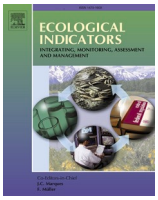
Ecological Indicators 123 (2021) 107334

Contents lists available at ScienceDirect

Ecological Indicators

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



Predicting coastal algal blooms with environmental factors by machine learning methods

Peixuan Yu, Rui Gao *, Dezhen Zhang, Zhi-Ping Liu *

School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China

ARTICLE INFO

Keywords:
Harmful algal bloom
Machine learning
Feature selection
GBDT
Feature importance

ABSTRACT

Harmful algal blooms are a major type of marine disaster that endangers the marine ecological environment and human health. Since the algal bloom is a complex nonlinear process with time characteristics, traditional statistical methods often cannot provide good predictions. In this paper, we propose a method based on machine learning with the aim to predict the occurrence of algal blooms by environmental parameters. The features related to algal bloom growth have been experimented for achieving a good prediction of algal concentrations by a combination strategy. We validate the prediction performance on two real datasets from two locations in US and China, i.e., Scripps Pier, California and Sishili Bay, Shandong, respectively. The models and feature subsets have been selected to complete the missing data and predict the phytoplankton concentration. The results demonstrate the efficiency of our method in the short-term prediction of concentrations by selecting appropriate features. The comparison studies prove the advantage of our developed machine learning method. The importance of every features for the prediction performance reveals crucial factors for the outbreak of harmful algal blooms.

1. Introduction

Algal bloom in marine water system is a phenomenon of rapid outbreak of phytoplankton concentration (Sellner et al., 2003). As a food of many aquatic animals, phytoplankton is an indispensable part of the ocean ecosystem with significant importance. However, when a large amount of planktonic algae proliferate over a certain limit, harmful algal blooms (HABs) will occur. HABs not only cause huge economic losses of aquaculture and tourism, but also endanger human health (Hallegraeff, 1993). In the past few decades, the problem of HABs has shown be with a clear expansion trend worldwide (Anderson et al., 2002). Since the earliest documented in 1746, the southern California bight in USA has been experiencing irregular red tides, and many occasions have been accompanied with the observations of fish and shellfish deaths (Gregorio and Pieper, 2000). After the 1980s, the number of HABs in China's coastal waters has been increasing, which caused dramatic economic losses (Tang et al., 2006). Sishili bay in China, located in Yantai city and adjacent to the North Yellow Sea, is an important marine aquaculture area in Shandong province, China. In recent years, several serious algal blooms have occurred, which resulted in big impacts on the local aquaculture industry (Hao et al., 2011). Therefore, it is of prominent interest to understand the environmental factors affecting the outbreak of algal blooms and make precise predictions by building up quantitative models on them.

Ecological dynamic model is expected to provide accurate equations which can be employed to predict the temporal and spatial distribution of organisms after the parameter recognitions (Everbecq et al., 2001). Based on Environmental Fluid Dynamics Code (EFDC), Wu et al. predicted the chlorophyll *A* concentration and algal blooming prediction in the Daoxiang lake (Wu and Xu, 2011). Kim et al. used the EFDC and water quality model to predict the algal bloom on Han River, Korea (Kim et al., 2017), but when the concentration of chlorophyll *A* overcomes a certain threshold, the prediction accuracy will decrease. In real environment, the relationship between phytoplankton growth and environmental variables is complicated and nonlinear (Wang et al., 2017). It is difficult to obtain some specific relationships between phytoplankton concentration and environmental characteristics. Conventionally, statistical methods are used in the prediction of HABs, such as time-series regression analysis. Qin et al. combined autoregressive integrated moving average (ARIMA) model and deep belief network (DBN), proposed an ARIMA-DBN method to predict the time series of red tides in Zhoushan and Wenzhou, China (Qin et al., 2017). Çamdevýren et al.

* Corresponding authors.
E-mail addresses: gaorui@sdu.edu.cn (R. Gao), zpliu@sdu.edu.cn (Z.-P. Liu).

https://doi.org/10.1016/j.ecolind.2020.107334
Received 3 July 2020; Received in revised form 23 December 2020; Accepted 30 December 2020
Available online 14 January 2021
1470-160X/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

employed principal component source (PCS) with multiple linear regression (MLR) to predict the concentration of chlorophyll *A* (Çamdevýren et al., 2005). Mathematically, ARIMA requires that the data sequence must be stationary (Zhang, 2003). Moreover, it cannot reach the intricate relationships between responsible predicting variables and environmental independent factors.

Recently, machine-learning-based predictions have been popularized in many fields, especially has been employed in the ecological modeling to meet the complex nonlinear phenomena (Huettmann et al., 2018). Artificial neural network (ANN) technology is a widely-used technique for solving nonlinear problems. In order to select the model parameters, it often integrates genetic programming (GP) for the optimization of functional fitting (Muttil and Chau, 2007). Guallar et al. used ANN to predict the abundance of harmful algae in Alfacs Bay (Guallar et al., 2016). To explain the incomprehensibility of ANN, neural interpretation diagram and connection weight approach methodologies were applied to get the relationship between variables and harmful algae. Support vector machine (SVM) is another popular machine learning algorithm (Cortes and Vapnik, 1995). In order to train the parameters of SVM, a variety of optimization methods can be available. Lou et al. proposed a PSO-SVM model, combined SVM with particle swarm optimization (PSO) to predict the algal bloom in the Macau Main Storage Reservoir (Lou et al., 2017). Li et al predicted the 1-week and 2-week trend of algal biomass based on BP neural network with generalized regression neural network (GRNN) and SVM respectively (Li et al., 2014). The experimental results show that SVM achieved a better prediction, but the running time needs to be further optimized. Ensemble learning is another commonly-used machine learning technology, which can significantly improve the generalization ability of the model by combining multiple weak learners. Common integrated learning models include AdaBoost, random forest, gradient boosted descent tree (GBDT), etc. Nieto et al. used gradient boosted regression tree model (GBRT) model predicted the cyanotoxin contents in a reservoir located in the north of Spain (Nieto et al., 2018).

Since the available methods are mostly based on a single dataset, whether they have strong generalization ability is still unknown. There are still rooms for improving the prediction accuracy and performance. Especially, when the number of features increases, the accuracy and calculation speed of the model will be affected. The finding of important features in prediction is expected to reveal the association between HABs and features, which will indicate the cause effect of environment factors for the occurrence of HABs. These will greatly benefit the deep understanding of HABs and the potential developing of control schemes for early interference (Genuer et al., 2010).

