

ORIGINAL RESEARCH ARTICLE

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

Supplementary File

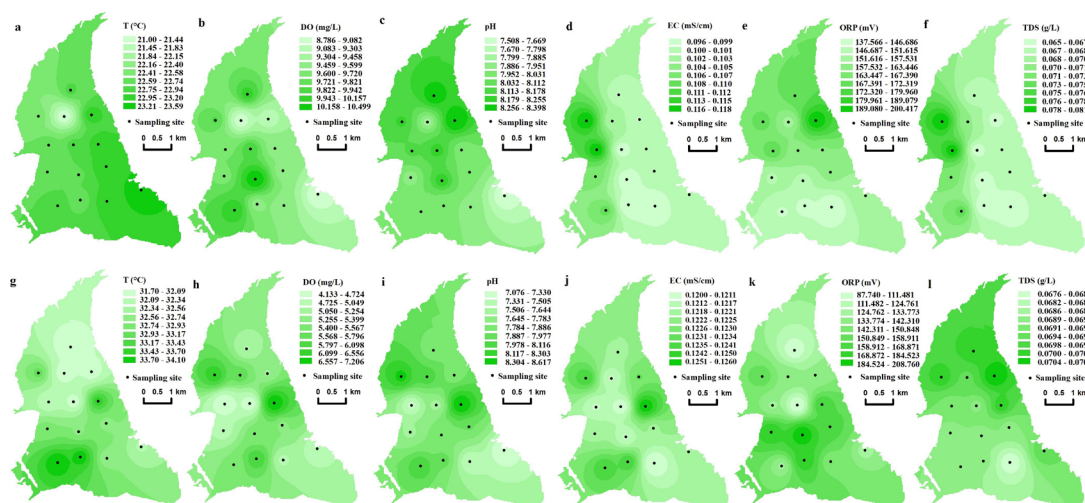


Figure S1. Spatiotemporal distribution of environmental parameters in the coastal area (Banghu Lake) of Poyang Lake, China. The (a–f) upper and (g–l) lower rows represent data collected on May 10 and August 10, 2024, respectively. (a, g) Water temperature (T). (b, h) Dissolved oxygen (DO). (c, i) Water pH. (d, j) Electrical conductivity (EC). (e, k) Oxidation–reduction potential (ORP). (f, l) Total dissolved solids (TDS). Black circles represent sampling sites. Site names correspond to those shown in Figure 1.

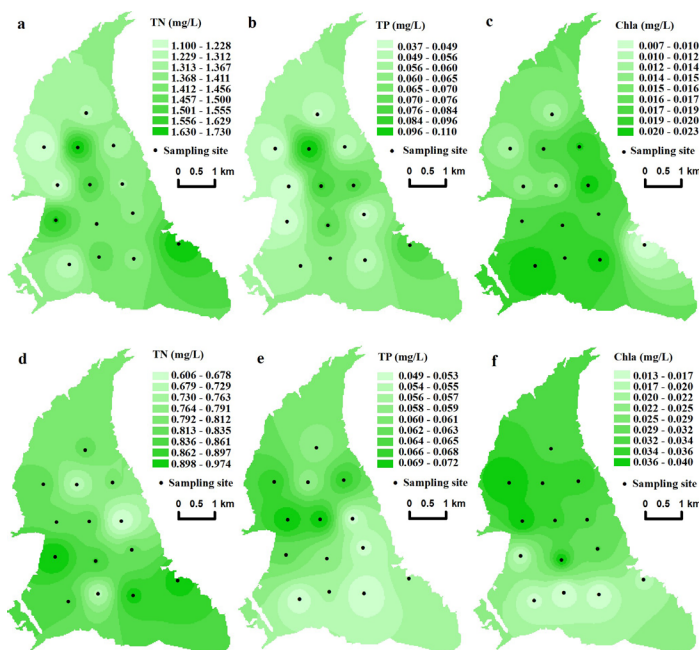


Figure S2. Spatiotemporal distribution of water quality indicators in the coastal area (Banghu Lake) of Poyang Lake, China. The (a–c) upper and (d–f) lower rows represent data collected on May 10 and August 10, 2024, respectively. (a, d) Total nitrogen (TN). (b, e) Total phosphorus (TP). (c, f) Chlorophyll-a (Chla). Black circles represent sampling sites. Site names correspond to those shown in Figure 1.

