-offs and seagrass degradation are the most prominent two threats caused by HABs to the ecosystem and economics in the bay. We selected 11 stations with long time-series water quality records to study the HABs in Biscayne Bay (Fig. 1a).

2.2. Data collection

Our study employed a consistent dataset with uniform features for all stations, utilizing monthly data to develop predictive models from 1997 to 2020. The target variable applied in all predictive models, chlorophyll-a (Chl-a), was collected and analyzed monthly by the Miami-Dade Division of Environmental Resources Management (DERM) using the standardized method SM 10200-H, involving manual sampling at a consistent depth of 0.10 m. The water quality data used in the S1 and S2 predictive models, including ammonia nitrogen (AN), nitrate/nitrite (NOx), dissolved oxygen (DO), pH, water temperature (WT), turbidity (TBD), and total phosphorus (TP), were provided by the DERM collected monthly for all stations. The climate data employed in the S1 and S2 predictive models comprised air temperature (AT), specific humidity (SH), wind speed (WS), precipitation (PR), and shortwave radiation flux (SRF), which were obtained from the North American Land Data Assimilation System (NLDAS) Primary Forcing Data L4 Monthly 0.125 * 0.125° V002, located in Greenbelt, Maryland, USA, at the Goddard Earth Sciences Data and Information Services Center (GES DISC) (Xia et al., 2012). The land use data for 2001, 2004, 2006, 2008, 2011, 2013, 2016, and 2019, applied in the S1 and S2 predictive models, were processed to obtain the developed percent (DP) feature (Dewitz and U.S. Geological Survey, 2021). This metric represents the ratio of developed area to the

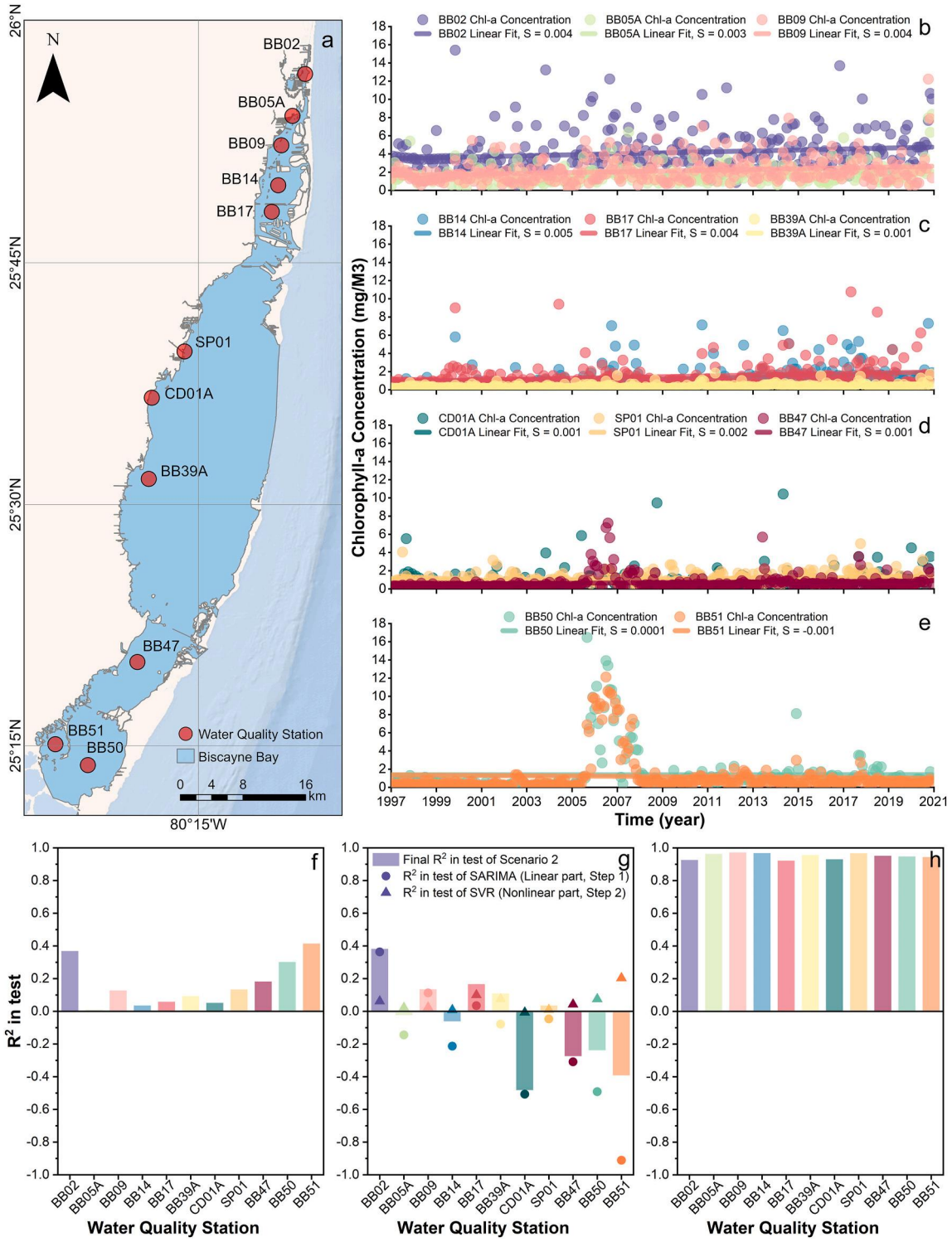

Fig. 1. Overview of water quality stations data and model results. **a** Study area and locations of water quality stations. **b** Monthly chlorophyll-a concentration and its trend of station BB02, BB05A, and BB09. **c** Monthly chlorophyll-a concentration and its trend of station BB14, BB17, and BB39A. **d** Monthly chlorophyll-a concentration and its trend of station CD01A, SP01, and BB47. **e** Monthly chlorophyll-a concentration and its trend of station BB50 and BB51. **f** R-squared results in test datasets of each station for Scenario 1 (S1). **g** R-squared results in test datasets of each station for Scenario 2 (S2). **h** R-squared results in test datasets of each station for Scenario 3 (S3).

total land area excluding open water and wetlands, which are significant in the regions studied. Excluding these areas is crucial to ensure comparability of the data across different watersheds. Because each watershed can have a very different percentage of water and wetlands, including these in the developed area calculation could lead to significant discrepancies when comparing the level of urbanization across regions. We divided Miami-Dade County (upstream of Biscayne Bay) into northern, central, and southern watersheds according to sub-watershed shapefile provided by the South Florida Water Management District (SFWMD). Details on the watershed division are available

in our previous study (Yan et al., 2024a). The DP was calculated for each watershed from the National Land Cover Database (NLCD) data provided by the U.S. Geological Survey. We kept the developed percent uniform across all downstream water quality stations within the same watershed. Each year's DP was applied uniformly across all months of that year. To address gaps in the land use data outside the NLCD-provided years, we employed linear interpolation and polynomial extrapolation methods to fill the data gaps in the study period (1997–2020). All collected data were subjected to rigorous quality assurance and quality control (QA/QC).