In this paper, we provided a study of predicting HABs by environmental factors based on machine learning methods. For showing our proposed concepts in the methods, we carried out the predictions on two cases of real datasets, which are from Scripps Pier of California, USA and Sishili Bay of Shandong, China, respectively. To achieve a better prediction of algal blooms, we built up several machine-learning-based predictors, e.g., AdaBoost, ANN, GBDT, KNN, and SVM for model selection. To find the major environmental factors that affect the prediction accuracy, we tested the performance of all feature subsets under these predictors using an exhaustive method. Thus we selected out the best forecasting predictor with the combinational integrations of machine learning algorithm and environmental factors in predicting the concentration of phytoplankton. After empirical training, we found the GBDT achieves the best performance on both datasets when using specific input combinations. For the missing concentration of phytoplankton data, we also used GBDT to complete the missing value, the R^2 of the data after the filling process reached to 0.997, and the concentration of phytoplankton was tested to be predicted by GBDT one or two weeks in advance. The comparison studies prove the effectiveness and advantages of our prediction strategies. The exhaustive method ensures the optimal feature subsets that can be obtained in different predictors. They are found to be crucial in the development of algae.

2. Materials and methods

2.1. Datasets

The data used in this article are from the two locations as shown in Fig. 1, i.e., the Scripps Pier in the southern California, USA and the SishiLi Bay in the eastern Shandong, China.

The first data from Scripps pier are downloaded from the southern California coastal ocean observing system (SCCOOS). SCCOOS provides the shore station water samples weekly. Every year, California coastal erupts with irregular red tides, dyeing the whole coast. According to the red tide that broke out in 1995 on the California coast, the main cause of this phenomenon is the massive outbreak of Lingulodinium polyedrum (Kudela and Cochlan, 2000). The data used here is from the harmful algal bloom project of SCCOOS, containing the Lingulodinium polyedrum concentration and meteorological and hydrological data from 2008 to 2015, totally 365 samples. The data contains 13 parameters, including cell concentration of Lingulodinium polyedrum (cells/L) ammonia (μM), chlorophyll (mg/L), chlorophyll 11 (mg/L), chlorophyll 2 (mg/L), nitrate (μM), nirtrite (μM), phaeophytin (mg/L), phaeophytin 1 (mg/L), phaeophytin 2 (mg/L), phosphate (μM), silicate (μM), water temperate (°C). It should be noted that to monitor the occurrence of harmful algal blooms of Scripps pier, the samples in SCCOOS were taken weekly for measurements of chlorophyll *A* and nutrient concentration, qualitative and quantitative analyses of phytoplankton species composition, temperature, and salinity. According to the harmful algal bloom plan of Scripps port, the chlorophyll, chlorophyll 1 and chlorophyll 2 in the data are regarded as chlorophyll *A* for simplicity (Mazzillo et al., 2015).

The second data from Sishili bay were collected by Shandong provincial marine environmental monitoring center. As an important port and fishing ground in Shandong Province, Sishili bay is of great economic significance. During July to August of 2007, this area broke out serious algal blooms existing about one month. In this period, the center collected 40 samples of monitoring data, with a total of 11 parameters for marine water quality. In details, they are water temperature (℃), diaphaneity (m), pH value, salinity (‰), chemical oxygen demand (COD, mg/L), dissolved oxygen (DO, mg/L), carbonate (mg/L), nitrite (NO2, mg/L), nitrate (NO3 mg/L,), ammonia (NH4, mg/L), chlorophyll *A* (mg/L). Daily phytoplankton concentrations were also measured (cells/L) accordingly.

2.2. Methods

Fig. 2 illustrates the framework of our proposed method. It includes four major parts, i.e., data preprocessing, feature and model selection, feature importance, and 1-week and 2-week predictions. We firstly preprocess the collected data and generate the combinatorial feature subsets and then we implement a series of machine learning methods with exhaustive feature combinations for selecting a predictor and a subset of feature with the best prediction performance. In the experiments, GBDT is found to be the machine learning method that obtain the best prediction. Then, we employ it to evaluate the feature importance of prediction that provides the indications of crucial environmental factors for the occurrence of algal blooms. We also provide 1-week and 2-week predictions based on the trained predictor based on GBDT and optimized features.

2.2.1. Machine learning based predictors

There are five machine learning algorithms used in this paper, i.e., Adaptive Boosting (AdaBoost), Artificial Neural Network (ANN), Gradient Boosting Decision Tree (GBDT), K-nearest Neighbor (KNN) and Support Vector Machine (SVM).

2.2.1.1. AdaBoost. AdaBoost is a type of ensemble learning method that

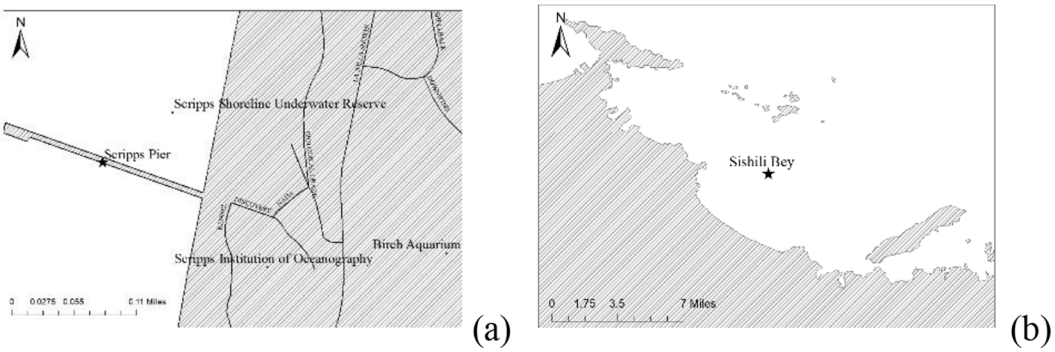


Fig. 1. The two costal locations in this paper. (a) Scripps Pier (b) SishiLi Bay.

