ACCSCIENCE PUBLISHING

Asian Journal of Water, Environment and Pollution

ORIGINAL RESEARCH ARTICLE

Interpretable application of machine learning techniques in the assessment of coastal water quality of Poyang Lake, China

Zhangtao Hu[1], Rong Yi[1]*, Xin Liu[2,3,4]*, Chao Du[1], Xinyue Di[1], Qiang Huang[2,3,4], Aimin Hao[5], and Yasushi Iseri[5]

[1]School of Geography and Environmental Engineering, Gannan Normal University, Ganzhou, Jiangxi, China

[2]Guangxi Key Laboratory of Marine Environmental Science, Guangxi Academy of Marine Sciences, Guangxi Academy of Sciences, Nanning, Guangxi, China

[3]Beibu Gulf Marine Industry Research Institute, Fangchenggang, Guangxi, China

[4]Guangxi Laboratory of Oceanography, Nanning, Guangxi, China

[5]College of Life and Environmental Science, Wenzhou University, Wenzhou, Zhejiang, China

***Corresponding authors:**
Rong Yi
(yirong@gnnu.edu.cn);
Xin Liu
(lx@gxas.cn)

Citation: Hu Z, Yi R, Liu X, *et al*. Interpretable application of machine learning techniques in the assessment of coastal water quality of Poyang Lake, China. *Asian J Water Environ Pollut*.
doi: 10.36922/AJWEP026260183

Received: June 27, 2026

Revised: July 21, 2026

Accepted: July 28, 2026

Published online: August 7, 2026

Abstract

Machine learning models (MLMs) have made substantial progress across diverse scientific domains owing to their strong predictive capabilities and effectiveness in extracting patterns from high-dimensional data. In lake environmental management, clarifying the relationships between environmental factors and water quality indicators (WQIs) is essential for improving assessment accuracy and operational efficiency. In this study, four MLMs (support vector regression [SVR], random forest [RF], extreme gradient boosting [XGBoost], and k-nearest neighbors [KNN]) were applied to identify the key environmental parameters influencing WQIs in the coastal zone of Poyang Lake, China, and to evaluate the predictive performance of each model. Model interpretability was enhanced using the SHapley Additive exPlanations (SHAP) method, which quantified the contributions of environmental variables to WQI variability. Total nitrogen (TN), total phosphorus (TP), and chlorophyll-a (Chla) concentrations were used as representative WQIs. Among all models, XGBoost consistently exhibited the highest predictive accuracy. Electrical conductivity (EC), water temperature (T), dissolved oxygen (DO), oxidation–reduction potential (ORP), and pH were identified as the five most influential parameters. Specifically, EC, T, and DO were the dominant drivers of TN and Chla, accounting for 92.8% and 84.2% of the total SHAP contributions, respectively. Meanwhile, TP was primarily governed by ORP, pH, and T, with a combined contribution of 70.6%. XGBoost accurately reproduced the observed WQIs using environmental parameters as inputs. These findings demonstrate that interpretable MLMs can effectively reveal the environmental drivers of lake water quality and offer a robust framework for ecosystem monitoring and forecasting in freshwater environments.

Keywords: Machine learning models; SHapley Additive exPlanations; Environmental parameter; Water quality modeling; Ecological forecasting; Lake ecosystem

1. Introduction

With increasing anthropogenic pressure, lake water quality has deteriorated globally, posing severe risks to the ecological integrity of watershed systems.[1,2] As vital components of regional hydrological networks, lakes are highly sensitive to multiple interacting environmental drivers.[3] Understanding the complex and spatially heterogeneous relationships between these drivers and water quality indicators (WQIs) is essential for elucidating lake ecosystem dynamics and improving water resource management strategies.[4]

Traditional statistical methods, including multiple regression[5] and ordination methods,[6] have been widely applied to investigate the relationships between environmental variables and water quality, generally under the assumption of linear or predefined relationships.[7] Heneash *et al.*[8] employed regression analysis, multiple correlation analysis, redundancy analysis (RDA), and canonical correspondence analysis (CCA) to identify water temperature (T), salinity, and silicate as the primary environmental factors influencing plankton community composition. Similarly, Jakovljević *et al.*[9] investigated the effects of environmental factors on diatom distribution and revealed complex ecological patterns using RDA and CCA. Beyond biological communities, multivariate statistical techniques, including Spearman and Pearson correlation analyses and RDA, have also been widely used to assess the impacts of land use and anthropogenic activities on nutrient dynamics in aquatic ecosystems. Mu *et al.*[10] demonstrated that agricultural and urban land uses were significantly associated with increased total nitrogen (TN) and total phosphorus (TP) concentrations in the Huai River (China), while Li *et al.*[11] identified upstream inflows and intensive aquaculture as major contributors to TP and ammonia nitrogen enrichment in the Daheiting Reservoir (China). Mebarkia and Boufekane[12] reported that anthropogenic salinization and agricultural pollution were important stressors affecting water quality in Aïnzeda Lake (Algeria). Although these approaches provide valuable insights into ecological relationships, they may have limited capacity to capture complex nonlinear relationships among multiple environmental variables. Therefore, more flexible frameworks have been increasingly applied to characterize nonlinear patterns in aquatic ecosystems.[13-15]

In recent years, machine learning models (MLMs) have made substantial progress across various scientific disciplines, primarily because of their strong predictive capabilities and ability to extract patterns from complex, high-dimensional datasets.[16,17] Algorithms, such as random forests (RFs), support vector machines (SVMs), and artificial neural networks (ANNs), have demonstrated robust performance in predicting WQIs and identifying influential environmental factors. For example, Thanh *et al.*[18] evaluated the effectiveness of RF and ANN in predicting groundwater quality and identifying calcium, magnesium, and bicarbonate as the primary pollutants. Uddin *et al.*[19] compared several MLMs, including SVM, naïve Bayes, RF, k-nearest neighbors (KNN), and extreme gradient boosting (XGBoost), to determine the most appropriate classifier for water quality categorization under the WQI framework. Irwan *et al.*[20] applied MLMs to estimate WQI in the Gombak River (Malaysia) using variables such as dissolved oxygen (DO) and pH, highlighting the utility of data-driven models for water-quality monitoring. Kruk[21] reported that XGBoost effectively detected cyanobacterial concentrations based on environmental inputs. Furthermore, Chen *et al.*[22] integrated physically based models with RF, SVM, and long short-term memory (LSTM) networks to forecast phytoplankton dynamics in Taihu Lake (China), while Huang *et al.*[23] identified electrical conductivity (EC) as a key predictor of chlorophyll-a (Chla) concentration through RF modeling.

Despite the high predictive performance of MLMs, their limited interpretability has impeded our understanding of underlying causal relationships among variables.[24] To address this challenge, the SHapley Additive exPlanations (SHAP) framework was developed to quantify the individual contributions of input features to model predictions.[25,26] Recent studies, such as that by Makumbura *et al.*,[27] have combined MLMs with SHAP methods to enhance interpretability in water-quality modeling. For example, chemical oxygen demand and biological oxygen demand were identified as the most influential factors affecting water quality in the Kelani River Basin (Sri Lanka). The SHAP algorithm offers distinct advantages, including the ability to clarify the overall effects of environmental drivers and to assess interactions among multiple parameters at specific sites. This interpretability aligns well with the spatial and temporal heterogeneity of water quality dynamics.[28]

Poyang Lake, the largest freshwater lake in China (with a surface area of approximately 4,100 km^2 and a depth of 23 m), plays a critical role in regulating regional hydrology, sustaining biodiversity, and supporting agricultural and economic activities across the Yangtze River Basin. The lake receives inflows from five major rivers and discharges through the Hukou outlet, forming a dynamic river–lake interaction system.[29] The surrounding area includes several marginal coastal lakes (e.g., Banghu Lake, Dahu Pond, and Zhonghu Pond), which are hydrologically connected to the main water body through seasonal fluctuations, exerting a substantial influence on water quality.[30] These marginal lakes typically merge with Poyang Lake during the rainy season (June–September) and become isolated during the dry season (October–March), forming a dynamic floodplain–lake system. Notably, the water quality of these seasonally isolated lakes is sensitive to environmental changes because of their limited water storage capacity and strong dependence on surface water connectivity.[29,31] Although considerable research has focused on water quality variability,[32,33] eutrophication processes,[34,35] and human-induced impacts,[36,37] few studies have systematically examined the relationship between environmental drivers and WQIs, particularly in dynamic marginal lake systems.[38] The integration of MLMs with SHAP offers a promising approach to disentangle these complex interactions, thereby improving our understanding of ecological responses in marginal lakes under shifting environmental conditions.

This study proposes an integrated analytical framework that combines MLMs with SHAP to evaluate the influence of environmental drivers on water quality in the coastal area of Poyang Lake. The analysis incorporated six environmental parameters (T, DO, pH, EC, oxidation–reduction potential [ORP], and total dissolved solids [TDS]) and three WQIs (TN, TP, and Chla). Both TN and TP serve as key nutrient indicators, as they directly govern nutrient enrichment and eutrophication dynamics in lake ecosystems.[39,40] Excessive inputs of these nutrients are primary drivers of accelerated phytoplankton growth, shifts in community composition, and altered ecosystem productivity.[39,41,42] Chla, a widely recognized proxy for phytoplankton biomass, effectively reflects the biological response to nutrient availability and serves as a core parameter for assessing trophic status.[43-45] Consequently, TN, TP, and Chla were selected as representative metrics to evaluate the overall water quality of the target aquatic ecosystems. Four MLMs (support vector regression [SVR], RF, XGBoost, and KNN) were applied to identify the most influential predictors and assess the predictive performance of each model. This integrative approach aims to elucidate the regulatory roles of environmental variables, enhance the interpretability of spatiotemporal water quality dynamics in freshwater lake systems, and evaluate the applicability of machine learning methods for WQI prediction.

2. Materials and methods

2.1. Study area and data collection

This study focused on the inshore marginal lake of Poyang Lake, specifically Banghu Lake, which has a surface area of approximately 80 km^2,[30] located on the western edge of Poyang Lake (29°10′N–29°17′N, 115°54′E–116°01′E; Figure 1). Water samples were collected from 14 sampling sites between May 10 and August 10, 2024. Six environmental parameters (T, DO concentration, pH, EC, ORP, and TDS) were measured using a Professional Plus portable water quality analyzer (605604, YSI, United States of America). Approximately 1.5 L of surface water was collected from each site using an acrylic water sampler and stored in polyethylene bottles for subsequent analyses. For Chla determination, 1 L of water was filtered through a 0.45 μm membrane filter (diameter: 47 mm, CN-CA, Haining Kewo Filter Equipment Co., Ltd., China). Unfiltered water samples (50 mL) were used to analyze TP and TN. All samples were preserved at −20 °C and transported to the laboratory for further analysis.