2.3. Core model development

All predictive scenarios aimed to predict the next time step, the following month's chlorophyll-a concentration. In our study, we created individual models for each sampling station, employing the dataset unique to each station to predict its chlorophyll-a levels. To develop and compare robust methods for predicting HABs, S1 and S2 utilized various environmental inputs and selected 11 water quality stations in Biscayne Bay based on data availability (Fig. 1a). In Scenario 1 (S1), we employed all variables from the preceding three months to predict the following month's Chl-a concentration by a single SVM model (Fig. 2). In the hybrid Scenario 2 (S2), we initially used SARIMA to predict the linear component of Chl-a time series. We used the SVM model with all variables to predict the residual or nonlinear section of Chl-a, obtained by subtracting the linear part from the original data. Next, we added the predicted nonlinear part to the linear component, resulting in the predicted Chl-a time series (Fig. 2). In the TD-ML hybrid Scenario 3 (S3), we started by applying the CEEMDAN algorithm to decompose the original Chl-a time series into high-frequency IMFs (IMF1 and IMF2) and low-frequency IMFs (all IMFs except IMF1 and IMF2). For the high-frequency IMFs, we further applied the VMD algorithm to perform a secondary decomposition, generating multiple modes. Using PACF analysis, we determined significant time steps as predictors by identifying when partial autocorrelation coefficients for modes and IMFs surpassed the confidence threshold. Finally, we summed all predicted sub-sequences to obtain the ultimate Chl-a time series (Fig. 2).

In developing our models, we adopted a temporal segmentation approach by designating the initial 80% of the chronological data from 1997 to 2015 as the training set, with the remaining 20% from 2016 to 2020 used for testing. As our models are created to predict future HABs based on past observations, employing a temporal split ensures that the training data exclude future information (Deng et al., 2021). This strategy crucially underpins the validity of our models in predicting the subsequent monthly time step, reflecting real-world scenario prediction capabilities and maintaining the temporal continuity essential for accurate prediction (Cerqueira et al., 2020). For S1 and S2, SVM models were trained using three distinct feature selection methods: all features, forward selection, and backward elimination. These feature selection methods iteratively determined the best combination of features, aiming to maximize model performance while minimizing redundancy among the predictors. Our entire codebase and models were developed in Python 3.8.8, with GIS visualizations implemented in ArcGIS Pro 3.0.1. The SVM regressor employed belongs to the 'sklearn.svm.SVR' module from the 'sklearn 1.1.3' library, while SARIMA modeling utilized the 'sm.tsa.arima.model.ARIMA' module from the 'statsmodels 0.13.5'

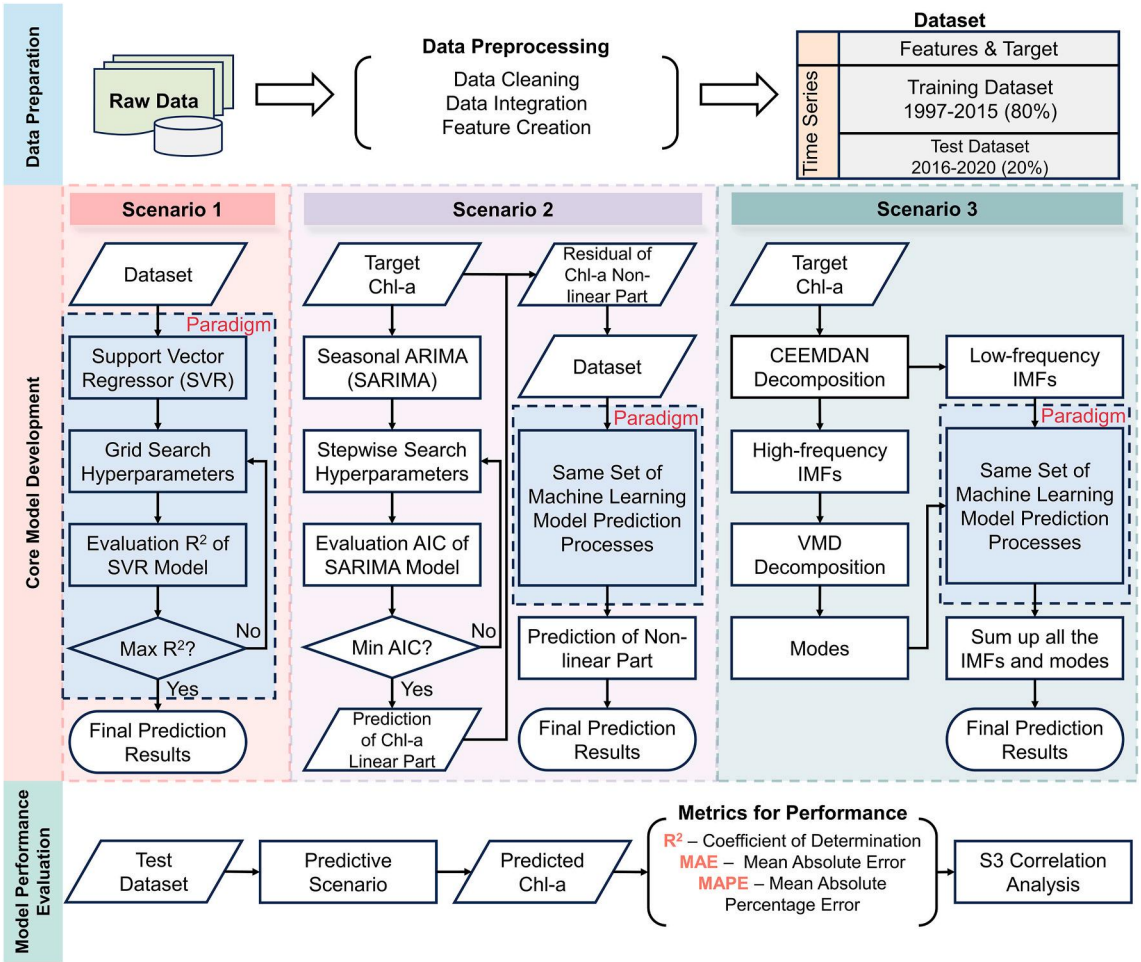

Fig. 2. Model development framework for all scenarios in chlorophyll-a concentration prediction.

library. The SARIMA hyperparameters were tuned using the 'pm.auto_arima' function from the 'pmdarima 2.0.2' library. For the decomposition, CEEMDAN was implemented using the 'PyEMD 1.4.0' library, and VMD employed the 'vmdpy' library.

2.3.1. Support vector regression (SVR)

SVR is a Support Vector Machine (SVM) used for regression challenges (Hearst et al., 1998). It is performed by mapping input features into a higher-dimensional space where a linear regression function is constructed. The SVR aims to find a function that deviates from the actual observed targets y by a value no greater than ϵ for each training example, and at the same time, is as flat as possible to ensure model generalization. It handles non-linear relationships by employing kernel functions, primarily using radial basis functions in our study, which effectively capture complex patterns in environmental data. The formulation for SVR focuses on minimizing the error within a certain threshold ϵ and can be represented as:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{1}$$

subject to:

$$\begin{cases} y_i - \langle w, x_i\rangle - b \le \epsilon + \xi_i, \\ \langle w, x_i\rangle + b - y_i \le \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \ge 0, \end{cases} \tag{2}$$

where w is the weight vector, b is the bias, ξ and ξ^* are slack variables, and C is the regularization parameter.