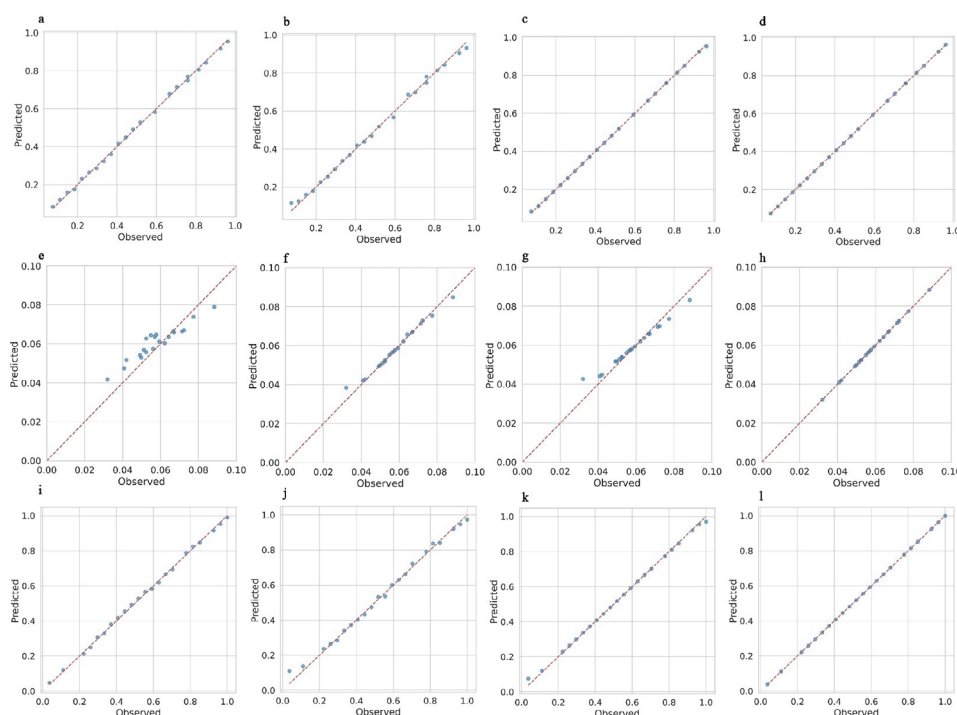


Figure S3. Scatter plots of observed versus predicted values of water quality indicators (total nitrogen [TN], total phosphorus [TP], chlorophyll-a [Chla]) using four machine learning models (support vector regression [SVR], random forest [RF], extreme gradient boosting [XGBoost], k-nearest neighbors [KNN]) for training. TN data from (a) SVR, (b) RF, (c) XGBoost, and (d) KNN. TP data from (e) SVR, (f) RF, (g) XGBoost, and (h) KNN. Chla data from (i) SVR, (j) RF, (k) XGBoost, and (l) KNN. A linear regression line (dashed red) was included in each model.