2.2. Chemical analysis and chlorophyll-a measurement

Total phosphorus concentration was determined using the national standard method (GB11893-89). A 15 mL unfiltered water sample was treated with 2.4 mL of 5% potassium persulfate and autoclaved at 120 °C for 30 min (G180DWS, ZEALWAY Instrument Inc., China). The TP concentration was measured using the molybdenum blue method (detection limit: 1 μg/L), and the absorbance was recorded at 700 nm using a UV/visible spectrophotometer (UV-2700, Shimadzu, Japan). The TN concentration (detection limit: 1 μg/L) was determined using a total organic carbon/TN analyzer (Multi N/C 2100, Analytik Jena AG, Germany). The Chla concentration in lake water (detection limit: 1 μg/L) was measured using the national standard method (HJ897-2017). Briefly, Chla from phytoplankton assemblages was extracted using 10 mL of 90% acetone in the dark at 4 °C for 12 h. The samples were then centrifuged at 4,000 rpm for 10 min, and the supernatant was used for measurement. Absorbance was recorded at 630, 647, 664, and 750 nm using the UV-2700 UV/visible spectrophotometer. All measurements were conducted in triplicate at each sampling site. The environmental parameter profiles are shown in Table S1.

2.3. Machine learning model selection

A sequential analytical framework was adopted, integrating machine learning modeling, model interpretability, and statistical validation. First, optimal MLMs were selected through a comparative performance evaluation. The SHAP analysis was performed to quantify the contributions of each input feature and to explore potential response patterns. Finally, a traditional statistical test was applied for model-independent validation to ensure consistency.

2.3.1. Support vector regression

Support vector regression is a regression-based extension of SVM that aims to identify a hyperplane that best fits

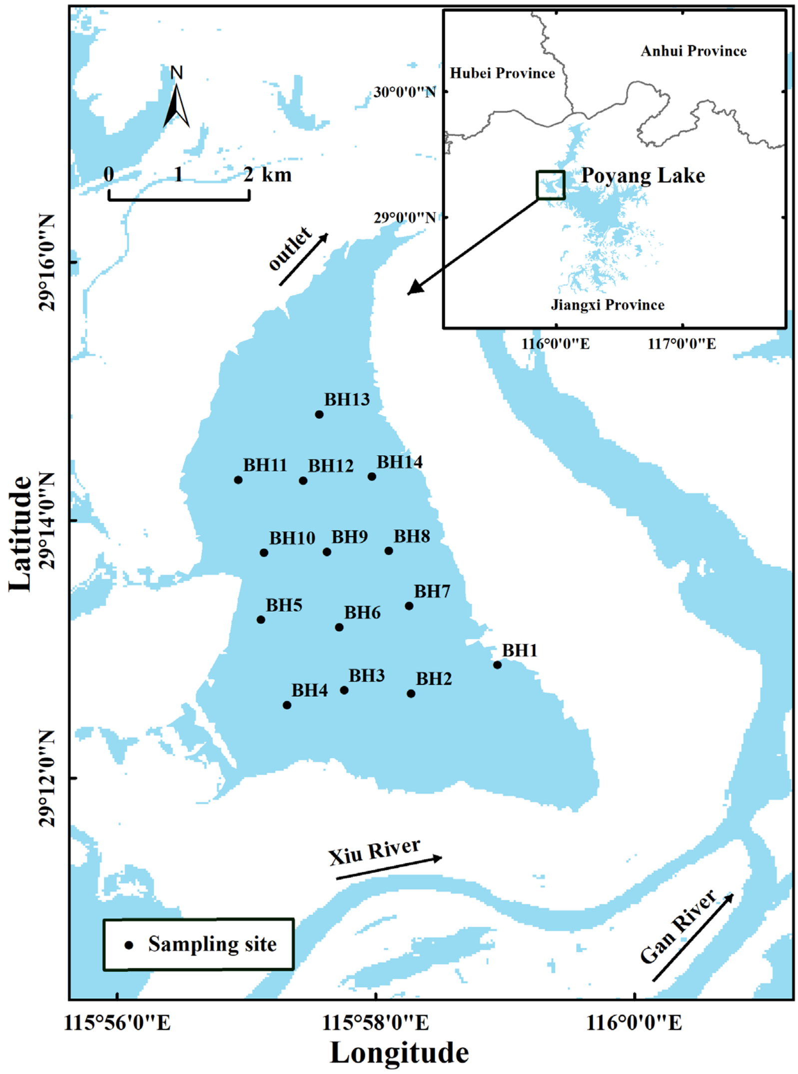

Figure 1. Location of Poyang Lake in Jiangxi Province, China, and 14 sampling sites situated in its coastal area (Banghu Lake)

the data while allowing a predefined margin of tolerance, known as ε-insensitive loss.[46] It supports various kernel functions (e.g., linear, radial basis, and polynomial) to capture nonlinear relationships.[47] The optimization objective is expressed as follows:[46,48]

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

$$\text{subject to: } \begin{cases} y_i - \langle w, x_i \rangle - b \le \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \quad (2)$$

where w is the weight vector that defines the orientation of the regression hyperplane; C is the regularization parameter (a large C emphasizes minimizing training errors, whereas a small C promotes generalization); n is the number of data points in the dataset; ξ_i and ξ_i^* are the slack variables representing the deviation from the ε-tube; ε is the margin of tolerance; y_i is the observed output (target) for the data point I; x_i is the input feature vector for the data point I; and b is the bias term that shifts the hyperplane.

2.3.2. Random forest

Random forest is an ensemble learning algorithm designed to mitigate the issues of overfitting and instability associated with single decision trees.[49] It constructs multiple decision trees on random subsets of the original training dataset and combines their outputs to enhance prediction accuracy and robustness. RF has been widely employed for both regression and classification tasks.[50] The inherent randomness in feature and sample selection improves generalization performance, and the final prediction is derived from the average output of all individual trees, as expressed below:

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} T_m(x) \tag{3}$$

where $\hat{y}$ is the final predicted value for input x, M is the number of decision trees, and $T_m(x)$ is the prediction from the m-th tree in the ensemble for a given input x.

2.3.3. Extreme gradient boosting

Extreme gradient boosting is a gradient boosting framework developed for high performance and scalability.[51-53] XGBoost constructs an ensemble of weak learners (typically decision trees) in an additive fashion by optimizing a regularized objective function.[54] The objective function (Obj) is defined as follows:[51]

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} T_m(x) \tag{4}$$

with the regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2 \tag{5}$$

where L is the differentiable convex loss function that measures the difference between the prediction and the target y_i, f_k is the decision tree k, the term Ω penalizes the complexity of the model (i.e., the regression tree function), γ is the regularization term for the number of leaves T, and λ is the penalty coefficient for the weight w.

2.3.4. K-nearest neighbors

K-nearest neighbors is a simple, nonparametric, and instance-based machine learning algorithm applicable to both regression and classification tasks.[55,56] KNN operates without model training by storing the entire dataset and predicting the output based on the proximity of neighboring samples. For each prediction, the algorithm identifies the nearest neighbors by calculating the distance between the test sample and all training samples, typically using a metric such as Euclidean distance.[57] The prediction can be determined based on either the average value (regression) or the majority vote (classification) of the KNN in the feature space, using a distance metric such as Euclidean distance.[55,58] In this study, for a given test sample x, the Euclidean distance (d) between x and a training sample x_i was calculated as:[59]

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} \left(x_j - x_{ij}\right)^2} \tag{6}$$

where $x = (x_1, x_2, \ldots, x_n)$ is the feature vector of the test sample, $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ is the feature vector of the i-th training sample, n is the number of features, x_j is the value of the j-th feature of the test sample, and x_{ij} is the value of the j-th feature for the i-th training sample. For a given test sample x, the predicted value is computed as the average of the target values of the K-nearest training samples as follows:

$$\hat{y} = \frac{1}{K}\sum_{i \in \mathcal{N}_K(x)} y_i \tag{7}$$

where $\hat{y}$ is the predicted value of the test sample x; K denotes the number of nearest neighbors; represents the set of indices corresponding to the K samples in the training set that are closest to x based on the Euclidean distance; and y_i is the actual observed target value of the i-th nearest neighbor.

2.4. Data processing

Prior to implementing MLMs, a Mantel test was conducted to examine the relationship between the multivariate structures of environmental factors and WQIs. This statistical approach is particularly suitable for assessing relationships within multivariate ecological and environmental datasets.[60-62] Specifically, the Mantel test evaluates the association between two distance matrices by computing the Spearman correlation coefficient and assessing its significance through permutation testing. The analysis was performed using the vegan package (version 0.9.8.1) in the R software (version 4.3.2), with the significance threshold set at $p < 0.05$.

Four MLMs were implemented using Scikit-learn and XGBoost libraries on the Google Colab platform (https://colab.research.google.com; June 26, 2026). As a cloud-based version of Jupyter Notebook, Colab was used to execute Python scripts. The input features comprised six environmental parameters: T, DO, pH, EC, ORP, and TDS. The target variables were three WQIs: TP, TN, and Chla. Before model training, the dataset was normalized and randomly partitioned into training (80%) and testing (20%) subsets, following a commonly adopted strategy that provides a reasonable balance between model training and independent performance evaluation.[63-65] Stratified sampling was employed to ensure consistent distribution of the target variables across subsets. Hyperparameters for each model were optimized via a grid search using a five-fold cross-validation on the training set.[66] For each target variable, the entire training dataset was randomly partitioned into five equal folds. In each iteration, four folds were used for model training, and the remaining fold served as the validation set to evaluate the performance of a given parameter combination. This process was repeated five times, and the average coefficient of determination (R^2) was used as the performance metric for that parameter set. The grid search enumerated all combinations within the predefined hyperparameter ranges for each model (Table S2), and the final optimal parameters were selected as the combination that yielded the highest average cross-validation R^2. This strategy was applied across all four algorithms, ensuring a standardized and rigorous model selection process while mitigating the risk of overfitting given the limited sample size.

Model performance was assessed on the testing set using three evaluation metrics: root-mean-squared error (RMSE), mean absolute error (MAE), and R^2, as defined below:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{10}$$

where y_i, $\hat{y}_i$, and $\bar{y}$ represent the observed, predicted, and mean of observed values, respectively.

Subsequently, SHAP was applied to interpret the selected best-performing model. The SHAP value is a weighted average of all possible differences and is calculated as follows:[25,67]

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\left(|N| - |S| - 1\right)!}{|N|!}\left[f_{(S \cup \{i\})} - f_{(S)}\right] \tag{11}$$

where ϕ_i is the SHAP value for feature i, N is the set of all features, S is a subset of features excluding feature i, $|S|$ is the number of features in subset S, $|N|$ is the total number of permutations of all features, $f_{(S\cup\{i\})}$ is the prediction after adding feature i, and $f_{(S)}$ is the model's prediction with feature subset S.