2.3.2. Seasonal Autoregressive Integrated Moving Average (SARIMA)

SARIMA is an extension of the ARIMA model that includes seasonal terms (Williams and Hoel, 2003). It is particularly useful for modeling time series data with seasonality. The model consists of several parameters: autoregressive (AR) terms, differencing order, moving average (MA) components, and additional seasonal elements. In our study, SARIMA was utilized to model the linear and seasonal components of chlorophyll-a concentrations, capturing underlying patterns that repeat over fixed periods. The SARIMA model is typically noted as SARIMA $(p,d,q)(P,D,Q)_s$, where the non-seasonal part of the model (ARIMA (p,d,q)) and the seasonal part (P,D,Q) are defined as follows:

$$\left(1 - \sum_{i=1}^{p}\varphi_i L^i\right)(1-L^d)\left(1 - \sum_{j=1}^{P}\Phi_j L^{js}\right)(1-L^D)^s X_t = \left(1 + \sum_{k=1}^{q}\theta_k L^k\right)\left(1 + \sum_{m=1}^{Q}\Theta_m L^{ms}\right)\varepsilon_t \tag{3}$$

where L is the lag operator, φ, Φ are the autoregressive terms, θ, Θ are the moving average terms, and d, D are the orders of differencing.

2.3.3. Complete ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

CEEMDAN is an advanced time series decomposition technique that improves upon the traditional Empirical Mode Decomposition (EMD) by adding noise-assisted data analysis (Torres et al., 2011). This method decomposes a time series into intrinsic mode functions (IMFs) that are simple oscillatory modes at different scales. CEEMDAN helps mitigate mode mixing and provides a more stable and representative decomposition, crucial for accurately capturing the multi-scale dynamics of environmental time series. While CEEMDAN is more of an algorithmic procedure than a formula-based one, its core lies in recursively applying noise-assisted data analysis to decompose a signal into IMFs. Each IMF C_i must satisfy two conditions: (1) The number of extrema and zero-crossings must differ at most by one. (2) The mean value of the envelope defined by the local maxima and minima is zero.

2.3.4. Variational Mode Decomposition (VMD)

VMD is a non-recursive technique for decomposing a time series into a predefined number of quasi-orthogonal modes called intrinsic mode functions (Dragomiretskiy and Zosso, 2014). Each mode is smooth and has a compact spectral bandwidth, optimized through an iterative process. VMD is particularly effective in handling non-stationary and non-linear signals, as it allows for extracting high-frequency components and trends without prior assumptions about the data. VMD decomposes a signal into K modes by solving the following constrained variational problem:

$$\min_{\{u_k\}}\left\{\sum_{k=1}^{K}\left\|\delta(t) + \frac{\partial}{\partial t}[(\delta(t) + j\pi t) * u_k(t)]e^{-j\omega_k t}\right\|^2\right\} \tag{4}$$

where $u_k(t)$ are the modes, and ω_k are the center frequencies of the modes, which are also optimized as part of the problem.

2.4. Model performance evaluation

The choice of metrics to evaluate the model performance is predicated on the continuous or categorical nature of the target variable within the study. As our study's continuous chlorophyll-a indicates HABs without a definitive threshold for categorizing HAB occurrences, we employed regression metrics to assess the model performance. After obtaining results from predicting the test dataset using the predictive models, we evaluated all scenarios across various stations using three key metrics: R^2 (Coefficient of Determination), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error). R^2 signifies the amount of target variance clarified by the features, offering an insight into the fit quality and its predictive capability of the created models for unseen data, gauged by the proportion of explained variance. The MAE is a risk measurement representing the anticipated value of absolute error loss or L1-norm loss. The MAPE serves as a regression evaluation metric designed to emphasize relative errors, remaining unaffected by overall scaling adjustments of the target variable. Higher R^2 values and smaller MAE and MAPE values indicate enhanced accuracy and stronger predictive capabilities of the model. To perform these assessments, we employed functions from the sklearn.metrics module, namely r2_score, mean_absolute_error, and mean_absolute_percentage_error (Pedregosa et al., 2011).

2.5. Correlation analysis

We initially employed Pearson correlation coefficients to examine the relationships between chlorophyll-a and its decomposed IMFs and other environmental variables. This method was chosen because it can detect linear associations between variables. It provides a useful overview of the potential connections with chlorophyll-a, which can complement exploratory environmental data analysis after developing high-precision TD-ML predictive models that cannot capture the environmental effects on chlorophyll-a. While Pearson correlation typically assumes that the underlying data distributions are normal, our preliminary assessments revealed that our data were not normally distributed. Despite this, Pearson coefficients were used to give an initial estimate of linear correlations, which are informative for our supplemented environmental relationship analysis. The Pearson correlation coefficient, developed in 'pandas.DataFrame.corrwith' module from the 'pandas 1.4.2' library, is a commonly used tool to describe the correlation between two variables. The correlation coefficient ranges from −1 to 1, with 0 indicating no correlation. The closer the absolute value is to 1, the stronger the correlation, where negative values indicate a negative correlation and positive values show a positive correlation.

3. Results and discussion

3.1. Limitations and challenges in HAB prediction using single machine learning models

Chl-a concentrations and upward trends were higher in the north bay than in other areas during the study period. The average Chl-a concentration recorded across all sampling stations from 1997 to 2020 was 1.37 μg/L. Notably, the most northern station BB02 exhibited the highest average concentration during this period, recording 4.12 μg/L. Most sites exhibited an increasing trend in Chl-a concentrations, except for BB51, while the north bay displayed particularly pronounced trends, with slopes exceeding 0.003 (Fig. 1b, c, 1d, 1e). Given the presence of extreme values of Chl-a concentrations and the lack of significant autoregressive characteristics, it is challenging to rely solely on conventional univariate time series forecasting methods for Chl-a concentration prediction. To improve the prediction accuracy of the HABs, Chl-a concentration in this study, we incorporated related environmental variables, such as ammonia nitrogen, water temperature, NOx, total phosphorous, turbidity, dissolved oxygen, pH, air temperature, wind speed, shortwave radiation, specific humidity, precipitation, and developed percent, to the S1 and S2 predictive models (Ly et al., 2021). Fig. 1f revealed that the single SVM prediction models did not perform satisfactorily in predicting HABs in Biscayne Bay. None of the stations achieved an R^2 exceeding 0.5, with BB51 having the highest at 0.41 and BB05A the lowest at 0.01. Additionally, almost all stations had a Mean Absolute Percentage Error (MAPE) exceeding 30%, with BB39A having the lowest at 32% and CD01A the highest at 85% (Table 1).

Machine learning (ML) has been extensively applied in geosciences and environmental science research fields, showing promise in predicting HABs (Bergen et al., 2019). However, regarding HAB prediction, ML model applications to monthly data have been limited in existing research. While researchers often interpolate monthly time series data to a finer scale for better accuracy, this may introduce errors (Deng et al., 2021; Lepot et al., 2017). Alternatively, modeling monthly data directly may not precisely capture the rapid life cycles of algal species (Silva et al., 2023; Yajima and Derot, 2017). These challenges highlight the need for effective HAB prediction strategies, particularly with sparse data. As a starting point, we applied S1 single ML models to evaluate their effectiveness in HAB prediction without the uncertainties of extensive interpolation.