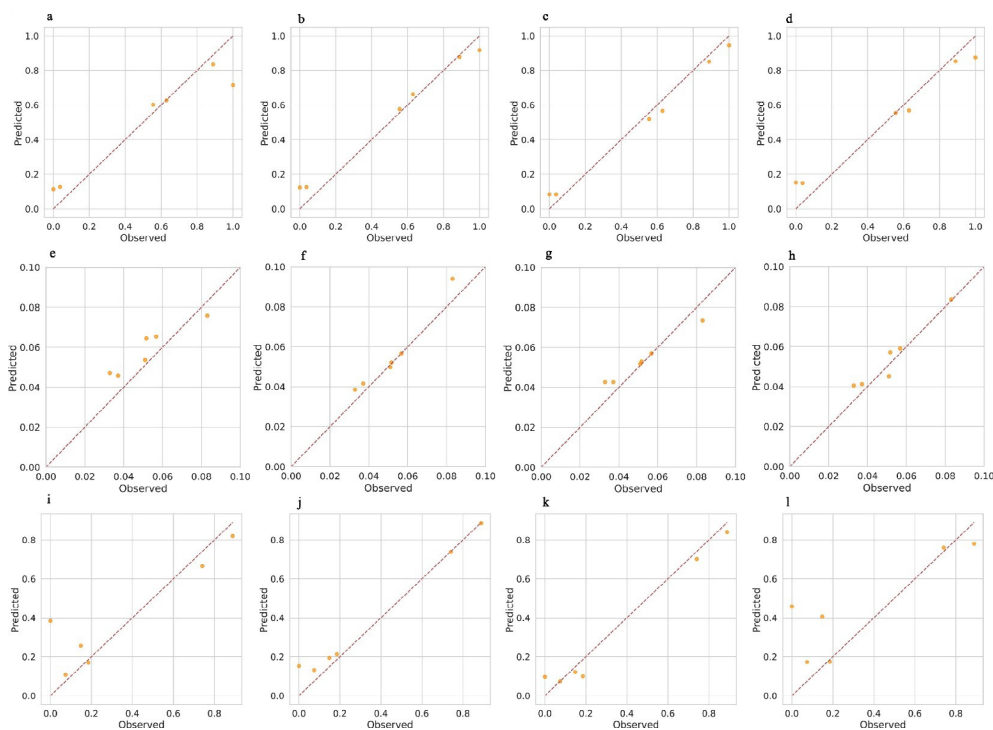


Figure S4. Scatter plots of observed versus predicted values of water quality indicators (total nitrogen [TN], total phosphorus [TP], chlorophyll-a [Chla]) using four machine learning models (support vector regression [SVR], random forest [RF], extreme gradient boosting [XGBoost], k-nearest neighbors [KNN]) for testing. TN data from (a) SVR, (b) RF, (c) XGBoost, and (d) KNN. TP data from (e) SVR, (f) RF, (g) XGBoost, and (h) KNN. Chla data from (i) SVR, (j) RF, (k) XGBoost, and (l) KNN. A linear regression line (dashed red) was included for each model.

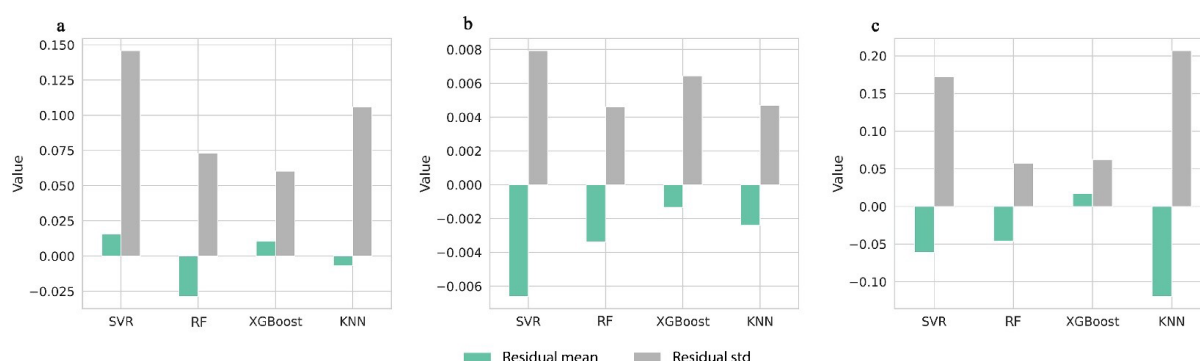


Figure S5. Residual means and residual standard deviations (Std) of water quality indicators: (a) total nitrogen (TN), (b) total phosphorus (TP), (c) chlorophyll-a (Chla), obtained from four machine learning models (support vector regression [SVR], random forest [RF], extreme gradient boosting [XGBoost], k-nearest neighbors [KNN]) during model testing.