Given that the XGBoost model exhibited the highest predictive performance among the tested models (see Section 3), it was selected to predict WQIs from environmental parameters and to compare the predicted values with the observed data. The scripts used for non-image data analysis and for MLMs figure generation are available in Data S1.

3. Results

3.1. Water profile and water quality indicators

Distinct spatiotemporal patterns of water profiles were observed across the study area (Figure S1). Relatively higher Ts were recorded in the southern basin than in the northern basin, ranging from 21.0 °C to 23.6 °C and 31.7 °C to 34.1 °C in May and August, respectively. The DO concentrations exceeded 8.0 mg/L (8.786–10.500 mg/L) in May, while values below 7.3 mg/L (4.132–7.208 mg/L) were recorded in August. The pH ranged similarly in May (7.5–8.4) and August (7.1–8.6), with average values of 8.0 and 7.8, respectively. The lowest pH consistently occurred at the BH1 site, whereas relatively higher pH levels were observed in the northern basin. A similar spatial distribution pattern was identified for EC and TDS in May, with elevated values detected in the western region compared with the eastern region, particularly at sites BH10 and BH11. However, no such spatial consistency was observed between EC and TDS in August. ORP exhibited an inverse spatiotemporal distribution pattern, with relatively higher values in the northern region in May and in the southern region in August. The distribution patterns of environmental parameters indicate considerable variability across the study area.

Similarly, spatiotemporal variations were also evident in WQIs (Figure S2). TN concentrations were relatively higher in May (1.100–1.730 mg/L) than in August (0.606–0.974 mg/L). Across sampling sites, TN concentrations were consistently higher in the southern region than in the north. TP concentrations ranged from 0.037–0.110 mg/L in May and 0.049–0.072 mg/L in August, with a mean value of 0.059 mg/L (August). Chla concentrations ranged from 0.007–0.023 mg/L in May and 0.013–0.040 mg/L in August. Higher Chla concentrations were observed predominantly in the southern and northern regions in May and August, respectively.

3.2. Mantel test

Six pairs of significant correlations (based on Spearman correlation, $p < 0.05$) were identified among the tested environmental parameters, including T and DO, T and EC, DO and pH, DO and EC, pH and ORP, and EC and TDS (Figure 2). The Mantel test further confirmed that WQIs were significantly associated with variations in environmental parameters. TN and Chla exhibited strong correlations with T, DO, and EC ($p < 0.05$), whereas TP showed a significant correlation only with T ($p < 0.05$). These results suggest that T, DO, and EC may serve as key environmental factors that influence lake water quality.

3.3. Machine learning models

For the training data, the scatter plots of observed and predicted values approximated a 1:1 linear fit line (R^2 > 0.85 for all), whereas TP values predicted by the SVR model showed the largest deviation (Figure S3, Table S3).

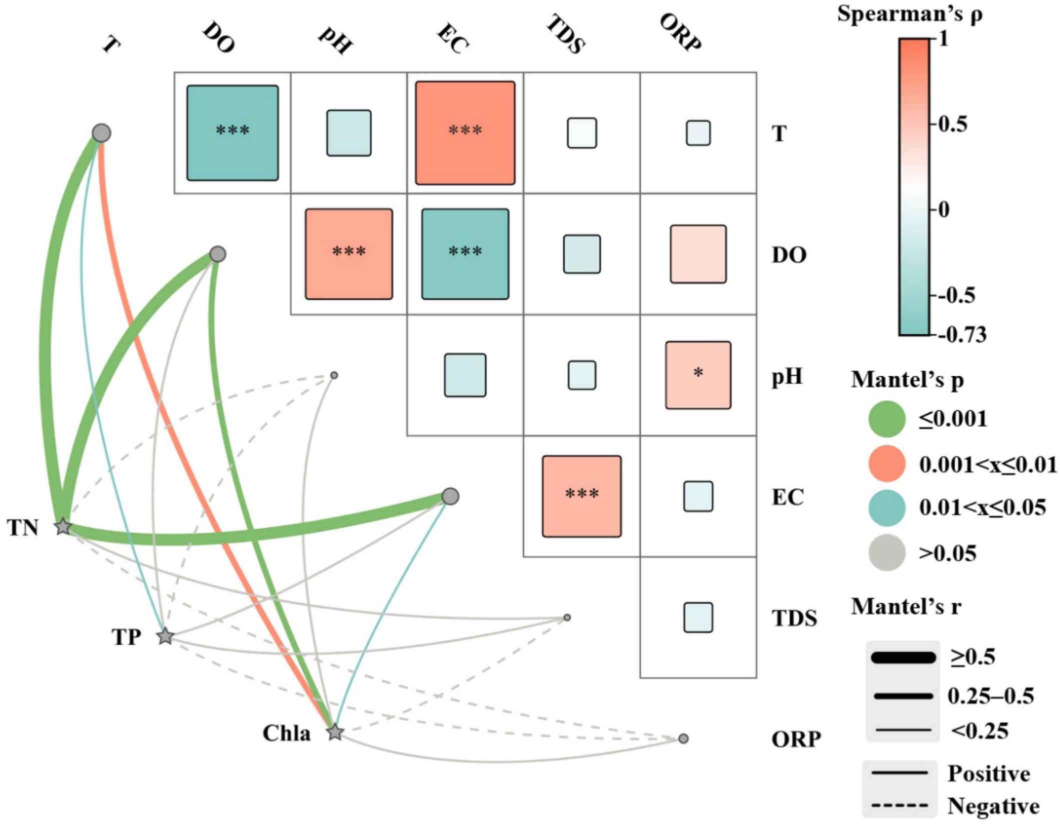

Figure 2. Mantel test results showing correlations among environmental parameters and their relationship with water quality indicators (total nitrogen [TN], total phosphorus [TP], chlorophyll-a [Chla]) and each environmental parameter (water temperature [T], dissolved oxygen [DO], water pH [pH], electrical conductivity [EC], oxidation–reduction potential [ORP], total dissolved solids [TDS]).
Notes: * $p < 0.05$, *** $p < 0.001$.

In the testing dataset, the predictive performance varied considerably across models (Figure S4). For TN, XGBoost achieved the best performance, with the lowest RMSE (0.06807) and the highest R^2 (0.97853) among all models (Table 1). For TP, SVR exhibited the lowest predictive accuracy, whereas RF, XGBoost, and KNN demonstrated comparable performance. Their RMSE values ranged from 0.00491 to 0.00603, MAE values ranged from 0.00381 to 0.00451, and R^2 values ranged from 0.86150 to 0.90818 (Table 1). For Chla, although RF yielded the lowest MAE (0.04489), XGBoost still demonstrated superior overall performance, with RMSE, MAE, and R^2 values of 0.05961, 0.04960, and 0.96993, respectively (Table 1).

The residual distributions also differed across prediction models. Broader and more dispersed residuals were observed in TN and Chla predictions from SVR and KNN, indicating poor reliability (Figure 3a and Figure 3c). In contrast, the TN residuals from XGBoost exhibited the narrowest range and the highest density among the three models (Figure 3a), indicating the best prediction accuracy for TN using XGBoost. For TP, the residuals from KNN, RF, and XGBoost displayed similar patterns, whereas those from SVR deviated significantly (Figure 3b). For Chla, both RF and XGBoost exhibited relatively low residual variability, confirming their excellent predictive performance (Figure 3c). These results indicate that XGBoost consistently produces reliable and acceptable predictions for simulating nutrient and phytoplankton dynamics. Residual skewness also varied across models. For TN, positive skewness was observed in SVR (1.146) and RF (0.204), whereas negative skewness was observed in XGBoost (−0.737) and KNN (−0.284) (Figure 3d). For TP, RF demonstrated negative skewness (−0.646), whereas the other models exhibited positive skewness (Figure 3e). For Chla, all four models yielded negative skewness, ranging from −1.225 to −0.661 (Figure 3f).

The SHAP analysis was conducted on the XGBoost model owing to its superior predictive performance. The SHAP values revealed the predicted characteristics of observations and their contributions to WQI estimation accuracy. EC, T, and DO were identified as the dominant predictors for both TN and Chla concentrations, contributing 35.2%, 34.6%, and 23.0% (92.8% in total) for TN (Figure 4a and 4b) and 43.7%, 24.7%, and 15.8% (84.2% in total) for Chla (Figure 4g and 4h). In contrast, ORP, pH, and TDS exhibited minor contributions (<5.5%) to TN and Chla. For TP, a distinct set of key predictors was identified, with ORP, pH, and T contributing 27.1%, 23.7%, and 19.8% (70.6% in total), respectively, whereas EC and DO contributed moderately (13.2% and 13.0%, respectively), and TDS accounted for only 3.2% (Figure 4d and 4e). In the decision diagram ranking predictor importance, individual prediction characteristics were ordered based on SHAP values (Figure 4c, 4f, and 4i). Predictors ranked higher in importance, such as EC, T, and DO for TN and Chla and ORP, pH, and T for TP, exhibited more pronounced SHAP value fluctuations than the lower-ranked variables.

Table 1. Parameter metrics of the four machine learning models (MLMs)

WQIs	MLMs	Performance metrics		
		RMSE	MAE	R^2
TN	SVR	0.13398	0.09869	0.87808
	RF	0.07467	0.06087	0.96213
	XGBoost	0.06807	0.06313	0.97853
	KNN	0.09706	0.08132	0.93602
TP	SVR	0.00979	0.00903	0.63477
	RF	0.00539	0.00381	0.88930
	XGBoost	0.00603	0.00451	0.86150
	KNN	0.00491	0.00433	0.90818
Chla	SVR	0.16868	0.11398	0.75919
	RF	0.06638	0.04489	0.96270
	XGBoost	0.05961	0.04960	0.96993
	KNN	0.22340	0.15930	0.57764

Abbreviations: Chla: Chlorophyll-a; KNN: k-nearest neighbors; MAE: Mean absolute error; R^2: Coefficient of determination; RMSE: Root-mean-squared error; RF: Random forest; SVR: Support vector regression; TN: Total nitrogen; TP: Total phosphorus; XGBoost: Extreme gradient boosting.

The complete hyperparameter ranges and the optimal parameters for all four models are summarized in Table S2. Notably, the optimal XGBoost configurations varied across the three target variables, reflecting the varying complexities and data distributions of the three WQIs. The predicted spatiotemporal distributions of WQIs closely matched the observed values (Figures 5 and S2). Linear regression analyses demonstrated a strong agreement between the predicted and observed target WQIs, with R^2 values exceeding 0.988 for all parameters (Figure 6). Most predicted values followed the temporal trends of observed data, confirming the model's ability to capture dynamic patterns effectively. Although relatively large deviations were observed at specific sampling sites (Figure 6), such as TN at BH8 in August and Chla at BH1 in May (Figures 5 and S2), the overall agreement between observed and predicted values confirmed high predictive reliability.