The limitations of low prediction accuracy in S1 underscore the challenges of predicting HABs and provide valuable support for the complexities of the environmental factors influencing Chl-a concentration. Key challenges include the difficulty of modeling sporadic algal growth events with extreme Chl-a values, as conventional ML methods struggle to predict these irregularities accurately across various aquatic environments (Qi and Majda, 2020; Yajima and Derot, 2017). Also, sparse monthly data fails to capture the rapid life cycles of diverse algal species, including various algal species such as chlorophytes, cyanobacteria, and diatoms, in Biscayne Bay (Alexandre et al., 2021; Wachnicka et al., 2020). Studies found that these species often have relatively short life cycles, and bloom events can last from a few days to several weeks or months that do not align perfectly with monthly intervals, leading to inaccuracies in HAB predictions (Pokrzywinski et al., 2022; Silva et al., 2023). Additionally, the lack of significant autoregressive characteristics in Chl-a data, implying that Chl-a might not exhibit strong temporal dependencies, hampers the model's capabilities to utilize past observations effectively. Our previous research in Biscayne Bay indicated that most stations exhibited weak autoregressive characteristics, contributing to the lower prediction accuracy (Yan et al., 2024a). Last, critical features inadequacies limit the models' ability to fully capture the complex HAB dynamics. The complexity of algal growth in aquatic ecosystems involves various interacting factors, posing significant challenges in our study due to limited data availability (Wells et al., 2020). The scarcity of hydrodynamic and biological data and a small sample size hindered our algorithm's learning of Chl-a concentration patterns, underscoring the need for comprehensive data collection to improve predictive model effectiveness in complex environmental studies like HAB prediction (Xia et al., 2020). In particular, the absence of salinity data, a critical factor in the HAB dynamics, notably impacted the S1 scenario's predictive accuracy (Wells et al., 2020). As Biscayne Bay is influenced by various inflows from the upstream rivers and canals, these sources modify the salinity gradients within the bay, especially affecting algal bloom dynamics (Chin, 2020). The lack of consistent salinity measurements across all stations, due to limitations in the data collection, meant that our models could not account for this variability. All these absence of hydrodynamic and biological data (such as algae predators) and inadequate temporal resolution of datasets likely contributed to the less robust predictive performance observed, as evidenced by the low R^2, high MAE, and high MAPE values across most stations, indicating a need for more sophisticated and integrated modeling approaches for accurate HAB prediction.

3.2. Evaluation of SARIMA and SVM hybrid models for HAB prediction

Fig. 1g showed that the hybrid models combining SARIMA and SVM failed to predict Chl-a concentration accurately, and the results exhibited significant differences. In the first stage of S2, the R^2 of the SARIMA models, except for BB02, BB09, and BB17, was negative. The R^2 of SVM models was close to zero in the second stage. When combining both, the R^2 results at all stations remained below 0.4, with BB02 achieving the highest at 0.38 and CD01A the lowest at −0.48. The MAPE for all stations exceeded 30%, with BB39A having the lowest at 32% and BB50 the highest at 80% (Fig. 1g and Table 1). Interestingly, except for a few stations, the performance of S1 was even superior to that of S2. Given the less-than-optimal results with S1 and S2, we pursued further enhancements by introducing an additional scenario, S3, which integrates temporal decomposition and machine learning.

Despite integrating linear and nonlinear components, the S2 hybrid models underperformed in predicting Chl-a concentrations, with most stations showing negative R^2 values in the first stage using SARIMA

Table 1
Metrics of the prediction results of test sets for all scenarios and water quality stations.

Stations	S1 R^2	S1 MAE	S1 MAPE	S2 R^2	S2 MAE	S2 MAPE	S3 R^2	S3 MAE	S3 MAPE
BB02	0.37	1.51	0.40	0.38	1.38	0.37	0.93	0.51	0.16
BB05A	0.01	1.03	0.44	−0.02	1.10	0.55	0.96	0.22	0.13
BB09	0.13	1.05	0.59	0.13	1.08	0.58	0.97	0.22	0.13
BB14	0.04	0.74	0.48	−0.06	0.86	0.60	0.97	0.17	0.13
BB17	0.06	1.07	0.65	0.17	1.07	0.73	0.92	0.40	0.27
BB39A	0.09	0.17	0.32	0.11	0.17	0.32	0.96	0.05	0.12
CD01A	0.05	0.65	0.85	−0.48	0.72	0.75	0.93	0.19	0.24
SP01	0.13	0.53	0.44	0.04	0.55	0.53	0.97	0.11	0.10
BB47	0.18	0.29	0.36	−0.27	0.43	0.62	0.95	0.10	0.19
BB50	0.30	0.41	0.46	−0.24	0.61	0.80	0.95	0.12	0.16
BB51	0.41	0.22	0.45	−0.39	0.32	0.57	0.94	0.07	0.13

models. This inadequacy might stem from the complex algal dynamics and the non-obvious seasonal patterns in Biscayne Bay (Gobler, 2020; Xia et al., 2019). In contrast to other successful prediction applications, consistent periodicity datasets, such as the sunspot and Canadian lynx datasets, aid in accurate S2-similar predictions (Belmahdi et al., 2020; Júnior et al., 2019), Chl-a time series are characterized by high variability and unpredictable environmental influences (Millette et al., 2019). These substantial disturbances from external factors obscure underlying periodicity, weakening the S2 models' capacity for effective prediction. The unsuccessful linear modeling in S2's first stage led to meaningless residuals in the second stage, reflecting the overall inadequacy of SARIMA and SVM components in this context. These findings highlight the necessity for advanced and continuously refined modeling techniques in hybrid approaches to accurately predict HABs, particularly for complex environmental datasets without apparent periodicity. Additionally, given the challenges encountered, the S2 predictive framework is not recommended for other complex time series with similar irregularities.

3.3. Superior performance of TD-ML hybrid model in HAB prediction

The limitations of the S1 and S2 predictive models led us to adopt a novel hybrid approach involving two-stage temporal decomposition of Chl-a concentration time series into IMFs and modes, coupled with ML algorithm fitting and aggregation for prediction. This method enhances prediction accuracy by capturing the complex dynamics of Chl-a concentrations by temporal decomposition (see decomposition results in *SI* Figs. S1, S2, S3), and utilizes SVM for fitting each nonlinear smoother IMF, showcasing its effectiveness even in data with extreme values and noise (Ding et al., 2022; Potnuri et al., 2023). The S3 framework,