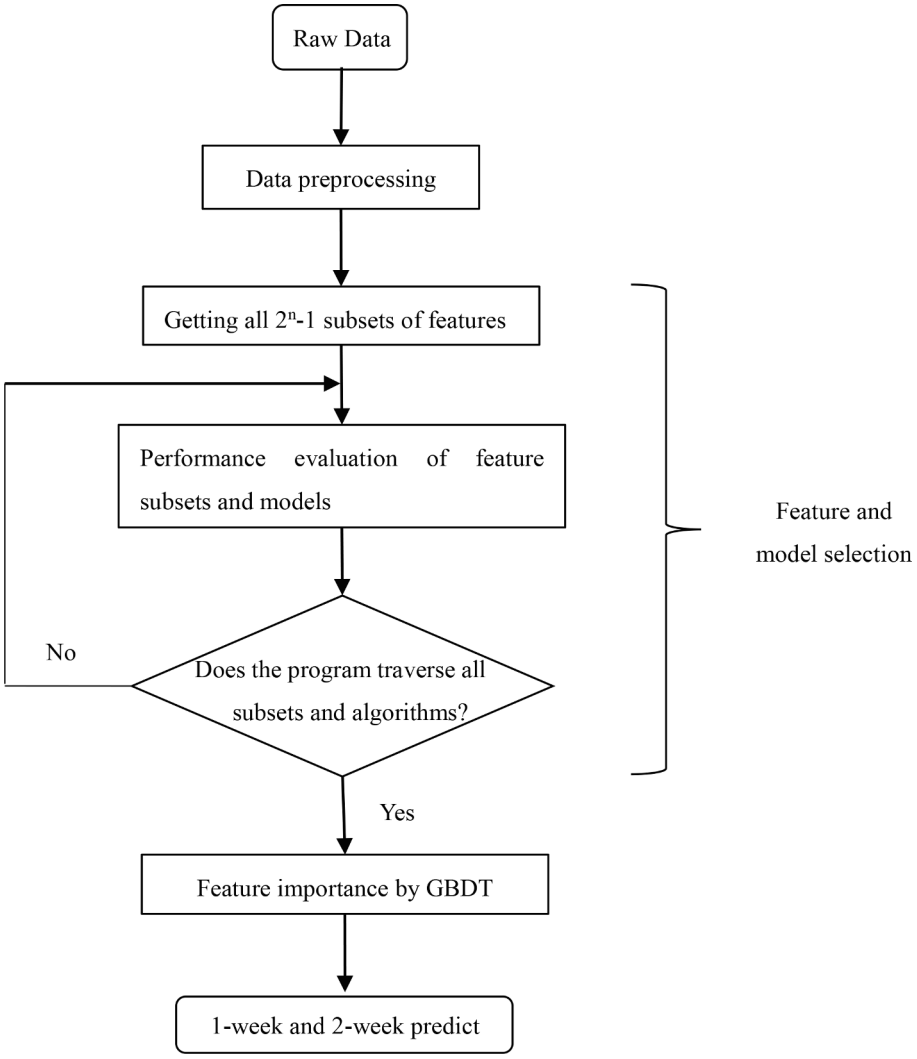


Fig. 2. The framework of predicting HABs.

combines several weak learners into one strong learner to improve classification or regression accuracy (Freund and Schapire, 1997). Common ensemble techniques are boosting and bagging. Decision tree and BP neural network are often employed as the weaker learners. Since the loss function is different, boosting algorithm has different types. AdaBoost is a boosting algorithm whose loss function is exponential loss. There are many varieties of AdaBoost for regression problems, and here we take AdaBoost R2 algorithm as the standard one (Drucker, 1997). AdaBoost pays more attention to the wrong samples, so that they can get more attention in the next training, and they are more likely to be predicted correctly.

The form of AdaBoost's ultimate strong learner is as follows

$$f(x) = \sum_{k=1}^{K} \alpha_k G_k(x) \tag{1}$$

where the f(x) is the final strong learner, K is the number of weak learner, α_k is weight of each weak learner, G_k(x) is a weak learner.

2.2.1.2. ANN. ANN, also called Multi-Layer Perceptron (MLP), is one of the most typical machine learning models. By simulating the construction of human brain, through building multilayer network, ANN achieves the goal of prediction. A typical neural network usually consists of input layer, hidden layer and output layer, based on different learning tasks, the number of hidden layers can reach any depth. In the simulation of human brain neurons, the commonly used activation functions are sigmoid and ReLU functions (Zhang and Woodland, 2016). In order to solve the calculating problem brought by network layers, back propagation (BP) algorithm is used to make it feasible to build a deep network (Hecht-Nielsen, 1989).

2.2.1.3. GBDT. GBDT, is another a widely used ensemble learning algorithm. It is also a boosting algorithm, but the specific process is different from AdaBoost. The goal of GBDT is to continuously reduce the loss of each iteration.

For the fitting problem of loss function, Freidman proposed a method of using the negative gradient of loss function to fit the approximation of the loss (Friedman, 2001) to fit a CART regression tree. The negative gradient of loss function of the i-th sample in the t-th round can be expressed as

$$\tau_{ti} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)} \tag{2}$$

where the $L(y_i, f(x_i)$ is the loss function, $\tau_{ti}$is the negative gradient, $f_{t-1}(x)$is the strong learner from the last round.

2.2.1.4. KNN. KNN is a simple but powerful machine learning algorithm to solve classification and regression problems. It represents each sample by its closet K nearest neighbors. When solving a regression problem, the value of the sample is to be obtained by averaging the values of the K nearest neighbors, i.e.,

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_k \tag{3}$$

where k is the number of neighbors, $y_k$is the value of neighbor. Often, different distance metrics such as Euclidean distance and Manhattan distance are employed to determine the neighbors.

2.2.1.5. SVM. SVM, is also a frequently used machine learning algorithm. If there have a set of linearly separable data, the line that separates the data set is called separating hyperplane. In two dimensions, it is a straight line, and in the case of high dimensions, it is called hyperplane. The nearest point to the boundary is called the support vector, their distance to the hyperplane is the margin. The purpose of SVM is to find a dividing line or hyperplane with maximum margin. For the simplest two-dimensional plane, the hyperplane is:

$$f(x) = \omega^T x + b \tag{4}$$

In order to make the classification as high as possible, the selected hyperplane needs to be able to maximize the margin, the objective function is:

$$max\frac{1}{||\omega||}, s.t., y_i(\omega^T x_i + b) \geqslant 1, \quad i = 1, 2, ..., n \tag{5}$$

where the ω is the support vector.

If the data is linearly inseparable, SVM introduces a technique called kernel trick can project data into high dimensional space to make them linearly separable.

2.2.2. Feature selection and model selection

For identifying the optimal subset of features for achieving good prediction, we implement a feature and model selection process in our framework. The selected features enable us to better understand and explain the method (Guyon and Elisseeff, 2003), as well as analyze the important factors for the occurrence of HABs. Due to the number of features are only 14 and 11 respectively in our experiments, exhaustive methods can be used to find the best feature combinations. Fig. 3 demonstrates the process. Supposing there are n features, the total number of feature subsets of all features is 2^n–1, combined with cross-validation, can quickly calculate the performance of all subsets in the model while preserving the generalization ability.

Increase the number of input features from *1* to *n*, the following *n* feature subsets can be generated, the *k*-th feature subset T_k can be expressed as,

$$T_k = \left\{(X_1, Y), (X_2, Y), ..., \left(X_{C_n^k}, Y\right)\right\} \quad \left(X_i = (x_1, x_2, ..., x_n), x_i \in X \subseteq \mathbb{R}^k\right) \tag{6}$$

Where k is the number of features, n is the number of samples.

To choose the best feature combination and best performance model, 50 rounds of 10-fold cross validation are done on all feature subsets, select the optimal combination under each number of feature subsets and use the leave-one-out method to test its MSE and R^2 on each algorithm individually. Finally, the best feature combinations and corresponding model are chosen as the final prediction model, it can be used to make the short-term prediction.

2.2.3. Performance evaluation

In this paper, we use mean squared error (MSE) and coefficient of determination (R^2) to evaluate the prediction accuracy of the model. MSE is a convenient way to measure the average error, defined as the expectation of the square of the difference between the estimated value and the true value, is a commonly used regression evaluation method.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{7}$$

Where$\hat{y}_i$is the predicted value, y_i is the true value.