4. Discussion

4.1. Model comparison

Among the four evaluated MLMs, XGBoost consistently outperformed the others in terms of predictive accuracy and robustness, confirming its suitability for modeling complex and nonlinear environmental systems.[68] A previous study utilized five machine learning algorithms to predict nitrate nitrogen concentrations in the lower Ganga River (Pakistan) based on water quality and meteorological data,[69] in which XGBoost achieved the highest predictive performance (R^2 = 0.85), followed by RF (R^2 = 0.80). In contrast, traditional models such as multiple polynomial regression[70] and generalized additive models yielded lower accuracy. Niazkar *et al.*[71] assessed the predictive performance of 11 MLMs for groundwater quality in the arid province of Yazd,[72] and XGBoost demonstrated stable reliability across varying scenarios. These findings further emphasize the strength of XGBoost in capturing complex environmental interactions and underscore its broad potential for water-quality prediction. In our study, XGBoost achieved high R^2 values for both the training set (0.907–0.998; Table S3) and the test set (0.861–0.979; Table 1). Moreover, the training and validation plots show comparable predictive performance between the training and test datasets, providing no substantial overfitting (Figures S3 and S4). In contrast, the SVR and KNN models exhibited high R^2 values (>0.855) in the training set but substantially lower R^2 values (approximately 0.6) in the test set, suggesting potential overfitting. Although multicollinearity among environmental variables may exist because some water quality parameters are ecologically interrelated, only a limited proportion of variable pairs showed strong correlations in this study (Figure 2), suggesting that severe multicollinearity was unlikely among the selected predictors. Moreover, XGBoost is generally less sensitive to multicollinearity than linear regression approaches because it does not rely on estimating independent regression coefficients for correlated predictors.[51] However, multicollinearity may still influence the interpretation of feature importance when correlated predictors provide redundant information. Because the SHAP values in this

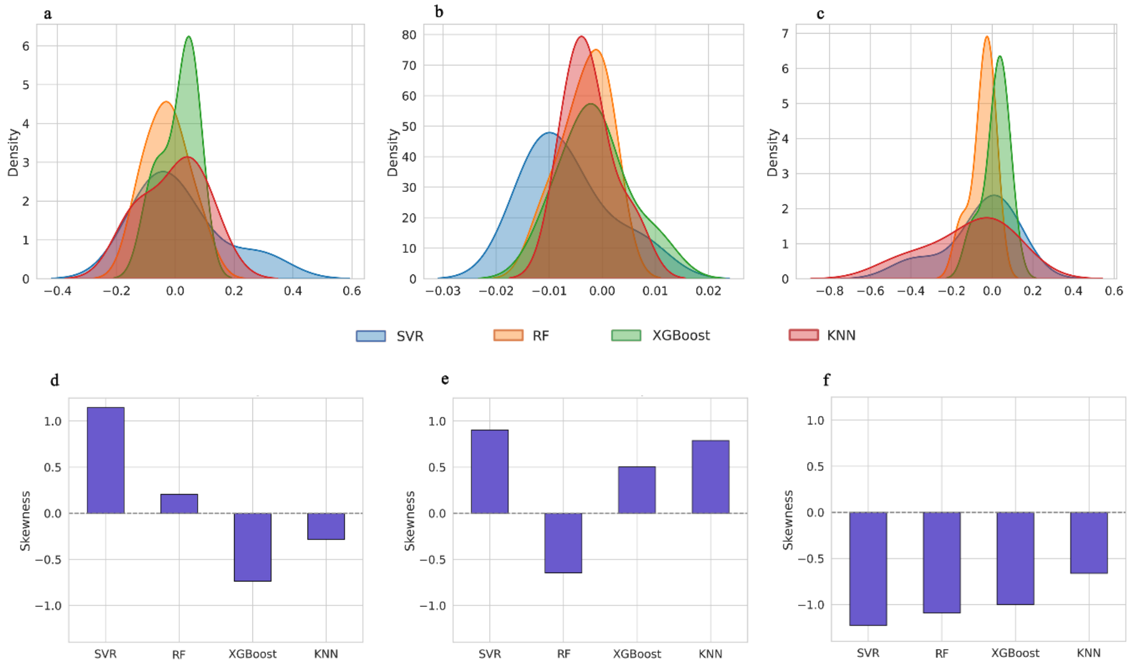

Figure 3. Density distribution of residuals from four machine learning models for predicting water quality indicators using test data: (a) TN, (b) TP, and (c) density. Skewness values of the residual distributions are also shown for (d) TN, (e) TP, and (f) density.
Abbreviations: Chla: Chlorophyll-a; KNN: k-nearest neighbors; RF: Random forest; SVR: Support vector regression; TN: Total nitrogen; TP: Total phosphorus; XGBoost: Extreme gradient boosting.

study were interpreted as the contribution of each variable to the model predictions rather than as evidence of causal relationships, additional time-series analyses, such as wavelet analysis and convergent cross mapping, are needed to further investigate the potential causal relationships among environmental variables.[13,15,73,74]

Although the dataset from Poyang Lake used in this study was relatively small and lacked long-term temporal observations, the close agreement between predicted and observed values suggested that the selected input features, including physicochemical parameters, contained sufficient predictive information to capture the dynamics of water quality fluctuations.[75] Tree-based ensemble methods generally require large datasets to reduce the risk of overfitting; recent studies have demonstrated that XGBoost can achieve competitive performance and robustness even with relatively small datasets.[76,77] XGBoost incorporates several mechanisms to reduce the risk of overfitting, including regularized objective functions, tree pruning, shrinkage, and subsampling strategies. These approaches constrain model complexity and improve the generalization ability of the model, allowing XGBoost to achieve robust predictive performance even with relatively small datasets.[51] These findings highlight the applicability of ensemble learning approaches for analyzing complex ecological datasets and support the potential use of XGBoost as a promising tool for water quality assessment and environmental decision-making.[78] As this study represents an initial exploratory effort to evaluate the feasibility of MLMs for water-quality prediction in Poyang Lake, the application of XGBoost provides a preliminary predictive framework under current data limitations. Water quality is influenced by multiple factors, including seasonality, spatial scale, and land-use characteristics, and water quality parameters are generally better explained during the rainy season than during the dry season.[79] Future research will focus on expanding long-term monitoring datasets and performing independent temporal validation. With larger datasets, bootstrap-based confidence intervals can also be employed to quantify the uncertainty in model predictions and to further assess the robustness, reliability, and transferability of the proposed model.

The residual distribution of XGBoost showed negative skewness for both TN and Chla predictions, indicating a general underestimation of these variables. This pattern may be attributed to the conservative regularization constraints embedded in XGBoost, which can effectively reduce overfitting while limiting the model's

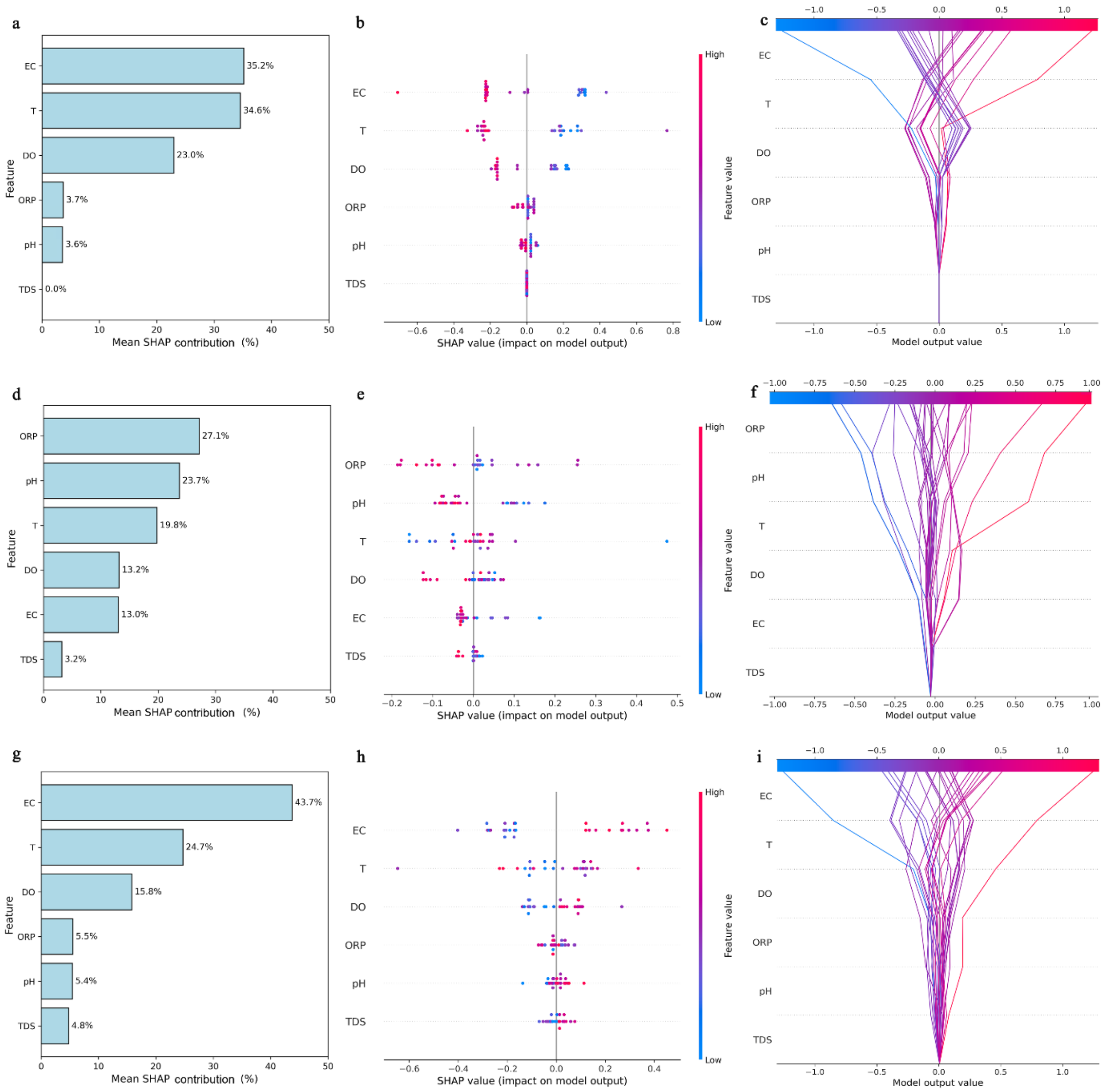

Figure 4. SHapley Additive exPlanations (SHAP) analysis of the effect of six input variables (T, DO, pH, EC, ORP, TDS) on the XGBoost model's predictive ability for three water quality indicators: (a–c) TN, (d–f) TP, and (g–i) Chla. (a, d, g) SHAP bar plots. (b, e, h) SHAP dot plots. (c, f, i) SHAP decision plots. Each dot represents the SHAP value for one prediction (28 in total), with one dot for each feature per prediction. Variables were ranked in descending order of importance. Dot color indicates the actual value of each feature, with red representing higher feature values and blue representing lower values.
Abbreviations: Chla: Chlorophyll-a; DO: Dissolved oxygen; EC: Electrical conductivity; ORP: Oxidation–reduction potential; pH: Water pH; T: Water temperature; TDS: Total dissolved solids; TN: Total nitrogen; TP: Total phosphorus; XGBoost: Extreme gradient boosting.