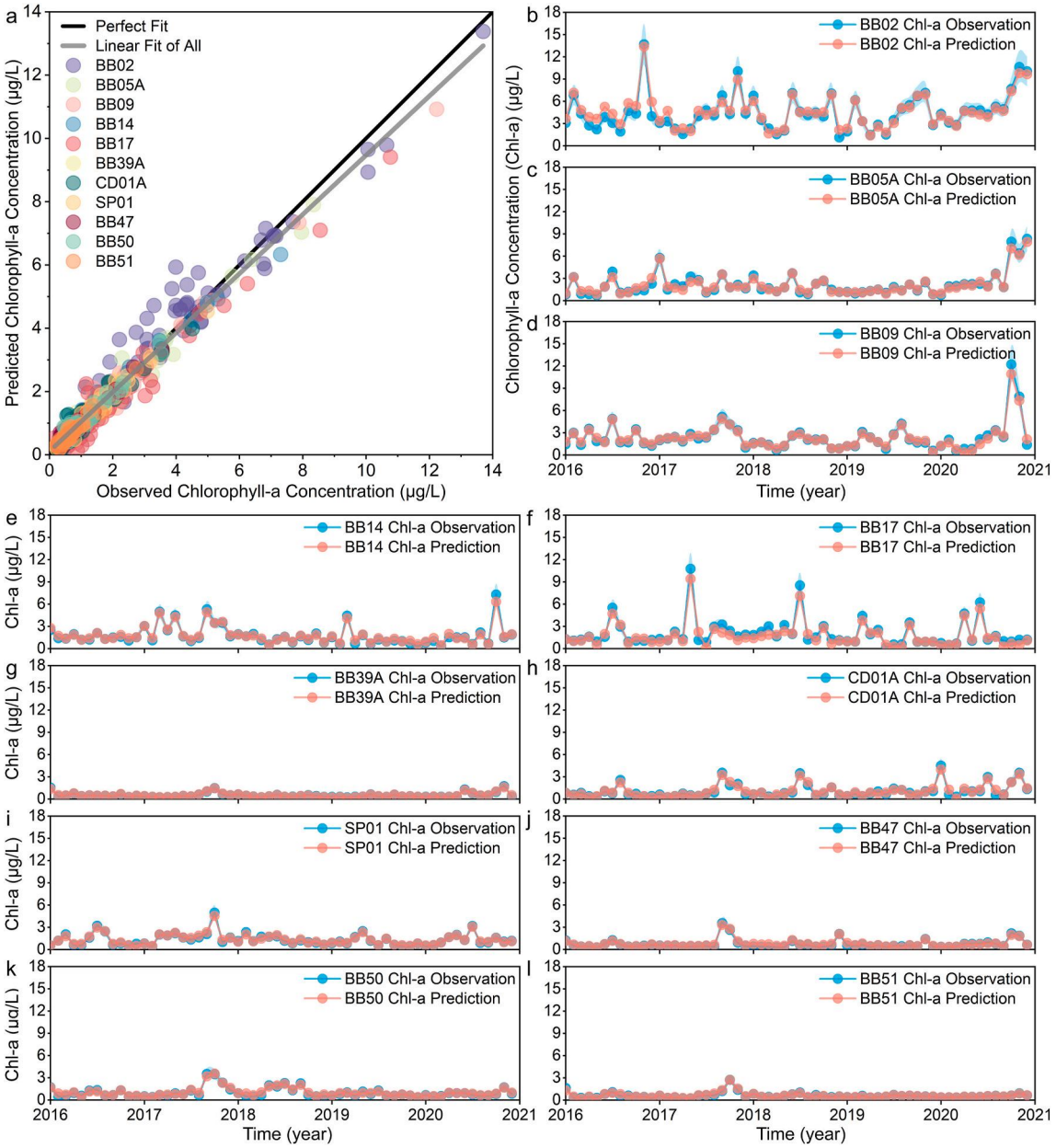

Fig. 3. Prediction results of the test dataset for each water quality station in Scenario 3 (S3). a Test dataset scatter plots of observation chlorophyll-a concentration (x-axis) versus prediction chlorophyll-a concentration (y-axis) for all water quality stations. The sample point lies on the perfect fit when the prediction is identical to the observation. **b-l** Observation and prediction time-series plots in the 2016–2020 test period for each water quality station. The blue shading is the ±20% error range of the observation data.

requiring minimal data and avoiding additional environmental variables, emerges as a promising, cost-effective solution for predicting and managing HABs.

Fig. 1h and 3 demonstrated the exceptional performance of the hybrid model in S3 for predicting HABs at all stations, surpassing the accuracy achieved in S1 and S2. The R^2 values consistently exceeded 0.9 in S3, with BB09, BB14, and SP01 reaching the highest at 0.97 and BB17 achieving the lowest at 0.92 (Fig. 1h). Furthermore, the MAPE remained below 30% in S3, with SP01 having the lowest at 0.10 and BB17 the highest at 0.27 (Table 1). The remarkable precision of the S3 hybrid model was further substantiated by Fig. 3, where all the predicted test samples exhibited an impressive fit to the observed values, indicating high accuracy in predicting both moderate and extreme Chl-a concentrations. This outcome solidified the robust predictive capability of the hybrid models within the S3 prediction framework across all stations in Biscayne Bay, which significantly surpasses that of the previous related study (Yan et al., 2024a).

TD-ML hybrid models are gaining traction in various fields, including geosciences, for applications like runoff and rainfall prediction (Chen et al., 2021; Unnikrishnan and Jothiprakash, 2020). Despite limited research in HAB prediction, our study demonstrates that the TD-ML hybrid model significantly outperforms earlier models, offering robust predictions for Chl-a concentrations (Table 1). Note that our TD-ML approach leverages the inherent dynamics of monthly data without resorting to extensive data interpolation techniques, which may often introduce errors (Li et al., 2014). Unlike some studies that may generate frequent data points through interpolation from sparser datasets, our method respects the original monthly data frequency. This approach helps avoid creating pseudo-regular patterns that may not exist, ensuring that our model's high accuracy reflects its ability to discern genuine environmental dynamics from the real data provided (Lepot et al., 2017). Consequently, this method enhances the reliability and applicability of our predictions in real-world scenarios, where true data variability and unpredictability are common.

Our S3 results, exhibiting robust predictive performance in Chl-a prediction, are consistent with a similar study in the coastal areas of Hong Kong (Zhu et al., 2023). This successful application demonstrates that the efficacy of TD-ML, which is particularly effective for predicting time series characterized by extremes, noise, and subtle linear and seasonal trends - common challenges in HAB prediction (Khan et al., 2020). The novel hybrid model described by Yu et al. (2024), based on two-stage data processing and machine learning, further supports our findings, demonstrating significant improvements in forecasting Chl-a in reservoirs. Similarly, the ensemble deep learning model developed by Zhang et al. (2023), which employs a two-layer decomposition and attention mechanisms, validates our model's capacity to handle complex datasets effectively. TD-ML represents a promising approach for predicting complex time series in geosciences, combining various techniques to enhance insights into intricate temporal dynamics (Zhang et al., 2024). By leveraging predictive frameworks similar to those recent innovative studies, the TD-ML model can forecast the occurrences of HABs and offer potential applications in broader environmental management contexts (Wang et al., 2023). These include integrating predictive models with real-time monitoring systems to enable proactive measures and deploying them in varied ecological settings to tailor specific mitigation strategies based on predicted bloom dynamics. The continued exploration and refinement of these models will be crucial for effectively advancing our ability to manage and mitigate the impact of HABs.