R^2 is another measure for evaluating the fitting effect of the model. It is a number between 0 and 1, the closer to 1, the better the fitting effect of the model is.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{8}$$

$\hat{y}_i$ is the predicted value, y is the true value, $\bar{y}$ is the average of the original data.

2.2.4. Feature importance

The quantification of each feature in one model is also an important impact need to be considered. It can help enhance the explanatory of model and better understanding the data. In order to sort the importance of each features in the selected optimal feature subset, the GBDT was used to obtain the feature importance because it obtains the best prediction performance in the former empirical feature and model selection.

In the GBDT model, the importance of features is calculated by averaging the importance of each feature in each single decision tree. Here, Gini index is used as an index to judge the importance of features (Breiman, 2017). It can be defined as,

$$Gini(p) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2 \tag{9}$$

Where K is the number of output categories, $p_k$is the probability that sample belongs to class k.

3. Results and discussion

According to the above description, experiments were carried out on the datasets of Scripps Pier and Sishili bay respectively. The concentration of phytoplankton was predicted and the key influencing factors

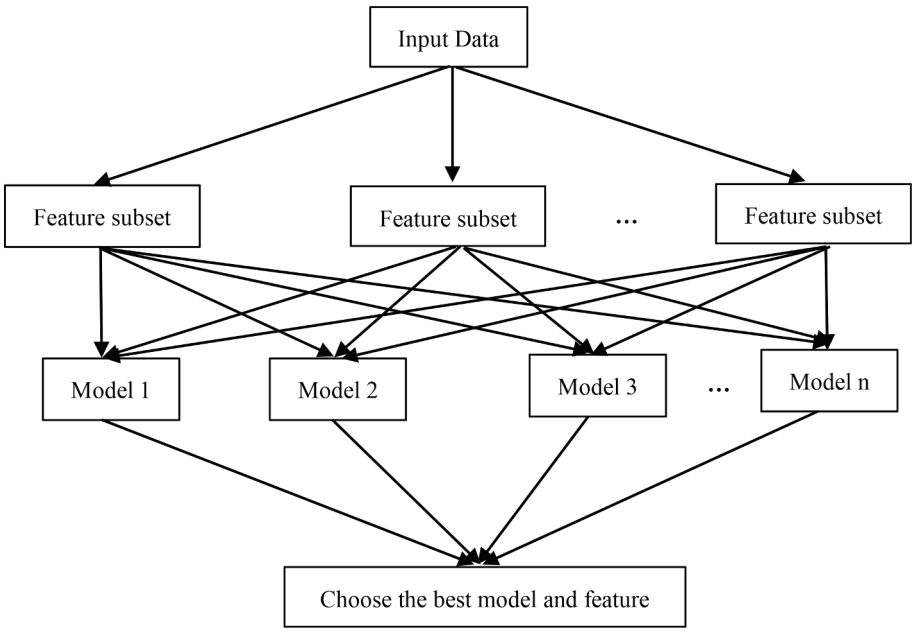


Fig. 3. The framework of feature selection and model selection.

were identified accordingly. Since the environmental factors are measured weekly, the concentration of this week can be predicted from last week's data. The prediction value was compared with the real concentration of phytoplankton for calculating the performance metrics.

3.1. Feature and model selection

In feature selection, all feature subsets were tested under different algorithms. The optimal feature subset with different feature number is evaluated for each algorithm.

Fig. 4 shows the impact of the number of features on prediction results. Each point represents the MSE of the optimal feature combination with the corresponding number of features. It can be seen that the number of features has an important impact on the final results. On the Scripps Pier data, the performance of ANN, KNN and SVM fluctuates with the change of feature number, while Adaboost and GBDT have smaller MSE, and GBDT is more stable than Adaboost. When the number of input features is 4, all models achieved the minimum MSE, and GBDT gets the best prediction with MSE of 0.031. On the Sishili Bay data, the performance of the model changes more obviously with the number of input features, especially ANN. When the number of features is 6, the prediction result is very poor. Based on ensemble learning, GBDT still has a stable performance. When the number of input features is 3, also the GBDT reaches the minimum MSE of 0.383. Due to the different number of samples in the two datasets, the performance of each algorithm is different, but it certainly can be seen that only specific environmental factors are related to the concentration of phytoplankton. Moreover, the predictors based on ensemble learning are less affected by the number of features, and their prediction results are more stable. Overall, GBDT obtains the best prediction performance in both datasets. Fig. 5 shows the standard deviation of 50 rounds of 10-fold cross validations for different models under the optimal feature combinations. As shown, except for ANN, other models are stable in each round of cross validations.

Tables 1 and 2 shows the improvement of prediction results after feature selection on the two datasets, respectively. After feature selection, the accuracy of each algorithm has been significantly improved. In Scripps Pier data, ANN, KNN and SVM improved most obviously. Both

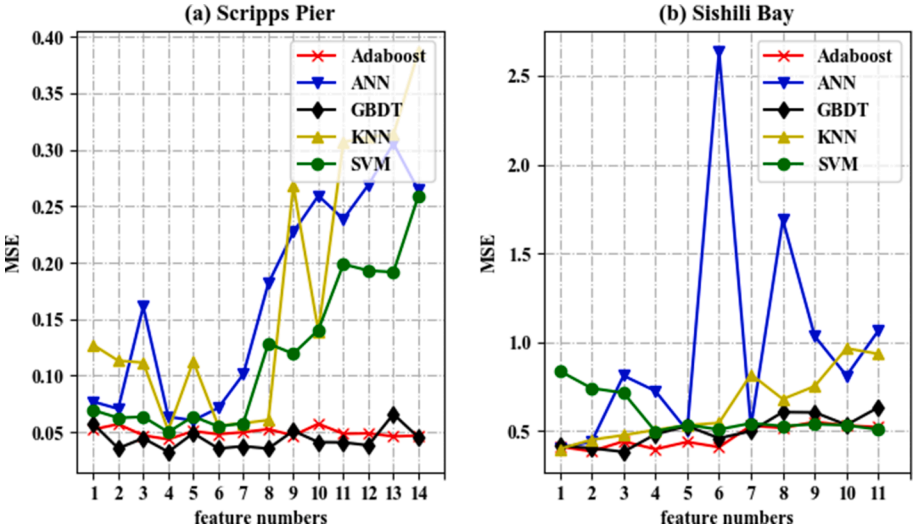


Fig. 4. MSE of optimal feature combination under different models.