responsiveness to elevated target values.[51] Although RF, XGBoost, and KNN exhibited acceptable predictive performance for TP, XGBoost displayed the lowest skewness, suggesting that it is the most reliable model. Nonetheless, a slight positive skewness was observed in TP predictions using XGBoost, suggesting a potential overestimation tendency. Phosphorus, serving as a key limiting nutrient regulating primary production in aquatic ecosystems, can be influenced by complex interactions between internal loading and external sources related to biological processes.[45,80] These interactions contribute to the variability in the phosphorus cycle across natural waters, posing challenges for TP prediction with current environmental parameters. Consequently, appropriate

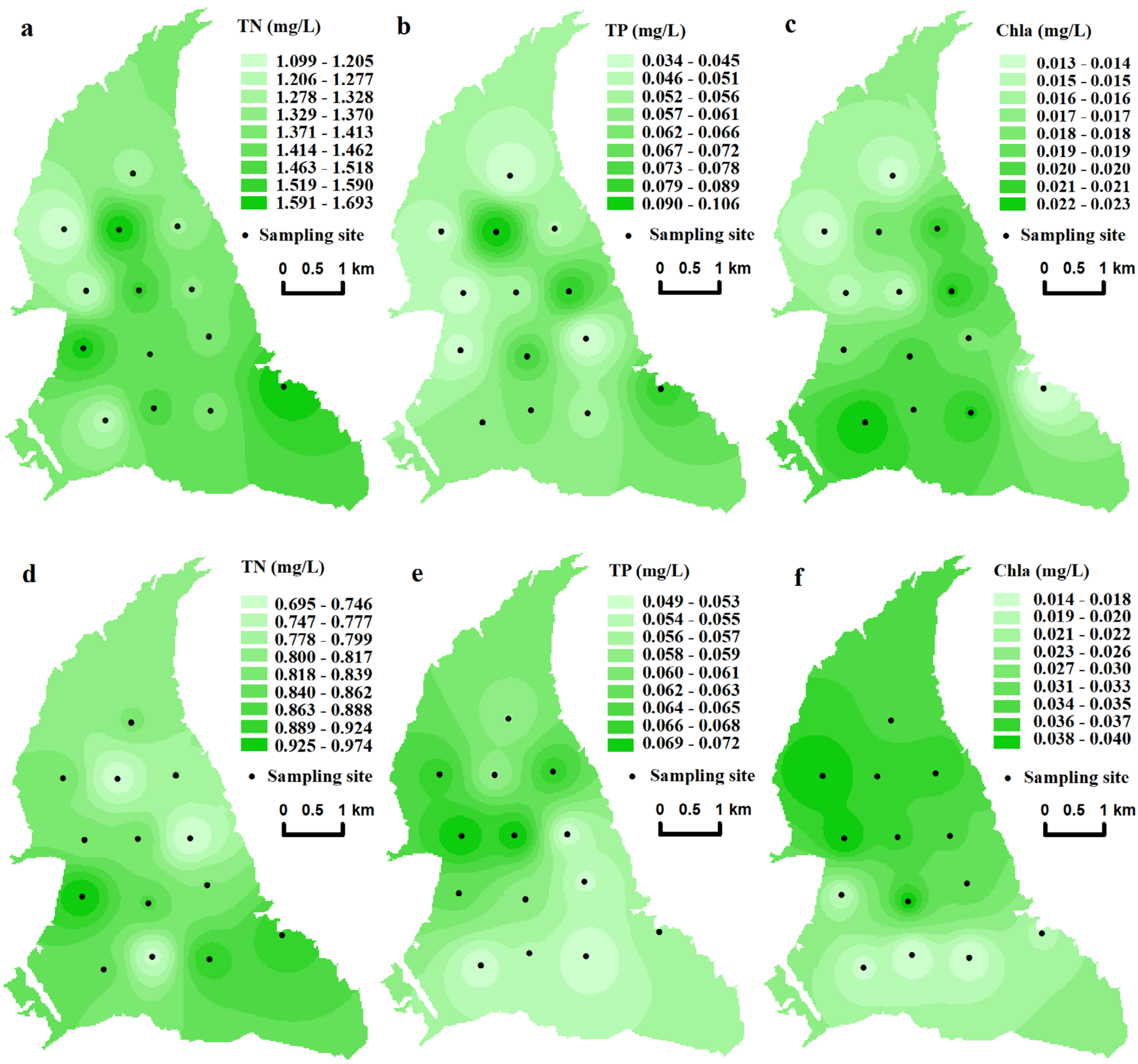

Figure 5. Predicted spatiotemporal distribution of water quality indicators in the coastal area of Poyang Lake using the extreme gradient boosting (XGBoost) model. The (a–c) upper and (d–f) lower rows represent the predictions for May 10 and August 10, 2024, respectively. (a, d) Total nitrogen (TN). (b, e) Total phosphorus (TP). (c, f) Chlorophyll-a (Chla).

MLMs should be selected with caution. Notably, all MLMs demonstrated negatively skewed residual distributions for Chla, reflecting a consistent underestimation pattern. This result indicates that current models may struggle to capture the full scope of Chla dynamics, possibly because of the intricate interplay between environmental drivers and this biotic indicator.

Support vector regression exhibited suboptimal performance in water-quality prediction, characterized by wide residual spreads and elevated residual standard deviations across all evaluated WQIs (Figure S5). SVR has been applied in various fields, such as time series analysis, financial prediction, and the approximation of complex engineering analyses.[48] A previous study demonstrated that SVR outperformed RF and linear regression in predicting key irrigation water quality parameters, highlighting its effectiveness for real-time groundwater quality monitoring and management; however, limitations include reliance on data from specific periods and potential long-term or extreme weather effects.[47] KNN and RF occasionally produced predictions similar to XGBoost,

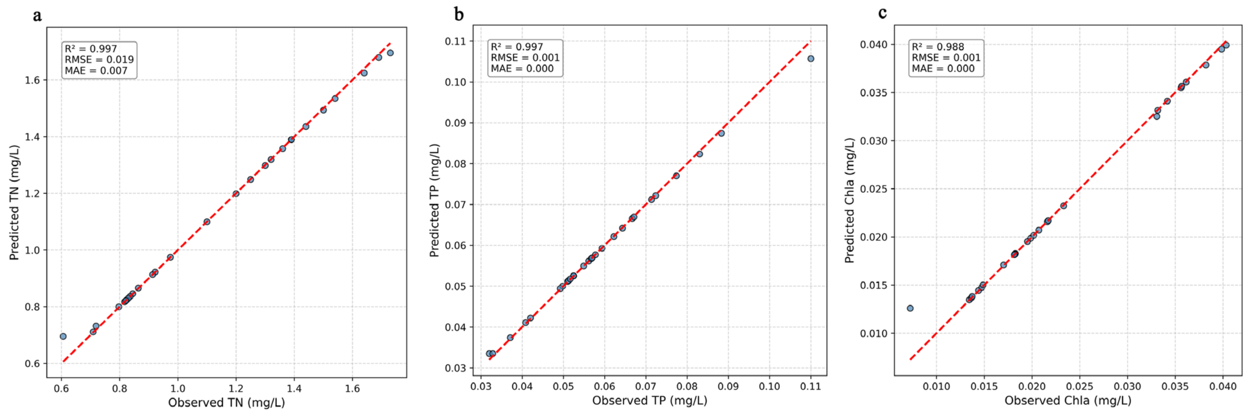

Figure 6. Scatter plots of observed versus predicted values of water quality indicators using the extreme gradient boosting (XGBoost) model. (a) Total nitrogen (TN). (b) Total phosphorus (TP). (c) Chlorophyll-a (Chla). A linear regression line (dashed red) and corresponding statistical coefficients are also provided.
Abbreviations: MAE: Mean absolute error; R^2: Coefficient of determination; RMSE: Root-mean-squared error;

such as in the case of TP. The limited capacity of SVR to capture nonlinear relationships under sparse parameter tuning, combined with KNN's sensitivity to local noise and outliers, may partially account for these deficiencies.[58] In the present study, RF consistently showed negative skewness, indicating a general underestimation tendency. Prior studies have revealed that RF is sensitive to local data fluctuations, and its ensemble nature leads to overfitting in areas with high variance or extreme values.[49] Yu *et al.*[81] demonstrated that integrating RF with optimization techniques, such as adaptive boosting (AdaBoost), the Kepler optimization algorithm (KOA), and the genetic algorithm (GA), significantly enhanced its predictive accuracy for TP and TN concentrations, outperforming the conventional RF model. Bradter *et al.*[82] suggested that incorporating the area under the curve (AUC) metric effectively reduced the adverse effects of small sample sizes on the inference and predictive performance of RF models, thereby substantially improving their reliability and robustness. These findings collectively emphasize the critical role of algorithmic optimization and the selection of appropriate performance metrics in enhancing model adaptability to complex and data-limited environmental systems.[83] Recently, Transformer-based neural networks have emerged as a promising alternative for environmental prediction because their self-attention mechanism enables the learning of complex nonlinear relationships and long-range temporal dependencies within multivariate time-series data.[84,85] Transformer-based architectures have shown superior performance compared with conventional MLMs when large and continuous monitoring datasets are available. However, these models generally require substantially larger datasets for effective training, which may limit their applicability to relatively small environmental datasets, such as that used in the present study. Comparisons with established deep learning models, including XGBoost and LSTM, have indicated that these approaches performed well for short-term forecasting, whereas their performance may be less satisfactory for long-term prediction.[84]

4.2. Influential environmental parameters and predictions

Both the Mantel test and SHAP analysis consistently identified EC, T, and DO concentrations as key environmental drivers influencing TN and Chla concentrations. Significant correlations (Spearman correlation, $p < 0.001$ for all) among EC, T, and DO concentrations were further confirmed (Figure 2). Because the Mantel test was not designed to quantify the contribution of each parameter in predictive modeling, XGBoost combined with SHAP analysis was employed to evaluate their influence. The results demonstrated that EC, T, and DO were the three most influential predictors, jointly accounting for 92.8% of the total SHAP importance for TN prediction. Similarly, for Chla prediction, their combined SHAP importance reached 84.2%, with the remaining 15.8% attributed to other environmental variables. Notably, correlations among environmental variables are common in aquatic ecosystems because many physicochemical parameters are governed by the same ecological and biogeochemical processes.[40] Consequently, a certain degree of multicollinearity is expected in environmental datasets. Although EC and TDS were highly correlated, only EC showed a substantially greater contribution in the XGBoost–SHAP analysis, particularly in TN and Chla (Figure 4). These findings suggest that XGBoost can effectively capture complex nonlinear relationships and distinguish the relative predictive importance of correlated environmental variables.