Table S1. Means, standard deviations (SD), skewness, and kurtosis of the environmental parameters used in this study

Parameters	Mean	SD	Skewness	Kurtosis
Total nitrogen (TN, mg/L)	1.119	0.337	0.319	-1.268
Total phosphorus (TP, mg/L)	0.059	0.017	1.001	1.847
Chlorophyll-a (Chla, mg/L)	0.023	0.010	0.504	-1.111
Temperature (T, °C)	27.711	5.082	0.015	-2.064
Dissolved oxygen (DO, mg/L)	7.490	2.256	-0.107	-1.844
pH	7.923	0.352	-0.492	0.410
Electrical conductivity (EC, mS/cm)	0.112	0.011	-0.363	-1.783
Oxidation–reduction potential (ORP, mV)	153.501	25.866	-0.381	1.246
Total dissolved solids (TDS, g/L)	0.069	0.004	1.915	4.343

Table S2. Hyperparameter search ranges and their optimal values for the different machine learning models (MLMs) across the three water quality indicators (WQIs)

WQI	MLM	Search ranges	Optimal hyperparameter values
TN	SVR	kernel: ['rbf']	
		C: [0.1, 1, 5, 10]	C = 1
		epsilon: [0.01, 0.05, 0.1]	epsilon = 0.01
	RF	gamma: ['scale', 'auto']	gamma = 'auto'
		n_estimators: [100, 200, 500]	n_estimators = 500
		max_depth: [None, 4, 6, 8]	max_depth = 4
	XGBoost	min_samples_split: [2, 4, 6]	min_samples_split = 2
		n_estimators: [100, 200, 500]	n_estimators = 100
		learning_rate: [0.01, 0.05, 0.1]	learning_rate = 0.1
	KNN	max_depth: [3, 5, 7]	max_depth = 5
		reg_lambda: [1, 2, 5]	reg_lambda = 5
		n_neighbors: [3, 5, 10]	n_neighbors = 3
		weights: ['uniform', 'distance']	weights = 'distance'

(cont'd...)

Table S2. (Continued)

WQI	MLM	Search ranges	Optimal hyperparameter values
TP	SVR	kernel: ['rbf']	
		C: [0.1, 1, 5, 10]	C = 0.1
		epsilon: [0.01, 0.05, 0.1]	epsilon = 0.01
		gamma: ['scale', 'auto']	gamma = 'auto'
	RF	n_estimators: [100, 200, 500]	n_estimators = 100
		max_depth: [None, 4, 6, 8]	max_depth = None
		min_samples_split: [2, 4, 6]	min_samples_split = 2
		n_estimators: [100, 200, 500]	n_estimators = 200
	XGBoost	learning_rate: [0.01, 0.05, 0.1]	learning_rate = 0.01
		max_depth: [3, 5, 7]	max_depth = 5
		reg_lambda: [1, 2, 5]	reg_lambda = 1
		n_neighbors: [3, 5, 10]	n_neighbors = 3
	KNN	weights: ['uniform', 'distance']	weights = 'distance'
		kernel: ['rbf']	
Chla	SVR	C: [0.1, 1, 5, 10]	C = 5
		epsilon: [0.01, 0.05, 0.1]	epsilon = 0.01
		gamma: ['scale', 'auto']	gamma = 'auto'
		n_estimators: [100, 200, 500]	n_estimators = 500
	RF	max_depth: [None, 4, 6, 8]	max_depth = None
		min_samples_split: [2, 4, 6]	min_samples_split = 2
		n_estimators: [100, 200, 500]	n_estimators = 100
		learning_rate: [0.01, 0.05, 0.1]	learning_rate = 0.05
	XGBoost	max_depth: [3, 5, 7]	max_depth = 3
		reg_lambda: [1, 2, 5]	reg_lambda = 1
		n_neighbors: [3, 5, 10]	n_neighbors = 3
		weights: ['uniform', 'distance']	weights = 'distance'
	KNN		

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.

Table S3. Parameter metrics (RESE, MAE, R^2) for four machine learning models (MLMs) based on training data for water quality indicators (WQIs)

WQIs	MLMs	RMSE	MAE	R^2
TN	SVR	0.00956	0.00945	0.99876
	RF	0.01592	0.01235	0.99656
	XGBoost	0.01118	0.00426	0.99830
	KNN	0	0	1
TP	SVR	0.00624	0.00530	0.85556
	RF	0.00263	0.00128	0.97441
	XGBoost	0.00500	0.00289	0.90724
	KNN	0	0	1
Chla	SVR	0.00941	0.00905	0.99878
	RF	0.01982	0.01427	0.99460
	XGBoost	0.01088	0.00577	0.99837
	KNN	0	0	1

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.