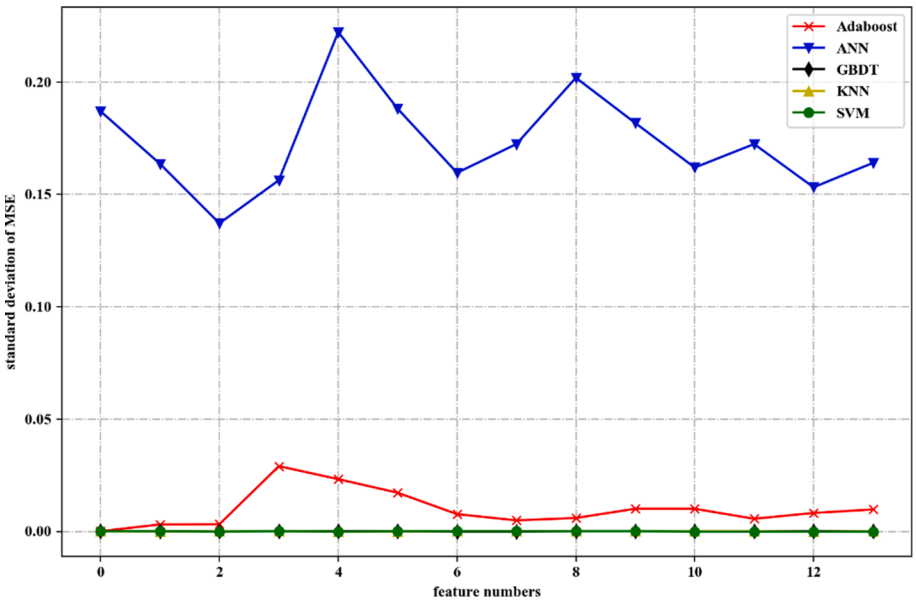


Fig. 5. Standard deviation of each model with different number of features.

Table 1
Comparison of Prediction Results between Optimum Feature Combination and Raw Data (Scripps Pier).

	Raw data		After feature selection	
	MSE	R^2	MSE	R^2
AdaBoost	0.047	0.953	0.044	0.956
ANN	0.240	0.760	0.061	0.939
GBDT	0.045	0.951	0.031	0.969
KNN	0.131	0.869	0.052	0.948
SVM	0.174	0.826	0.050	0.950

Table 2
Comparison of Prediction Results between Optimum Feature Combination and Raw Data (Sishili Bay).

	Raw data		After feature selection	
	MSE	R^2	MSE	R^2
AdaBoost	0.522	0.386	0.386	0.614
ANN	1.424	−0.424	0.418	0.582
GBDT	0.634	0.366	0.383	0.617
KNN	0.935	0.065	0.451	0.549
SVM	0.509	0.491	0.499	0.501

the MSE and R^2 have improved greatly. As the optimal prediction model, the MSE of GBDT decreased to 0.031 from 0.045, and R^2 increased to 0.969. The same effect can be achieved in Sishili Bay data. After feature selection, the MSE of GBDT decreased to 0.383, and the R^2 increased to 0.617. Because the sample number of Sishili Bay data is small, the prediction effect of the GBDT cannot get the performance as that in Scripps Pier data.

Table 3 shows the final input features after feature selection under two datasets. Although the accuracy of five models is different, the feature combination that obtains good prediction results under each model has the same overlapping part. In Scripps Pier dataset, ammonia, chlorophyll1 and temperature are the features with the highest selection rate. In Sishili Bay dataset, DO is the most significant feature. In the KNN model, the best feature combination even only contains DO. Combined with the feature combination of two datasets, there are still some similarity. Ammonia, chlorophyll *A* and nitrite are selected as the input features under both datasets. This indicates the importance of these environmental factors in the occurrence of HABs.

Table 3
Best feature combination under each model.

	SCCOOS	Sishili Bay
AdaBoost	Ammonia + Chlorophyll1 + Nitrite + temp	DO + Chlorophyll *A*
ANN	Ammonia + Chlorophyll 1 + Chlorophyll 2 + Nitrate	salinity + DO
GBDT	Ammonia + Chlorophyll1 + Nitrite + temp	DO + Nitrite + Chlorophyll *A*
KNN	Ammonia + Chlorophyll + Chlorophyll1 + temp	DO
SVM	Ammonia + Chlorophyll1 + Chlorophyll1 + temp	Diaphaneity + PH + DO + NH4

3.2. *Feature importance analysis*

Fig. 6 shows the importance of each feature evaluated by GBDT. In two datasets, chlorophyll *A* and nitrite have a certain relationship with phytoplankton concentrations. In Scripps Pier data, as a primary pigment involved in phytoplankton photosynthesis, the feature importance of chlorophyll *A* occupies 0.527, far beyond other features. As shown, ammonia, temperature and nitrite are also related to the prediction of phytoplankton concentration. In Sishili Bay data, the feature importance of DO, nitrite and chlorophyll Are close, reaching 0.35, 0.337and 0.312 respectively. The selected features have been reported in literature with close relationship with HABs from ecosystem perspectives. Sivapragasam et al. predicted the concentration of chlorophyll *A* by a genetic programming method. They found that chlorophyll *A* is the most sensitive variable in their method and DO also plays an important role because it is necessary for organisms in water to breathe (Sivapragasam et al., 2010). Roy et al found that Lingulodinium polyedrum is very sensitive to low temperature through experiments (Roy et al., 2014). When the temperature drops to 8 ℃, all cells stop swimming, and when returned to normal culture temperature (18 ℃), all cells were found swimming. In addition, the increase of N and P in water will lead to eutrophication and HAB (Blaas and Kroeze, 2016). Through feature selection, the environmental factors related to the concentration of phytoplankton can be identified.

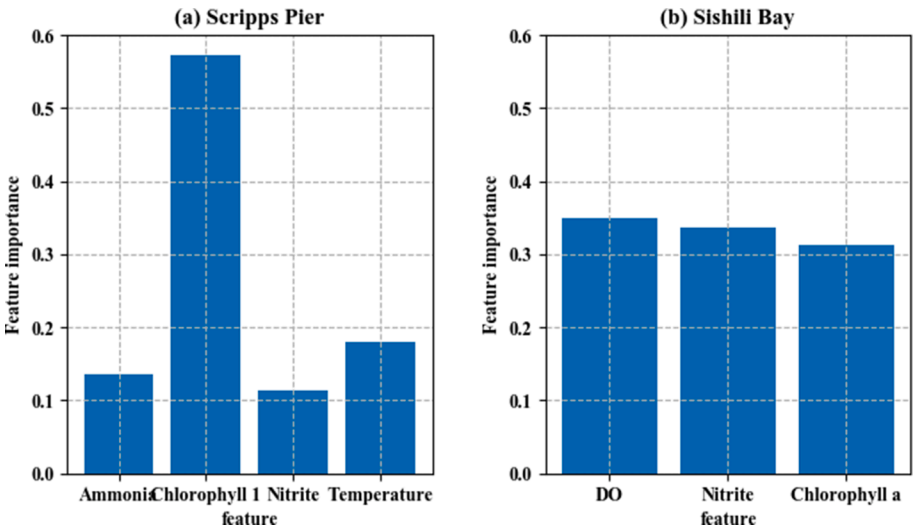


Fig. 6. Feature importance based on GBDT.