3.4. Physical mechanisms interpretation of hybrid models in scenario 3

Environmental factors significantly influence HABs and are crucial for understanding their temporal dynamics (Yu et al., 2021). Factors, such as nutrient levels, water temperature, and salinity, directly affect algal growth, a key aspect in HAB prediction (Anderson et al., 2021). This study primarily involves models developed independently for each station. Particularly in the case of traditional ML modeling from S1, our results were unfavorable, with most station models showing low accuracy. Thus, we did not focus on feature importance in models with poor accuracy because this could lead to biased outcomes. Our novel TD-ML approach involved decomposing the Chl-a into less interpretable sub-sequences for machine learning fitting to enhance prediction accuracy. This strategy involved a trade-off, potentially sacrificing the limited interpretability of ML methods. To address this limitation, we preliminarily explored the physical implications of our S3 models by analyzing the correlation between Chl-a and its IMFs and other environmental variables (Zhu et al., 2023). We acknowledge that the dynamics of HABs exhibit significant nonlinearity, so this analysis is just a glance, capturing only the more apparent linear relationships. This investigation may provide insights into protecting Biscayne Bay and similar ecosystems.

The correlation matrix reveals a generally weak association between most environmental variables and Chl-a concentration, with some climate factors like water and air temperature, specific humidity, and precipitation showing positive but weak correlations, suggesting their potential role in exacerbating HABs (Fig. 4). The dual-directional correlations between inorganic nitrogen and shortwave radiation with Chl-a signified the ecosystem complexity. Specifically, excess nitrogen might incite algal growth while potentially altering the dominance of species by shifting phosphorus or silicon to limiting nutrients, thus suppressing algal growth (Medina et al., 2022; Wang et al., 2021). The convoluted correlation of shortwave radiation could unveil how elevated levels of shortwave radiation either stimulate (below the light saturation point) or inhibit (above the saturation point) algal growth (Fu et al., 2012). The positive correlations of total phosphorus and developed percent with Chl-a concentration at most stations might reflect the impacts of human activities on downstream coastal environments by enhancing nutrient influx, thereby stimulating algal proliferation (Glibert, 2020). In the oligotrophic Biscayne Bay, phosphorus levels are vital in phytoplankton growth (Millette et al., 2019). The rising levels of phosphorus, coupled with an increase in developed percent, are likely to augment Chl-a concentration, aligning with our previous research findings (Yan et al., 2024a). In our previous research, we developed a predictive model using aggregated data from all stations (instead of individual models for each station). The results were slightly better than S1, and based on SHAP values, we found that developed percentage and total phosphorus were the most important positive predictors of current chlorophyll-a levels. However, the previous study's model was nonlinear, and the linear relationships shown in Fig. 4's correlation analysis did not reveal such strong connections. Climate change tends to increase the frequency and severity of many HABs in the future (Glibert, 2020). The negative correlations of pH with Chl-a across most stations may indicate the acidic environmental preference of local algae. Moreover, the positive correlations of climatic factors – water temperature, air temperature, relative humidity, and precipitation – with Chl-a at all stations reflected the boosting influence of climate change on algal growth, contributing to more favorable conditions for algal growth (Wells et al., 2020). Coupled with the fact that bay algae favor acidic environments, it is conceivable that climate change could intensify the HABs in Biscayne Bay.

IMF1 and IMF2 represented high-frequency sub-sequences that contained large-amplitude rapid oscillations and noise, making it challenging to establish their correlation with environmental variables. The correlation coefficients between all variables and IMF1 and IMF2 were below 0.4 at all stations, indicating a weak association (Fig. 4). Other moderate and low-frequency IMFs also displayed unexplainable periodicity, showing little correlation with environmental variables. The lowest-frequency IMFs primarily depict linear trends (Karijadi and Chou, 2022). They illustrated strong association and high correlation coefficients with pH (mostly negative) and developed percent (mostly positive). These results indicated that for most stations, as the water environment became more acidic and the developed percent of the

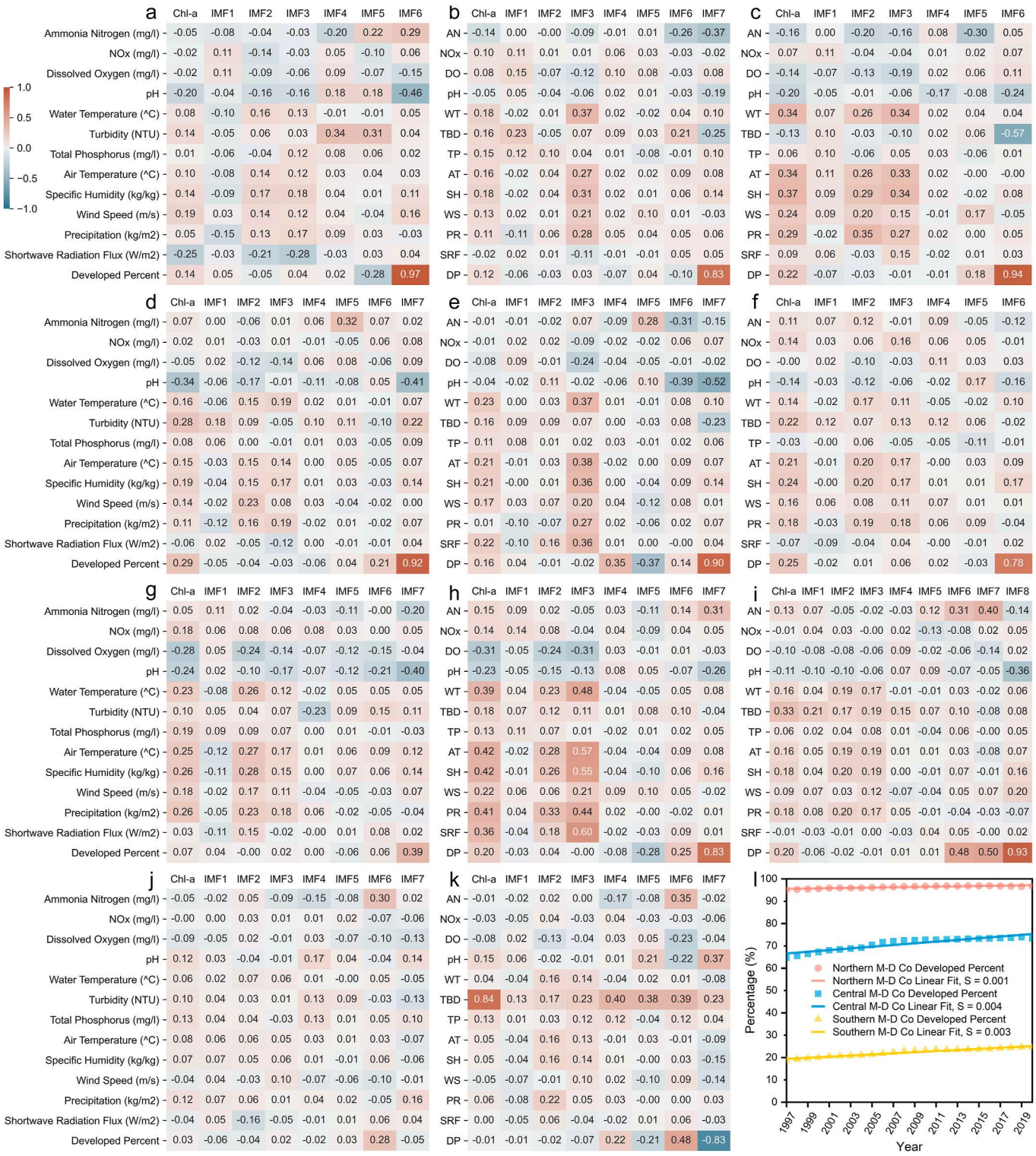

Fig. 4. The heatmaps of correlation analysis for each station and the developed percent change of upstream Biscayne Bay. a-k Correlation heatmaps between environmental variables and chlorophyll-a and its IMFs in BB02, BB05A, BB09, BB14, BB17, BB39A, CD01A, SP01, BB47, BB50, and BB51, respectively. l Developed percent of northern, central, and southern watersheds of Miami-Dade County, upstream of Biscayne Bay. The developed percent represented the developed area divided by the total area, and we excluded open water and wetlands to ensure comparability across watersheds. The northern, central, and southern watersheds were divided by the combination of sub-watersheds provided by the South Florida Water Management District (SFWMD).

upstream land use increased, Chl-a tended to rise.