In aquatic ecosystems, EC reflects the total concentration of dissolved ions and therefore serves as an integrative indicator of watershed inputs and hydrological conditions.[86-88] Seasonal fluctuations in water level and river-lake connectivity strongly regulate the transport, exchange, and dilution of dissolved substances.[89] The dominant role of EC in nitrogen dynamics corroborates previous findings from agricultural watershed studies.[90] The coastal region of Poyang Lake, which is characterized as a typical agricultural watershed, is subject to substantial external nutrient input.[91] In this study, a significant positive correlation between TN and EC was observed (Mantel test, $r = 0.791$, $p < 0.001$; Figure 2). As EC commonly reflects the level of external nutrient loading, it is typically associated with TN concentration.[92] Unlike TN, Chla represents phytoplankton biomass rather than nutrient concentration. The importance of EC in predicting Chla potentially reflects its role as an integrated indicator of the physicochemical environment that supports phytoplankton growth.[93,94] Therefore, EC may indirectly capture these combined environmental effects, thereby conferring high predictive importance for Chla. The effects of T and DO on TN and Chla are attributed to both abiotic and biotic processes. A recent machine learning-based study in the lower Ganga River (India) identified T as a major driver of nitrogen concentration.[69] Elevated temperatures promote nutrient release from sediments and stimulate nitrifying bacteria, leading to nitrate accumulation.[95] Furthermore, DO was reported to significantly enhance nitrification efficiency, thereby regulating nitrogen dynamics.[96] Increased T and DO levels can often promote phytoplankton proliferation and consequently elevate Chla concentrations.[97,98] These observations are consistent with findings from stochastic transfer function modeling, which confirm the strong predictive power of T and DO for Chla concentrations.[99] Using linear mixed-effects models, Li *et al.*[100] further verified the dominant influence of T, DO, and EC on controlling TN variations along hydrological gradients in Poyang Lake.

The environmental factors influencing TP exhibited distinct patterns compared with those influencing TN and Chla. While the Mantel test identified T as the most influential variable for TP, SHAP analysis further revealed that ORP and pH were two additional important predictors. A significant positive correlation was observed between ORP and pH (Spearman correlation, $\rho = 0.453$, $p < 0.05$; Figure 2). The potential of MLM-based analysis to capture complex ecological relationships is supported by SHAP results.[67] In contrast, the Mantel test is limited to identifying monotonic associations and may overlook such nonlinear patterns.[101] A decrease in ORP was generally accompanied by an increase in TP concentration. The changes in ORP under resulting redox conditions indirectly affect TP dynamics by influencing the release and deposition of phosphorus between sediments and the overlying water column.[102] In addition, reductive conditions can dissolve iron compounds and subsequently mobilize phosphorus into the water column.[103] Buffering mechanisms in aquatic environments may further decouple phosphorus cycling through pH fluctuations.[104] Despite these nonlinear interactions, the consistent identification of T across both traditional statistical approaches and MLMs highlights its fundamental role in regulating phosphorus dynamics. Liikanen *et al.*[105] reported that a 3 °C increase under both aerobic and anaerobic conditions led to a maximum increase in TP release of 57%. Similar observations have been reported in mesocosm experiments.[106] As TP predictions by the XGBoost model relied on moderate contributions from ORP, pH, and T (approximately 20%), these findings underscore the need to incorporate additional variables, such as sediment phosphorus and inflow rates, to capture TP dynamics more comprehensively and improve model accuracy.

5. Conclusion

This study systematically compared the predictive performance of four MLMs (SVR, RF, XGBoost, and KNN) for estimating WQIs (TN, TP, Chla) in the coastal zone of Poyang Lake and interpreted the underlying environmental drivers using SHAP analysis. Among the evaluated models, XGBoost consistently achieved the highest predictive accuracy, demonstrating its suitability for water-quality prediction based on physicochemical parameters. SHAP analysis further identified EC, T, DO, ORP, and pH as the dominant predictors, although the key controlling factors differed among TN, TP, and Chla. These results demonstrate that integrating XGBoost with SHAP not only provides accurate predictions but also improves model interpretability by quantifying the relative contributions of individual environmental variables. This study provides a practical and interpretable machine learning framework for water-quality assessment and environmental management in freshwater ecosystems. From a management perspective, monitoring key environmental parameters may provide a cost-effective approach for assessing eutrophication risk, consequently reducing the need for frequent and labor-intensive nutrient analyses. However, because the present analysis was based on a relatively short monitoring period, future research should incorporate long-term monitoring datasets and temporal validation to further improve model reliability and transferability.

Acknowledgments

We thank the staff of Gannan Normal University for their assistance with field sampling and for their support of this study.

Funding

This work was supported by the Natural Science Foundation of Jiangxi Province (Grant number 20212BAB213012), the Natural Science Foundation of Guangxi Zhuang Autonomous Region (Grant number 2025GXNSFAA069312), and the National Natural Science Foundation of China (Grant number 41907331).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Zhangtao Hu, Rong Yi, Xin Liu
Data curation: Zhangtao Hu
Formal analysis: Zhangtao Hu
Investigation: Zhangtao Hu, Rong Yi
Methodology: Zhangtao Hu, Xin Liu
Project administration: Rong Yi, Xin Liu, Chao Du
Supervision: Rong Yi, Xin Liu
Visualization: Zhangtao Hu
Writing–original draft: Zhangtao Hu
Writing–review & editing: Rong Yi, Chao Du, Xinyue Di, Xin Liu, Qiang Huang, Aimin Hao, Yasushi Iseri

Availability of data

The link to the Python script and the data used can be found in the GitHub repository (https://github.com/liuxinrk/LPoyang_MLMs).

References

1. Kawanabe H, Nishino M, Maehata M, eds. *Lake Biwa: Interactions between Nature and People*. Dordrecht, Netherlands: Springer Nature; 2012.

   doi: 10.1007/978-94-007-1783-1

2. Yadav SC. Water Pollution: The Problems and Solutions. *Sci Insights.* 2024;44(2):1245-1251.

   doi: 10.15354/si.24.re905

3. Sheng Y, Chen F, Sheng G, Fu J. Water quality remediation in a heavily polluted tidal river in Guangzhou, South China. *Aquat Ecosyst Health Manag.* 2012;15(2):219-226.

   doi: 10.1080/14634988.2012.687674

4. Zhu DT, Cheng XJ, Li WH, Niu FJ, Wen JH. Temporal and spatial variation characteristics of water quality in the middle and lower reaches of the Lijiang River, China and their responses to environmental factors. *Int J Environ Res Public Health.* 2022;19(13):8089.

   doi: 10.3390/ijerph19138089

5. Cohen J. Multiple regression as a general data-analytic system. *Psychol Bull.* 1968;70(6, pt 1):426-443.

   doi: 10.1037/h0026714

6. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol.* 1933;24(6):417-441.

   doi: 10.1037/h0071325

7. Tang ZH, Cao YH, Jiang QF. Spatial and temporal changes of water environmental factors and water quality assessment in coastal waters of Jiangsu Province. *J Sea Res.* 2025;204:102570.

   doi: 10.1016/j.seares.2025.102570

8. Heneash AM, Alprol AE, El-Naggar HA, *et al*. Multivariate analysis of plankton variability and water pollution in two highly dynamic sites, southeastern Mediterranean (Egyptian coast). *Arab J Geosci.* 2022;15(4):330.

   doi: 10.1007/s12517-022-09595-1

9. Jakovljević OS, Popović SS, Živić IM, *et al*. Epilithic diatoms in environmental bioindication and trout farm's effects on ecological quality assessment of rivers. *Ecol Indic.* 2021;128:107847.

   doi: 10.1016/j.ecolind.2021.107847

10. Mu M, Gao LM, Zhang HQ, *et al*. Effects of land use on water quality at different spatial scales in the middle reaches of Huaihe River. *J Freshw Ecol.* 2023;38(1):2176373.

    doi: 10.1080/02705060.2023.2176373

11. Li BD, Chen KQ, Liu XB, Liu C, Wang SY. Evolution Characteristics of Water Quality in Plain Reservoirs and Its Relationship with the Economic Development Response: A Case Study of Daheiting Reservoir in Northern China. *Water.* 2023;15(18):3229.

    doi: 10.3390/w15183229

12. Mebarkia A, Boufekane A. Human activity impact on surface water quality in semi-arid regions: a case study of Aïnzeda lake (North-East Algeria). *Water Supply.* 2020;20(5):1726-1744.

    doi: 10.2166/ws.2020.083

13. Sugihara G, May R, Ye H, *et al*. Detecting causality in complex ecosystems. *Science.* 2012;338(6106):496-500.

    doi: 10.1126/science.1227079

14. Liu X, Dur G, Ban S, Sakai Y, Ohmae S, Morita T. Planktivorous fish predation masks anthropogenic disturbances on decadal trends in zooplankton biomass and body size structure in Lake Biwa, Japan. *Limnol Oceanogr.* 2020;65(3):667-682.

    doi: 10.1002/lno.11336

15. Chang CW, Miki T, Ye H, *et al.* Causal networks of phytoplankton diversity and biomass are modulated by environmental context. *Nat Commun.* 2022;13(1):1140.

doi: 10.1038/s41467-022-28761-3

16. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science.* 2015;349(6245):255-260.

doi: 10.1126/science.aaa8415

17. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. *SN Comput Sci.* 2021;2(3):160.

doi: 10.1007/s42979-021-00592-x

18. Thanh NN, Chotpantarat S, Ngu NH, Thunyawatcharakul P, Kaewdum N. Integrating machine learning models with cross-validation and bootstrapping for evaluating groundwater quality in Kanchanaburi province, Thailand. *Environ Res.* 2024;252:118952.