3.3. Day prediction by GBDT

To show the necessity of feature selection, Fig. 7 illustrates the day prediction of phytoplankton cell concentration by feature selection and non-feature selection in Scripps Pier data. When a specific combination of input features (Ammonia + Chlorophyll1 + Nitrite + temp) is used, the prediction effect is improved obviously, GBDT can accurately learn the mapping between environmental factors and phytoplankton concentration, the MSE decreased from 0.045 to 0.031, R^2 increased from 0.955 to 0.969.

Fig. 8 shows the prediction results of the GBDT algorithm with DO + Nitrite + Chlorophyll *A* as the input on Sishili Bay data. After feature selection, the MSE gets 0.383, and R^2 is 0.617. The prediction curve after feature selection can capture the peak of the concentration. Since there are only 40 data samples and the number of features is also small on Sishili Bay data, the prediction performance still has a large room for improvement. The results show that the concentration of phytoplankton in water can be accurately predicted by the selected features and machine learning model. The selected features are of great significance for the development of algae.

3.4. One-week and two-week prediction

In the Scripps Pier dataset, it was found that phytoplankton concentrations were 0 for many days and the sampling period were nearly one week, considering that there may be data missing or the concentration is close to 0. According to the analysis in the former Section 3.3, the concentration of phytoplankton can be accurately predicted by using the environmental parameters of the day. If the environmental parameters of every day are complete, the missing concentration can be imputed by GBDT using specific feature combination.

Fig. 9 shows the comparison between the concentration after data imputation and the original data. We found the predicted data using GBDT with Ammonia + Chlorophyll1 + Nitrite + temp as the input is very close to the original raw data, the R^2 reached 0.996. The zero point in the original data set does correspond to the lower level of phytoplankton concentration. Through data filling, the time interval between two adjacent sample points is about one week, so phytoplankton concentration can be predicted by using the data of one week or two weeks ago.

By using the Ammonia + Chlorophyll1 + Nitrite + temp of the previous week as the input of GBDT, the short-term phytoplankton

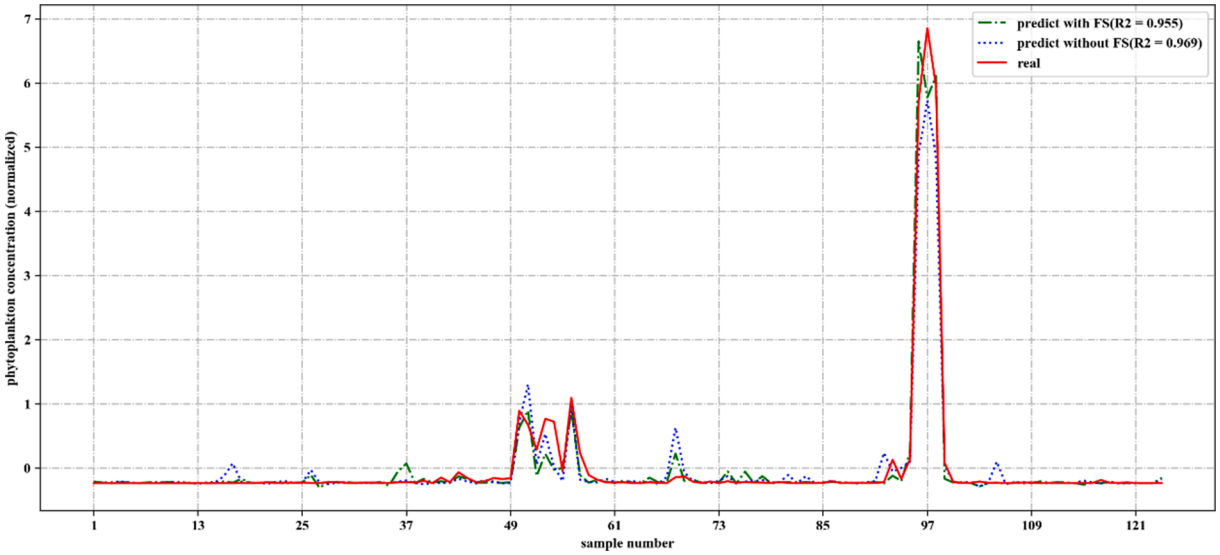


Fig. 7. Prediction results on Scripps Pier.

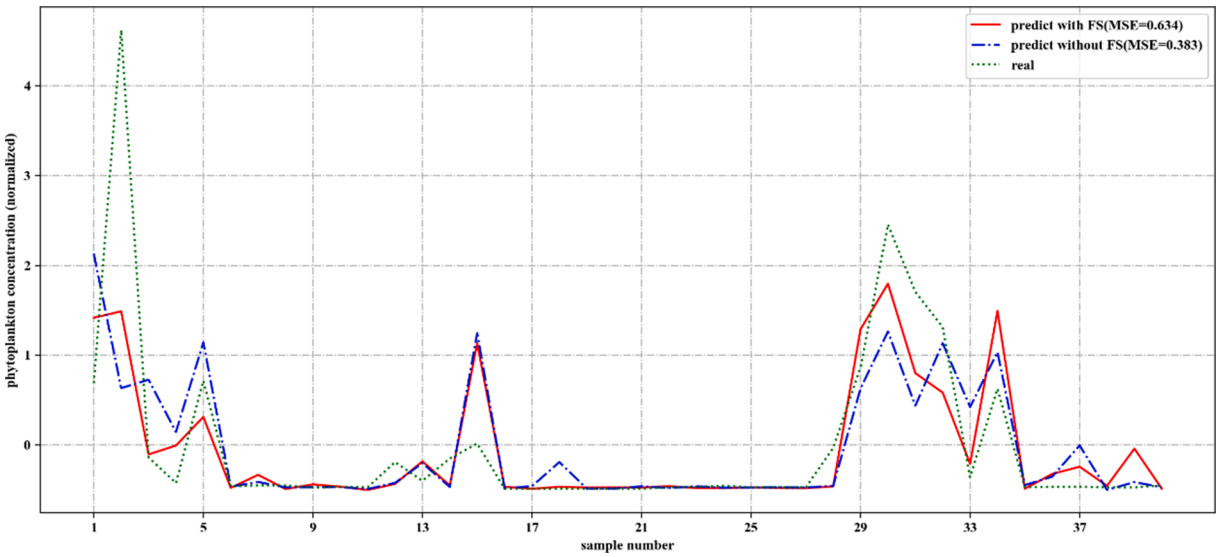


Fig. 8. Prediction results on Sishili Bay.