Fig. 4l shows an upward trend in urbanization across Miami-Dade County from 1997 to 2020, with the northern watershed increasing slightly from 95.1% to 96.7%, the central watershed more rapidly from 64.7% to 73.5%, and the southern watershed from 18.7% to 24.5%. The results indicated that Miami-Dade County was undergoing rapid urbanization, particularly with faster growth rates in the southern and central watersheds compared to the northern. This might be attributed to the fact that north Miami had already been almost entirely urbanized. Increased nutrient inputs into the bay due to urbanization may affect

water quality and chlorophyll-a concentrations or HABs. Fig. 4 showed that, except for the BB50 and BB51, there was a positive correlation between Chl-a concentration trends (represented by the lowest-frequency IMFs) and the developed percent at almost all stations. This suggested that as the level of development increased, Chl-a concentration tended to rise. This correlation might be caused by the impacts of human activities, such as growing air pollution aerosol and wastewater discharge on the water environment.

Given the high accuracy of the TD-ML predictive framework, decomposing the target (chlorophyll-a in this paper) appears to be a beneficial approach. Therefore, we hypothesize that with data that have a finer temporal frequency, such as daily or hourly, and can better capture HAB dynamics, we should be able to optimize subsequent predictive models by first decomposing the target and establishing a correlation matrix with its IMFs. When environmental variables show strong linear relationships with the target itself and IMFs, representing seasonality or trends, we should intensify our observation and interpretation of these variables. For instance, it may be necessary to ensure that feature selection methods will not eliminate these variables. Or we could explore and incorporate additional environmental features related to those with strong linear relationships to enhance the model's predictive capacity. Although the TD-ML model in this paper does not depend on environmental data, traditional machine learning predictive models could greatly benefit from this approach.

4. Conclusions

This research focused on predicting HABs in Biscayne Bay using a novel hybrid modeling approach. Traditional machine learning methods and SARIMA-ML cannot accurately predict Chl-a. We utilized a unique temporal decomposition and machine learning (TD-ML) hybrid model. This approach effectively decomposed Chl-a time series into subsequences, improving prediction accuracy remarkably. Subsequently, the study identified the influence of multiple environmental factors on Chl-a and its IMFs. We proposed the potential increasing trends of HABs in the bay under the global warming scenario with urbanization. The comparative findings highlighted the need for advanced modeling techniques to predict HABs and the potential impacts of human activities and climate change on algal outbreaks. Overall, the study provided valuable insights into predicting and managing HABs, offering a robust predictive framework for understanding complex temporal dynamics and safeguarding coastal ecosystems against HABs. Furthermore, this predictive framework holds the prospect of being applied to coastal regions worldwide, aiding in the prediction of challenging HABs.

CRediT authorship contribution statement

Zhengxiao Yan: Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Nasrin Alamdari:** Writing – review & editing, Resources, Project administration, Methodology, Funding acquisition.

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

Data will be made available on request.

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.jenvman.2024.121463.

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., 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. Harmful Algae, Global Harmful Algal Bloom Status Reporting 102, 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.

Barzegar, R., Aalami, M.T., Adamowski, J., 2020. Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. Stoch. Environ. Res. Risk Assess. 34, 415–433. https://doi.org/10.1007/s00477-020-01776-2.

Belmahdi, B., Louzazni, M., Bouardi, A.E., 2020. A hybrid ARIMA–ANN method to forecast daily global solar radiation in three different cities in Morocco. Eur. Phys. J. Plus 135, 925. https://doi.org/10.1140/epjp/s13360-020-00920-9.

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.

Cerqueira, V., Torgo, L., Mozetič, I., 2020. Evaluating time series forecasting models: an empirical study on performance estimation methods. Mach. Learn. 109, 1997–2028. https://doi.org/10.1007/s10994-020-05910-7.

Chen, S., Ren, M., Sun, W., 2021. Combining two-stage decomposition based machine learning methods for annual runoff forecasting. J. Hydrol. 603, 126945 https://doi.org/10.1016/j.jhydrol.2021.126945.

Chin, D.A., 2020. Source identification of nutrient impairment in north Biscayne bay, Florida, USA. J. Environ. Eng. 146, 04020101 https://doi.org/10.1061/(ASCE)EE.1943-7870.0001786.

Dai, Y., Yang, S., Zhao, D., Hu, C., Xu, W., Anderson, D.M., Li, Y., Song, X.-P., Boyce, D. G., Gibson, L., Zheng, C., Feng, L., 2023. Coastal phytoplankton blooms expand and intensify in the 21st century. Nature 615, 280–284. https://doi.org/10.1038/s41586-023-05760-y.

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., Geological Survey, U.S., 2021. National land cover Database (NLCD) 2019 products. https://doi.org/10.5066/P9KZCM54.

Ding, S., Zhang, H., Tao, Z., Li, R., 2022. Integrating data decomposition and machine learning methods: an empirical proposition and analysis for renewable energy generation forecasting. Expert Syst. Appl. 204, 117635 https://doi.org/10.1016/j.eswa.2022.117635.

Dragomiretskiy, K., Zosso, D., 2014. Variational mode decomposition. IEEE Trans. Signal Process. 62, 531–544. https://doi.org/10.1109/TSP.2013.2288675.

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. Global ecology and oceanography of harmful algal blooms 359–377.

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.

Glibert, P.M., 2020. Harmful algae at the complex nexus of eutrophication and climate change. Harmful Algae, Climate change and harmful algal blooms 91, 101583. https://doi.org/10.1016/j.hal.2019.03.001.

Gobler, C.J., 2020. Climate change and harmful algal blooms: insights and perspective. Harmful Algae, Climate change and harmful algal blooms 91, 101731. https://doi.org/10.1016/j.hal.2019.101731.

Guo, H., 2017. Big Earth data: a new frontier in Earth and information sciences. Big Earth Data 1, 4–20. https://doi.org/10.1080/20964471.2017.1403062.

Hallegraeff, G.M., Anderson, D.M., Belin, C., Bottein, M.-Y.D., Bresnan, E., Chinain, M., Enevoldsen, H., Iwataki, M., Karlson, B., McKenzie, C.H., Sunesen, I., Pitcher, G.C., Provoost, P., Richardson, A., Schweibold, L., Tester, P.A., Trainer, V.L., Yñiguez, A. T., Zingone, A., 2021. Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts. Commun Earth Environ 2, 1–10. https://doi.org/10.1038/s43247-021-00178-8.