doi: 10.1016/j.envres.2024.118952

19. Uddin MG, Nash S, Rahman A, Olbert AI. Performance analysis of the water quality index model for predicting water state using machine learning techniques. *Process Saf Environ Prot.* 2023;169:808-828.

doi: 10.1016/j.psep.2022.11.073

20. Irwan D, Ibrahim SL, Latif SD, *et al.* River water quality monitoring using machine learning with multiple possible in-situ scenarios. *Environ Sustain Indic.* 2025;26:100620.

doi: 10.1016/j.indic.2025.100620

21. Kruk M. SHAP-NET, a network based on Shapley values as a new tool to improve the explainability of the XGBoost-SHAP model for the problem of water quality. *Environ Model Softw.* 2025;188:106403.

doi: 10.1016/j.envsoft.2025.106403

22. Chen C, Chen Q, Yao S, *et al.* Combining physical-based model and machine learning to forecast chlorophyll-a concentration in freshwater lakes. *Sci Total Environ.* 2024;907:168097.

doi: 10.1016/j.scitotenv.2023.168097

23. Huang HJ, Wang W, Lv JP, *et al.* Relationship between chlorophyll a and environmental factors in lakes based on the random forest algorithm. *Water.* 2022;14(19):3128.

doi: 10.3390/w14193128

24. Isah AS, Bao PQ, Gaurav S, *et al.* Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. *Environ Sci Pollut Res.* 2020;27(33):41524-41539.

doi: 10.1007/s11356-020-09689-x

25. Shapley LS. A Value for n-Person Games. In: Kuhn HW, Tucker AW, eds. *Contributions to the Theory of Games (AM-28).* Volume II. Princeton, NJ: Princeton University Press; 1953:307-318.

doi: 10.1515/9781400881970-018

26. Li ZQ. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Comput Environ Urban Syst.* 2022;96:101845.

doi: 10.1016/j.compenvurbsys.2022.101845

27. Makumbura RK, Mampitiya L, Rathnayake N, *et al.* Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like SHapley Additive exPlanations (SHAP) for interpreting the black-box nature. *Results Eng.* 2024;23:102831.

doi: 10.1016/j.rineng.2024.102831

28. Talukdar S, Bera S, Naikoo MW, *et al.* Optimisation and interpretation of machine and deep learning models for improved water quality management in Lake Loktak. *J Environ Manage.* 2024;351:119866.

doi: 10.1016/j.jenvman.2023.119866

29. Wang YY, Yu XB, Zhang L, Lei GC. Seasonal variability in baseline δ15N and usage as a nutrient indicator in Lake Poyang, China. *J Freshw Ecol.* 2013;28(3):365-373.

doi: 10.1080/02705060.2013.763296

30. Liang Y, Liu XZ, Xiao HY, Gao XF, Li WH, Xiong J. Impact of high water level fluctuations on stable isotopic signature of POM and source identification in a floodplain lake—Bang Lake (Poyang Lake). *Environ Earth Sci.* 2016;75(3):255.

doi: 10.1007/s12665-015-4847-z

31. Yu X, Hawley-Howard J, Pitt AL, Wang J-J, Baldwin RF, Chow AT. Water quality of small seasonal wetlands in the Piedmont ecoregion, South Carolina, USA: effects of land use and hydrological connectivity. *Water Res.* 2015;73:98-108.

doi: 10.1016/j.watres.2015.01.007

32. Wu ZS, Cai YJ, Zhang L, Chen YW. Spatial and temporal heterogeneities in water quality and their potential drivers in Lake Poyang (China) from 2009 to 2015. *Limnologica.* 2018;69:115-124.

doi: 10.1016/j.limno.2017.12.001

33. Tian MM, Mao JQ, Wang K, Xu DD. Spatio-temporal heterogeneity of ecological water level in Poyang Lake, China. *Ecol Inform.* 2024;82:102694.

doi: 10.1016/j.ecoinf.2024.102694

34. Wang ZT, Sun FH, Sang YW, Wu FC. Drivers analysis and future scenario-based predictions of nutrient loads in key lakes and reservoirs of the Yangtze River Catchment. *J Environ Manage.* 2025;374:124078.

doi: 10.1016/j.jenvman.2025.124078

35. Fang N, You QH, Liu L, *et al*. Evaluation of eutrophication in Poyang Lake wetland during autumn based on the cloud model [in Chinese]. *Acta Ecol Sin.* 2019;39(17):1-8.

doi: 10.5846/stxb201808221790

36. Liu YB, Liu J, Yin XJ, *et al*. Human activities and ecosystem health: A historical analysis of Poyang Lake. *Ecol Indic.* 2024;166:112446.

doi: 10.1016/j.ecolind.2024.112446

37. Tan YB, Hu N, Huang MT, Xiao YT, Shan J, Li DJ. Spatial–Temporal Evolution and Correlation Analysis of Human Activity Intensity and Resource Carrying Capacity in the Region around Poyang Lake, China, from 2010 to 2020. *Land.* 2023;12(12):2139.

doi: 10.3390/land12122139

38. Yao Z, Wang Z, Huang J, Xu N, Cui X, Wu T. Interpretable prediction, classification and regulation of water quality: A case study of Poyang Lake, China. *Sci Total Environ.* 2024;951:175407.

doi: 10.1016/j.scitotenv.2024.175407

39. Smith VH, Schindler DW. Eutrophication science: where do we go from here? *Trends Ecol Evol.* 2009;24(4):201-207.

doi: 10.1016/j.tree.2008.11.009

40. Wetzel RG. *Limnology: Lake and River Ecosystems*. 3rd ed. San Diego, CA: Academic Press; 2001.

41. Yi R, Song P, Liu X, Maruo M, Ban S. Differences in dissolved phosphate in shallow-lake waters as determined by spectrophotometry and ion chromatography. *Limnology.* 2020;21(3):329-339.

doi: 10.1007/s10201-019-00574-2

42. Xu H, Paerl HW, Qin B, Zhu G, Gaoa G. Nitrogen and phosphorus inputs control phytoplankton growth in eutrophic Lake Taihu, China. *Limnol Oceanogr.* 2010;55(1):420-432.

doi: 10.4319/lo.2010.55.1.0420

43. TEZUKA Y. Seasonal variations of dominant phytoplankton, chlorophyll a and nutrient levels in the pelagic regions of Lake Biwa. *Jpn J Limnol.* 1984;45(1):26-37.

doi: 10.3739/rikusui.45.26

44. McQuatters-Gollop A, Raitsos DE, Edwards M, *et al*. A long-term chlorophyll dataset reveals regime shift in North Sea phytoplankton biomass unconnected to nutrient levels. *Limnol Oceanogr.* 2007;52(2):635-648.

doi: 10.4319/lo.2007.52.2.0635

45. Liu X, Yi R, Ohashi M, *et al*. Spatiotemporal variations of orthophosphate in a large mesotrophic lake and implications for phytoplankton productivity. *Inland Waters.* 2025;15(1):1-37.

doi: 10.1080/20442041.2025.2449763

46. Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. Support vector regression machines. In: *Advances in Neural Information Processing Systems 9*. 1996:155-161.

47. Raheja H, Goel A, Pal M. Analysis and prediction of groundwater quality using machine learning algorithm for irrigation purposes. *Environ Earth Sci.* 2025;84(7):180.

doi: 10.1007/s12665-025-12173-9

48. Basak D, Pal S, Patranabis DC. Support vector regression. *Neural Inf Process Lett Rev.* 2007;11(10):203-224.

49. Breiman L. Random forests. *Mach Learn.* 2001;45:5-32.

doi: 10.1023/A:1010933404324

50. He S, Wu J, Wang D, He X. Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. *Chemosphere.* 2022;290:133388.

doi: 10.1016/j.chemosphere.2021.133388

51. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM; 2016:785-794.

doi: 10.1145/2939672.2939785

52. Dhaliwal SS, Nahid A-A, Abbas R. Effective intrusion detection system using XGBoost. *Information.* 2018;9(7):149.

doi: 10.3390/info9070149

53. Niazkar M, Menapace A, Brentan B, *et al*. Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). *Environ Model Softw.* 2024;174:105971.

doi: 10.1016/j.envsoft.2024.105971

54. Ding F, Hao SL, Zhang WJ, *et al*. Using multiple machine learning algorithms to optimize the water quality index model and their applicability. *Ecol Indic.* 2025;172:113299.

doi: 10.1016/j.ecolind.2025.113299

55. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory.* 1967;13(1):21-27.

doi: 10.1109/tit.1967.1053964

56. Fix E. *Discriminatory Analysis: Nonparametric Discrimination, Consistency Properties*. Volume 1. San Antonio, TX: USAF School of Aviation Medicine; 1985.

57. Dhanabal S, Chandramathi S. A review of various k-nearest neighbor query processing techniques. *Int J Comput Appl.* 2011;31(7):14-22.

58. Zhang ZH. Introduction to machine learning: k-nearest neighbors. *Ann Transl Med.* 2016;4(11):218.

doi: 10.21037/atm.2016.03.37

59. Danielsson P-E. Euclidean distance mapping. *Comput Graph Image Process.* 1980;14(3):227-248.

doi: 10.1016/0146-664x(80)90054-4

60. Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 1967;27(2 pt 1):209-220.