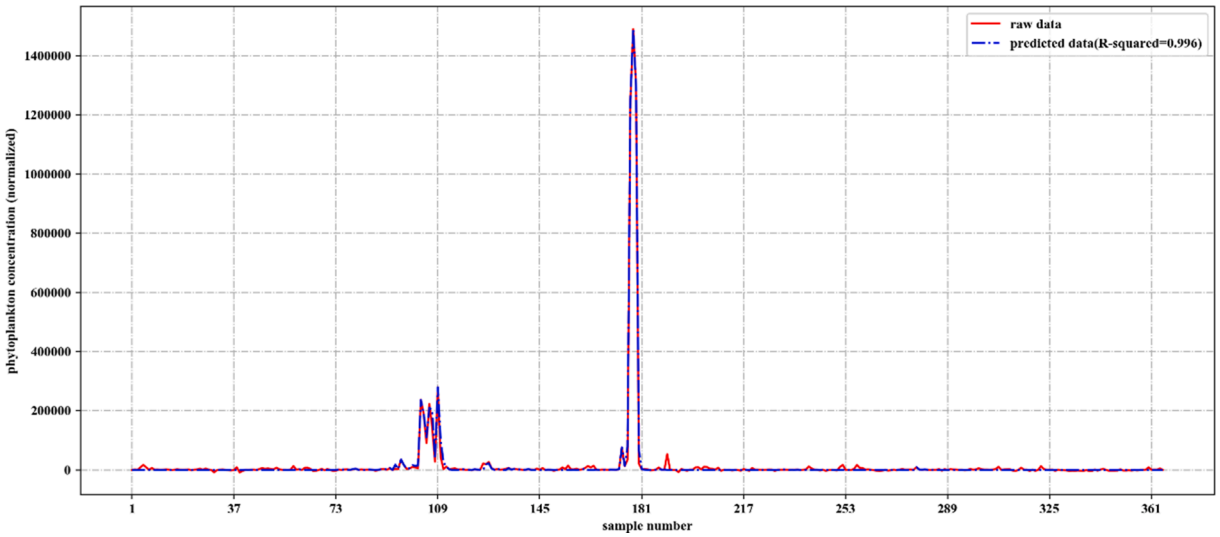


Fig. 9. Comparison between raw data and filled data.

concentration can be predicted. In prediction process, the data of the n–1 days are used as training to predict the data of the n-th day. Table 4 shows the comparison of 1-week prediction of GBDT with ARIMA. It can be found that the prediction results of GBDT is obviously better than ARIMA. Compared with ARIMA, GBDT has a stronger interpretability. Because the interval period of water quality sampling is basically 7 days in raw data, and there is no other data between two sampling points, the prediction effect is of course lower than the daily prediction of phytoplankton. However, it also has a certain prediction effect for the peak point of concentration change (shown in Fig. 10).

Fig. 10 shows the 1-week and 2-week prediction results by GBDT with the input of Ammonia + Chlorophyll1 + Nitrite + temp as features. As shown, the trend of concentration changes can be predicted based on the data from a week ago and two weeks ago, the R^2 of 1-week and 2-week predictions reached 0.904 and 0.833 respectively. When adding data from two weeks ago for prediction, the prediction accuracy is slightly lower than using the data only from one week ago. But for the peak point of the highest concentration, it can still be accurately predicted in advance. If there is more sampling data or collecting parameters within a week, the forecasting would be better.

Table 4
Comparisons with traditional methods.

	ARIMA (0,0,2)	GBDT(learning_rate = 0.1,loss="huber,n_estimators = 200)
MSE	0.430	0.096
R^2	0.570	0.904

4. Conclusions

In this paper, we proposed a machine learning method to predict the algal blooms with the investigation of feature importance of environmental factors. Through the selection of features and models, the best performance of feature combination and prediction models were chosen. Moreover, the method can be used to impute the missing data and predict the trend of phytoplankton concentration in advance simultaneously.

We validated our method on two real datasets. GBDT was identified as the final prediction model for its best empirical results. In Scripps Pier dataset, Ammonia + Chlorophyll1 + Nitrite + temp was selected as the

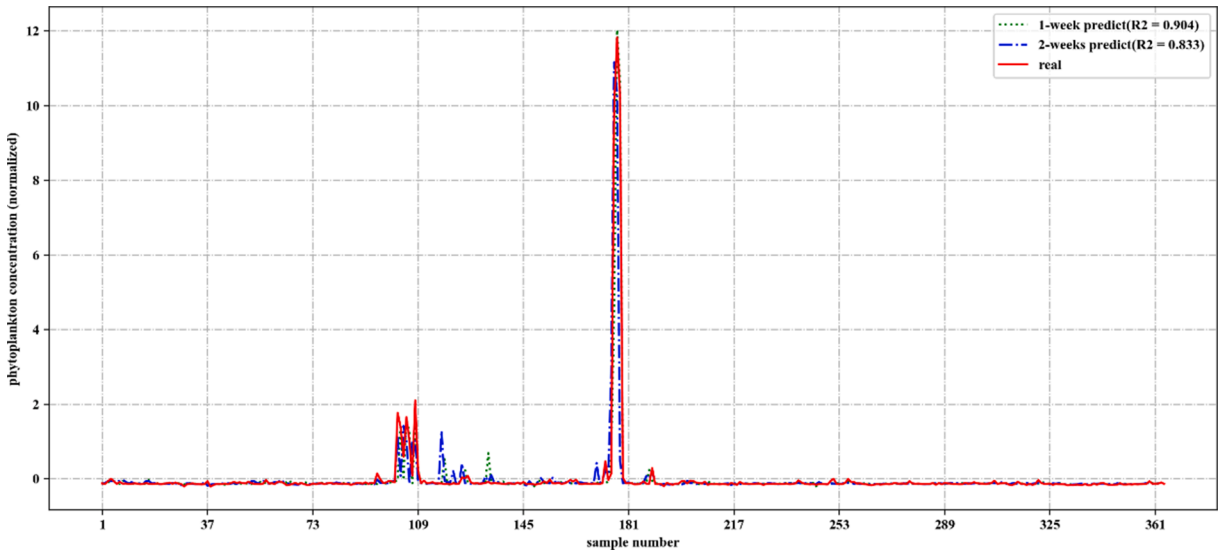


Fig. 10. 1-week and 2-week prediction.

input feature combination to impute the original raw data and predict the day concentration of phytoplankton. By the feature importance evaluation score of GBDT model, the impact of environmental factors on phytoplankton concentration have been further identified. After filling the missing data, the R^2 of prediction in the imputed data was improved to 0.967. And the R^2 reached 0.904 in the short-term phytoplankton concentration prediction using the data one or two weeks ago. While the data sample of Sishili Bay is small. It can't use GBDT to predict this kind of short-term concentration. It can be noted that due to the small sample size and the large sampling interval, there might be some errors in the predictions. The proposed method is expected to reach better prediction performance using a bigger dataset or combining the machine-learning-based strategy with the other methods.

CRediT authorship contribution statement

Peixuan Yu: Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review & editing. **Rui Gao:** Supervision, Project administration. **Dezhen Zhang:** Formal analysis. **Zhi-Ping Liu:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the NSFC-Shandong Province Joint Grant No. U1806202; Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project, 2019JZZY010423); National Natural Science Foundation of China (NSFC) under Grant No. 61533011.