Handy, S.M., Demir, E., Hutchins, D.A., Portune, K.J., Whereat, E.B., Hare, C.E., Rose, J. M., Warner, M., Farestad, M., Cary, S.C., Coyne, K.J., 2008. Using quantitative real-time PCR to study competition and community dynamics among Delaware Inland Bays harmful algae in field and laboratory studies. Harmful Algae 7, 599–613. https://doi.org/10.1016/j.hal.2007.12.018.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intell. Syst. Their Appl. 13, 18–28. https://doi.org/10.1109/5254.708428.

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.

Júnior, D.S., de Oliveira, J.F., de Mattos Neto, P.S., 2019. An intelligent hybridization of ARIMA with machine learning models for time series forecasting. Knowl. Base Syst. 175, 72–86.

Karijadi, I., Chou, S.-Y., 2022. A hybrid RF-LSTM based on CEEMDAN for improving the accuracy of building energy consumption prediction. Energy Build. 259, 111908 https://doi.org/10.1016/j.enbuild.2022.111908.

Khan, MdM.H., Muhammad, N.S., El-Shafie, A., 2020. Wavelet based hybrid ANN-ARIMA models for meteorological drought forecasting. J. Hydrol. 590, 125380 https://doi.org/10.1016/j.jhydrol.2020.125380.

Lee, D., Kim, M., Lee, B., Chae, S., Kwon, S., Kang, S., 2022. Integrated explainable deep learning prediction of harmful algal blooms. Technol. Forecast. Soc. Change 185, 122046. https://doi.org/10.1016/j.techfore.2022.122046.

Lepot, M., Aubin, J.-B., Clemens, F.H.L.R., 2017. Interpolation in time series: an introductive overview of existing methods, their performance criteria and uncertainty assessment. Water 9, 796. https://doi.org/10.3390/w9100796.

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.

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.

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.

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. Coast 42, 1792–1803. https://doi.org/10.1007/s12237-019-00610-5.

Mozo, A., Morón-López, J., Vakaruk, S., Pompa-Pernía, Á.G., González-Prieto, Á., Aguilar, J.A.P., Gómez-Canaval, S., Ortiz, J.M., 2022. Chlorophyll soft-sensor based on machine learning models for algal bloom predictions. Sci. Rep. 12, 13529 https://doi.org/10.1038/s41598-022-17299-5.

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.

Pokrzywinski, K., Johansen, R., Reif, M., Bourne, S., Hammond, S., Fernando, B., 2022. Remote sensing of the cyanobacteria life cycle: a mesocosm temporal assessment of a Microcystis sp. bloom using coincident unmanned aircraft system (UAS) hyperspectral imagery and ground sampling efforts. Harmful Algae 117, 102268. https://doi.org/10.1016/j.hal.2022.102268.

Potnuri, R., Rao, C.S., Surya, D.V., Kumar, A., Basak, T., 2023. Utilizing support vector regression modeling to predict pyro product yields from microwave-assisted catalytic co-pyrolysis of biomass and waste plastics. Energy Convers. Manag. 292, 117387 https://doi.org/10.1016/j.enconman.2023.117387.

Qi, D., Majda, A.J., 2020. Using machine learning to predict extreme events in complex systems. Proc. Natl. Acad. Sci. USA 117, 52–59. https://doi.org/10.1073/pnas.1917285117.

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.

Silva, E., Counillon, F., Brajard, J., Pettersson, L.H., Naustvoll, L., 2023. Forecasting harmful algae blooms: application to Dinophysis acuminata in northern Norway. Harmful Algae 126, 102442. https://doi.org/10.1016/j.hal.2023.102442.

Tian, W., Liao, Z., Zhang, J., 2017. An optimization of artificial neural network model for predicting chlorophyll dynamics. Ecol. Model. 364, 42–52. https://doi.org/10.1016/j.ecolmodel.2017.09.013.

Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P., 2011. A complete ensemble empirical mode decomposition with adaptive noise. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4144–4147.

Unnikrishnan, P., Jothiprakash, V., 2020. Hybrid SSA-ARIMA-ANN model for forecasting daily rainfall. Water Resour. Manag. 34, 3609–3623. https://doi.org/10.1007/s11269-020-02638-w.

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. Coast 43, 1217–1234. https://doi.org/10.1007/s12237-019-00592-4.

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, L., Xie, M., Pan, M., He, F., Yang, B., Gong, Z., Wu, X., Shang, M., Shan, K., 2023. Improved deep learning predictions for chlorophyll fluorescence based on decomposition algorithms: the importance of data preprocessing. Water 15, 4104. https://doi.org/10.3390/w15234104.

Wells, M.L., Karlson, B., Wulff, A., Kudela, R., Trick, C., Asnaghi, V., Berdalet, E., Cochlan, W., Davidson, K., De Rijcke, M., Dutkiewicz, S., Hallegraeff, G., Flynn, K.J., Legrand, C., Paerl, H., Silke, J., Suikkanen, S., Thompson, P., Trainer, V.L., 2020. Future HAB science: directions and challenges in a changing climate. Harmful Algae, Climate change and harmful algal blooms 91, 101632. https://doi.org/10.1016/j.hal.2019.101632.

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.

Williams, B.M., Hoel, L.A., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results. J. Transport. Eng. 129, 664–672. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:6(664.

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.

Xia, R., Zhang, Yuan, Wang, G., Zhang, Yongyong, Dou, M., Hou, X., Qiao, Y., Wang, Q., Yang, Z., 2019. Multi-factor identification and modelling analyses for managing large river algal blooms. Environ. Pollut. 254, 113056 https://doi.org/10.1016/j.envpol.2019.113056.

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.

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.

Yan, Z., Kamanmalek, S., Alamdari, N., 2024a. Predicting coastal harmful algal blooms using integrated data-driven analysis of environmental factors. Sci. Total Environ. 912, 169253 https://doi.org/10.1016/j.scitotenv.2023.169253.

Yan, Z., Kamanmalek, S., Alamdari, N., Nikoo, M.R., 2024b. Comprehensive insights into harmful algal blooms: a review of chemical, physical, biological, and climatological influencers with predictive modeling approaches. J. Environ. Eng. 150, 03124002 https://doi.org/10.1061/JOEEDU.EEENG-7549.

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

Yu, W., Wang, X., Jiang, X., Zhao, R., Zhao, S., 2024. A novel hybrid model based on two-stage data processing and machine learning for forecasting chlorophyll-a concentration in reservoirs. Environ. Sci. Pollut. Res. 31, 262–279. https://doi.org/10.1007/s11356-023-31148-6.

Zhang, C., Zou, Z., Wang, Z., Wang, J., 2023. Ensemble deep learning modeling for Chlorophyll-a concentration prediction based on two-layer decomposition and attention mechanisms. Acta Geophys. https://doi.org/10.1007/s11600-023-01240-z.

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0.

Zhang, L., Wang, C., Hu, W., Wang, X., Wang, H., Sun, X., Ren, W., Feng, Y., 2024. Dynamic real-time forecasting technique for reclaimed water volumes in urban river environmental management. Environ. Res. 248, 118267 https://doi.org/10.1016/j.envres.2024.118267.

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.

Zuo, G., Luo, J., Wang, N., Lian, Y., He, X., 2020. Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting. J. Hydrol. 585, 124776.