61. Cheng YX, Ding S, Shao Z, *et al.* Persistence of dissolved organic matter in sediments influenced by environmental factors: Implication for nutrition and carbon cycle. *J Environ Manage.* 2024;363:121387.

doi: 10.1016/j.jenvman.2024.121387

62. Zhang HY, Li N, Song JX, *et al.* Influence of land use on spatial distribution of primary productivity in aquatic environment in the Weihe River Basin, China. *J Arid Land.* 2025;17(3):304-323.

doi: 10.1007/s40333-025-0095-6

63. Moneera A, Maram A, Azizah A, *et al.* Click through Rate Effectiveness Prediction on Mobile Ads Using Extreme Gradient Boosting. *Comput Mater Contin.* 2021;66(2):1681-1696.

doi: 10.32604/cmc.2020.013466

64. Eun Choi J, Won Shin J, Wan Shin D. Vector SHAP Values for Machine Learning Time Series Forecasting. *J Forecast.* 2025;44(2):635-645.

doi: 10.1002/for.3220

65. Sharma A, Verbeke WJMI. Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081). *Front Big Data.* 2020;3:15.

doi: 10.3389/fdata.2020.00015

66. Zhang WG, Wu CZ, Zhong HY, Li YQ, Wang L. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geosci Front.* 2021;12(1):469-477.

doi: 10.1016/j.gsf.2020.03.007

67. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. *arXiv.* Preprint posted online 2017.

doi: 10.48550/arXiv.1705.07874

68. Shah FU, Khan AU, Khan AW, Ullah B, Khan MR, Javed I. Comparative analysis of ensemble learning algorithms in water quality prediction. *J Hydroinform.* 2024;26(12):3041-3059.

doi: 10.2166/hydro.2024.071

69. Das BK, Paul S, Mandal B, *et al.* Integrating machine learning models for optimizing ecosystem health assessments through prediction of nitrate–N concentrations in the lower stretch of Ganga River, India. *Environ Sci Pollut Res.* 2025;32(8):4670-4689.

doi: 10.1007/s11356-025-35999-z

70. Fei C, Cao X, Zang D, *et al.* Machine learning techniques for real-time UV-Vis spectral analysis to monitor dissolved nutrients in surface water. In: Kitayama K-i, Jalali B, eds. *Proceedings Volume 11703: AI and Optical Data Sciences II*. Bellingham, WA: SPIE; 2021:46.

doi: 10.1117/12.2577050

71. Niazkar M, Piraei R, Goodarzi MR, Abedi MJ. Comparative Assessment of Machine Learning Models for Groundwater Quality Prediction Using Various Parameters. *Environ Process.* 2025;12(1):10.

doi: 10.1007/s40710-025-00751-9

72. Fadiran AO, Dlamini SC, Mavuso. A comparative study of the phosphate levels in some surface and ground water bodies of Swaziland. *Bull Chem Soc Ethiop.* 2008;22(2):197-206.

doi: 10.4314/bcse.v22i2.61286

73. Liu X, Dur G, Ban S, Sakai Y, Ohmae S, Morita T. Quasi-decadal periodicities in growth and production of the copepod Eodiaptomus japonicus in Lake Biwa, Japan, related to the Arctic Oscillation. *Limnol Oceanogr.* 2021;66(10):3783-3795.

doi: 10.1002/lno.11918

74. Dur G, Liu X, Sakai Y, Hsieh C-h, Ban S, Souissi S. Disrupted seasonal cycle of the warm-adapted and main zooplankter of Lake Biwa, Japan. *J Great Lakes Res.* 2022;48(5):1206-1218.

doi: 10.1016/j.jglr.2022.06.001

75. Rammohan B, Partheeban P, Ranganathan R, Balaraman S. Groundwater Quality Prediction and Analysis Using Machine Learning Models and Geospatial Technology. *Sustainability.* 2024;16(22):9848.

doi: 10.3390/su16229848

76. Zou M, Jiang W-G, Qin Q-H, Liu Y-C, Li M-L. Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V parts manufactured by selective laser melting. *Materials.* 2022;15(15):5298.

doi: 10.3390/ma15155298

77. Zhu W, Shen S, Zhang Z. Improved multiclassification of schizophrenia based on xgboost and information fusion for small datasets. *Comput Math Methods Med.* 2022;2022(1):1581958.

doi: 10.1155/2022/1581958

78. Luan SL, Pan HX, Shen RQ, *et al.* High Resolution Water Quality Dataset of Chinese Lakes and Reservoirs from 2000 to 2023. *Sci Data.* 2025;12(1):572.

doi: 10.1038/s41597-025-04915-y

79. Sani Z, Katshiatshia HM, Tshimanga RM, Basamba TA, Odume ON. Relationship between land use and water quality in a tropical urban catchment of the Congo Basin: A case study of N'Djili River catchment. *Water Pract Technol.* 2024;19(4):1272-1289.

doi: 10.2166/wpt.2024.080

80. Boström B, Andersen JM, Fleischer S, Jansson M. Exchange of phosphorus across the sediment-water interface. *Hydrobiologia.* 1988;170(1):229-244.

doi: 10.1007/BF00024907

81. Yu YS, Ding P, Bian HY, Wei JS, Zhang H. Water quality parameters inversion based on multispectral remote sensing. *J Water Process Eng.* 2025;73:107707.

doi: 10.1016/j.jwpe.2025.107707

82. Bradter U, Altringham JD, Kunin WE, Thom TJ, O'Connell J, Benton TG. Variable ranking and selection with random forest for unbalanced data. *Environ Data Sci.* 2022;1:e30.

doi: 10.1017/eds.2022.34

83. Huang YP, Wang C, Wang YH, *et al.* Application of machine learning models in groundwater quality assessment and prediction: progress and challenges. *Front Environ Sci Eng.* 2024;18(3):29.

doi: 10.1007/s11783-024-1789-2

84. Tepetidis N, Koutsoyiannis D, Iliopoulou T, Dimitriadis P. Investigating the performance of the informer model for streamflow forecasting. *Water.* 2024;16(20):2882.

doi: 10.3390/w16202882

85. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. In: *Advances in Neural Information Processing Systems 30.* 2017;30.

86. Akinnigbagbe AE, Popoola SO, Oyatola OO, Oghenede EK, Nubi OA. Spatiotemporal patterns of water quality, nutrient dynamics and chlorophyll a concentrations in Five Cowries Creek from 2022 to 2024. *Discover Water.* 2025;5(1):63.

doi: 10.1007/s43832-025-00257-3

87. Tavakoli M, Bayati S. Seasonal assessment of surface and groundwater quality, heavy metals, and water quality indices (case study: Dehloran watershed, Ilam, Iran). *Appl Water Sci.* 2026;16(3):86.

doi: 10.1007/s13201-026-02781-6

88. Kumar V, Shekhar S, Arora NK. Integrated geostatistical and multiple linear regression modeling for assessment of groundwater quality parameters in the Burhi Gandak Basin. *Int J River Basin Manage.* 2026:1-29.

doi: 10.1080/15715124.2026.2626838

89. Li Y, Zhang Q, Cai Y, *et al.* Hydrodynamic investigation of surface hydrological connectivity and its effects on the water quality of seasonal lakes: Insights from a complex floodplain setting (Poyang Lake, China). *Sci Total Environ.* 2019;660:245-259.

doi: 10.1016/j.scitotenv.2019.01.015

90. Yamazaki Y, Muneoka T, Wakou S, Kimura M, Tsuji O. Evaluation of the Ion Components for the Estimation of Total Nitrogen Concentration in River Water based on Electrical Conductivity. *Int J Environ Rural Dev.* 2014;5(1):160-164.

doi: 10.32115/ijerd.5.1_160

91. Liang Y, Xiao HY, Liu XZ, *et al.* Variation in sources of inorganic nitrogen under different hydrological conditions in a floodplain lake: a case study of Bang Lake (Poyang Lake, Jiangxi Province, China). *Inland Waters.* 2018;8(2):176-185.

doi: 10.1080/20442041.2018.1457854

92. Wang JL, Fu ZS, Qiao HX, Liu FX. Assessment of eutrophication and water quality in the estuarine area of Lake Wuli, Lake Taihu, China. *Sci Total Environ.* 2019;650(pt 1):1392-1402.

doi: 10.1016/j.scitotenv.2018.09.137

93. De Matos Valerio A, Montanari JL, Kampel M, Ward ND, Richey JE, Cunha AC. Chlorophyll-a dynamics in the lower Amazon River: insights from in situ and hyperspectral remote sensing using OCI-PACE. *Int J Remote Sens.* 2026;47(1):124-149.

doi: 10.1080/01431161.2025.2596811

94. Çelekli A, Zariç ÖE. Ecological Assessments of Seven Lakes in Antalya River Basin: Application of Phytoplankton Indices and Multimetric Approach. *Ecohydrology.* 2026;19(1):e70174.

doi: 10.1002/eco.70174

95. Sun LF, Qiao YC, Song LQ, Peng YJ, Hu BQ. Responses of N2O production and associated functional genes to increasing temperature and moisture in surface and subsurface soils of a temperate forest. *Pedobiologia.* 2025;108:151020.

doi: 10.1016/j.pedobi.2025.151020

96. Luo ZX, Qiu ZZ, Wei QS, Du Laing G, Zhao YL, Yan CZ. Dynamics of ammonia-oxidizing archaea and bacteria in relation to nitrification along simulated dissolved oxygen gradient in sediment–water interface of the Jiulong river estuarine wetland, China. *Environ Earth Sci.* 2014;72(7):2225-2237.

doi: 10.1007/s12665-014-3128-6

97. Duan ZZ, Gao W, Cheng GW, Zhang Y, Chang XX. Warming surface and Lake heatwaves as key drivers to harmful algal Blooms: A case study of Lake Dianchi, China. *J Hydrol.* 2024;632:130971.

doi: 10.1016/j.jhydrol.2024.130971

98. Klose K, Cooper SD, Leydecker AD, Kreitler J. Relationships among catchment land use and concentrations of nutrients, algae, and dissolved oxygen in a southern California river. *Freshw Sci.* 2012;31(3):908-927.

doi: 10.1899/11-155.1

99. Romanowicz R, Petersen W. Statistical Modeling of Algae Concentrations in the Elbe River in the Years 1985—2001 Using Observations of Daily Dissolved Oxygen, Temperature, and pH. *Acta Hydrochim Hydrobiol.* 2003;31(4-5):319-333.

doi: 10.1002/aheh.200300499

100. Li B, Yang GS, Wan RR, Xu LG. Chlorophyll a variations and responses to environmental stressors along hydrological connectivity gradients: Insights from a large floodplain lake. *Environ Pollut.* 2022;307:119566.

doi: 10.1016/j.envpol.2022.119566

101. Legendre P, Fortin MJ. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Mol Ecol Resour.* 2010;10(5):831-844.

doi: 10.1111/j.1755-0998.2010.02866.x

102. Yi CY, Li JF, Zhang CR, Pan F, Zhou CF. In Situ Monitoring of a Eutrophicated Pond Revealed Complex Dynamics of Nitrogen and Phosphorus Triggered by Decomposition of Floating-Leaved Macrophytes. *Water.* 2021;13(13):1751.

doi: 10.3390/w13131751

103. House W. Factors influencing the measurement of equilibrium phosphate concentrations in river sediments. *Water Res.* 2000;34(4):1187-1200.

doi: 10.1016/s0043-1354(99)00249-3

104. Koski-Vahala J, Hartikainen H, Tallberg P. Phosphorus mobilization from various sediment pools in response to increased pH and silicate concentration. *J Environ Qual.* 2001;30(2):546-552.

doi: 10.2134/jeq2001.302546x

105. Liikanen A, Murtoniemi T, Tanskanen H, Väisänen T, Martikainen PJ. Effects of temperature and oxygenavailability on greenhouse gas and nutrient dynamics in sediment of a eutrophic mid-boreal lake. *Biogeochemistry.* 2002;59(3):269-286.

doi: 10.1023/a:1016015526712

106. Kankaala P. The relative importance of algae and bacteria as food for Daphnia longispina (Cladocera) in a polyhumic lake. *Freshw Biol.* 1988;19(3):285-296.

doi: 10.1111/j.1365-2427.1988.tb00351.x