References

Anderson, D.M., Glibert, P.M., Burkholder, J.M., 2002. Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. Estuaries 25 (4), 704–726. https://doi.org/10.1007/BF02804901.

Blaas, H., Kroeze, C., 2016. Excessive nitrogen and phosphorus in European rivers: 2000–2050. Ecol. Indicators 67, 328–337. https://doi.org/10.1016/j.ecolind.2016.03.004.

Breiman, L. 2017. Classification and Regression Trees. Routledge.

Çamdevýren, H., Demýr, N., Kanik, A., Keskýn, S., 2005. Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. Ecol. Model. 181 (4), 581–589. https://doi.org/10.1016/j.ecolmodel.2004.06.043.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. https://doi.org/10.1007/BF00994018.

Drucker, H., 1997. Improving regressors using boosting techniques. ICML 107–115.

Everbecq, E., Gosselain, V., Viroux, L., Descy, J.-P., 2001. Potamon: a dynamic model for predicting phytoplankton composition and biomass in lowland rivers. Water Res. 35 (4), 901–912.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recogn. Lett. 31 (14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.

Gregorio, D.E., Pieper, R.E., 2000. Investigations of red tides along the southern California coast. Bull. Southern Calif. Acad. Sci. 99 (3), 147.

Guallar, C., Delgado, M., Diogène, J., Fernández-Tejedor, M., 2016. Artificial neural network approach to population dynamics of harmful algal blooms in Alfacs Bay (NW Mediterranean): Case studies of Karlodinium and Pseudo-nitzschia. Ecol. Model. 338, 37–50. https://doi.org/10.1016/j.ecolmodel.2016.07.009.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3 (Mar), 1157–1182.

Hallegraeff, G.M., 1993. A review of harmful algal blooms and their apparent global increase. Phycologia 32 (2), 79–99. https://doi.org/10.2216/i0031-8884-32-2-79.1.

Hao, Y., Tang, D., Yu, L., Xing, Q., 2011. Nutrient and chlorophyll a anomaly in red-tide periods of 2003–2008 in Sishili Bay, China. Chin. J. Ocean. Limnol. 29 (3), 664–673. https://doi.org/10.1007/s00343-011-0179-3.

Hecht-Nielsen, R., 1989. Theory of the Back Propagation Neural Network, vol. 1. IEEE, New York, pp. 593–605. https://doi.org/10.1109/IJCNN.1989.118638.

Huettmann, F., Craig, E.H., Herrick, K.A., Baltensperger, A.P., Humphries, G.R.W., Lieske, D.J., Miller, K., Mullet, T.C., Oppel, S., Resendiz, C., Rutzen, I., Schmid, M.S., Suwal, M.K., Young, B.D., 2018. Use of machine learning (ML) for predicting and analyzing ecological and 'presence only' data: An overview of applications and a good outlook. In: Humphries, G., Magness, D.R., Huettmann, F. (Eds.), Machine Learning for Ecology and Sustainable Natural Resource Management. Springer International Publishing, Cham, pp. 27–61. https://doi.org/10.1007/978-3-319-96978-7_2.

Kim, J., Lee, T., Seo, D., 2017. Algal bloom prediction of the lower Han River, Korea using the EFDC hydrodynamic and water quality model. Ecol. Model. 366, 27–36. https://doi.org/10.1016/j.ecolmodel.2017.10.015.

Kudela, R.M., Cochlan, W.P., 2000. Nitrogen and carbon uptake kinetics and the influence of irradiance for a red tide bloom off southern California. Aquat. Microb. Ecol. 21, 31–47. https://doi.org/10.3354/ame021031.

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.

Lou, I., Xie, Z., Ung, W.K., Mok, K.M., 2017. In: Lou, I., Han, B., Zhang, W. (Eds.), Advances in Monitoring and Modelling Algal Blooms in Freshwater Reservoirs. Springer Netherlands, Dordrecht, pp. 125–141. https://doi.org/10.1007/978-94-024-0933-8_8.

Mazzillo, F., Carter, M., Busse, L., McGowan, J., 2015. Documenting a dinoflagellate bloom off Scripps pier-a report from the Pier Chlorophyll Program. 1–4.

Muttil, N., Chau, K.-W., 2007. Machine-learning paradigms for selecting ecologically significant input variables. Eng. Appl. Artif. Intell. 20 (6), 735–744. https://doi.org/10.1016/j.engappai.2006.11.016.

Nieto, P.J.G., García-Gonzalo, E., Sánchez Lasheras, F., Alonso Fernández, J.R., Díaz Muñiz, C., de Cos Juez, F.J., 2018. Cyanotoxin level prediction in a reservoir using gradient boosted regression trees: a case study. Environ. Sci. Pollut. Res. 25 (23), 22658–22671. https://doi.org/10.1007/s11356-018-2219-4.

Qin, M., Li, Z., Du, Z., 2017. Red tide time series forecasting by combining ARIMA and deep belief network. Knowl. Based Syst. 125, 39–52. https://doi.org/10.1016/j.knosys.2017.03.027.

Roy, S., Letourneau, L., Morse, D., 2014. Cold-induced cysts of the photosynthetic dinoflagellate lingulodinium polyedrum have an arrested circadian bioluminescence rhythm and lower levels of protein phosphorylation. Plant Physiol. 164 (2), 966–977. https://doi.org/10.1104/pp.113.229856.

Sellner, K.G., Doucette, G.J., Kirkpatrick, G.J., 2003. Harmful algal blooms: Causes, impacts and detection. J. Ind. Microbiol. Biotechnol. 30 (7), 383–406. https://doi.org/10.1007/s10295-003-0074-9.

Sivapragasam, C., Muttil, N., Muthukumar, S., Arun, V.M., 2010. Prediction of algal blooms using genetic programming. Mar. Pollut. Bull. 60 (10), 1849–1855. https://doi.org/10.1016/j.marpolbul.2010.05.020.

Tang, DanLing, Di, BaoPing, Wei, G., Ni, I.-H., Oh, I.S., Wang, SuFen, 2006. Spatial, seasonal and species variations of harmful algal blooms in the South Yellow Sea and East China Sea. Hydrobiologia 568 (1), 245–253. https://doi.org/10.1007/s10750-006-0108-1.

Wang, Y., Xie, Z., Lou, InChio, Ung, W.K., Mok, K.M., 2017. Algal bloom prediction by support vector machine and relevance vector machine with genetic algorithm optimization in freshwater reservoirs. Eng. Comput. 34 (2), 664–679. https://doi.org/10.1108/EC-11-2015-0356.

Wu, G., Xu, Z., 2011. Prediction of algal blooming using EFDC model: Case study in the Daoxiang Lake. Ecol. Model. 222 (6), 1245–1252. https://doi.org/10.1016/j.ecolmodel.2010.12.021.

Zhang, C., Woodland, P.C., 2016. DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions. IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), Shanghai, pp. 5300–5304.

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.