Supporting Materials

Utilizing machine learning models to grasp water quality dynamic changes in lake eutrophication through phytoplankton parameters

Yong Fang, Ruting Huang (✉), Yeyin Zhang, Jun Zhang, Wenni Xi, Xianyang Shi (✉)

Anhui Province Key Laboratory of Wetland Ecosystem Protection and Restoration, School of Resources and Environmental Engineering, Anhui University, Hefei 230601, China

✉ Corresponding authors
E-mail: rthuang@ahu.edu.cn (R. Huang); shixi381@163.com (X. Shi)

Text S1 Machine Learning Models (DT, RF, XGBoost, CatBoost, and LightGBM)

S1.1 DT

DT is a basic classification and regression method. The DT model is similar to the tree structure. In the process of prediction, it could be regarded as classifying samples based on features that correspond to the "if-then" rules. The main advantage of this approach is that it has readability and fast classification speed. When learning, the DT model is established by using the training data based on the principle of minimizing the loss function (Berk, 2016).

S1.2 RF

RF is an improvement of the bagging algorithm, still choosing DT as its base learner. The largest difference between the RF and bagging algorithms is the construction of the DT. The general DT algorithm selects an optimal feature from all of the sample features on the tree node to determine the split of the DT subtree. Meanwhile, RF randomly selects a part of the sample features from all of the features and then chooses one of the best features by mean square error (MSE) to determine the division of the DT subtree, further strengthening the model's generalization ability (Belgiu and Drăguţ, 2016).

S1.3 XGBoost

XGBoost is a model adopting a boosting strategy that differs from the bagging strategy in that it constructs DT in series rather than in parallel. This means that the new DT fits the error between the predicted value of all previous DTs and the true value during the model training process. One advantage of XGBoost is its high efficiency, that is, massively parallel boosting DT, showing a higher calculation speed and requiring much less computational effort than RF (Sheridan et al., 2016).

S1.4 CatBoost

CatBoost is a gradient-boosting decision tree (GBDT) framework, constructed upon the foundation of symmetric DT as the base learner, and the main point addressed is the efficient and rational dealing with categorical features. CatBoost exhibits a lower number of parameters, the capacity to accommodate categorical variables, along with a high degree of accuracy. Moreover, CatBoost addresses two key issues: gradient bias and prediction shift. By addressing these issues, CatBoost reduces the likelihood of model overfitting (Prokhorenkova et al., 2018).

S1.5 LightGBM

LightGBM represents a framework for the implementation of the GBDT algorithm. The underlying concept may be described as follows: the continuous features, which are of a floating-point nature, are first discretized and then k discrete values are constructed. A histogram of width k is subsequently formed, and the training data are traversed over to compute the cumulative statistics for each discrete value. Based on the distinct values observed in the histogram, it is only essential to traverse the data set to identify the optimized clustering points in the process of feature extraction. In this process, a leaf-wise method with depth restriction is employed. The framework is designed to enable efficient parallel training and offers a number of benefits in this regard. These include a reduction in training time, greater accuracy, and the ability to process data sets distributed across multiple processors (Ke et al., 2017).

S1.6 Comparison of model selection rationale and limitations

In this study, we selected five models, DT, RF, XGBoost, CatBoost, and LightGBM, for lake eutrophication prediction analysis. Our main reasons for selecting these models include their advantages in dealing with nonlinear problems and high-dimensional data, as well as their high accuracy and interpretability, which are widely recognized within the field of environmental science (Berk, 2016). However, there are limitations of these models, such as the possibility of overfitting in DT, the high computational cost of RF when processing data sets, and the complexity and sensitivity to hyperparameters of the parameter tuning process of XGBoost, CatBoost, and LightGBM, despite their fast-processing speed and good learning results. These limitations need to be considered in practical applications and appropriate models and tuning strategies should be selected according to specific situations (Sheridan et al., 2016; Ke et al., 2017).

Text S2 Five references for evaluating the performance of machine learning models (DT, RF, CatBoost, XGBoost, and LightGBM)

● MSE (Mean Square Error): The expected value of the squared differences between the predicted and actual values is indicative of the accuracy of the model. A smaller value is indicative of a higher level of accuracy.

● RMSE (Root Mean Square Error): The square root of MSE, where a smaller value signifies greater model accuracy.

● MAE (Mean Absolute Error): The mean of the absolute errors provides insight into the actual scenario of prediction errors. A lower value indicates greater accuracy in the model.

● MAPE (Mean Absolute Percentage Error): A variant of MAE, represented as a percentage. Lower values denote higher model accuracy.

● R^2: This measure compares the predicted values with the scenario of using only the mean. A result closer to 1 indicates greater accuracy in the model.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \quad , \quad \text{(S1)}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \quad , \quad \text{(S2)}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}^{N} (\hat{y}_i - y_i)^2} \quad , \quad \text{(S3)}$$

$$MAPE = \frac{1}{N}\sum_{i}^{N} \frac{|\hat{y}_i - y_i|}{y_i} \quad , \quad \text{(S4)}$$

$$R^2 = 1 - \frac{\sum_i^N (\hat{y}_i - y_i)^2}{\sum_i^N (\bar{y} - y_i)^2} \quad , \quad \text{(S5)}$$

where y_i is the real water quality index, $\hat{y}_i$ is the predicted water quality index; $\bar{y}$ is the average value of the water quality index; N is the number of water quality monitoring sites. Values close to 0 for MAE, MSE, RMSE, and MAPE, or an R^2 value approaching 1, indicate exceptional predictive performance of the model. Furthermore, a negative R^2 value suggests that the fitted function's predictive error exceeds that of the function itself.

Table S1 Detailed information of all monitoring sites in this study.

Number	Sites	Latitude	Longitude
1	Siqian river	116.2268953	30.57181294
2	Qingshi river	116.1366955	30.56204589
3	Shuyan river	116.1331933	30.46564993
4	Dianqian river	116.0726224	30.56111623
5	Anle river	116.0480788	30.54423458
6	Wan river	116.2478646	30.46421536
7	L1	116.2810135	30.5424519
8	L2	116.2676239	30.52500475
9	L3	116.2502861	30.49749697
10	L4	116.2151814	30.54821763
11	L5	116.2088299	30.52692705
12	L6	116.1984444	30.50903344
13	L7	116.1368179	30.55427868
14	L8	116.1349297	30.5434868
15	L9	116.1283207	30.53158476
16	L10	116.0920143	30.54880897
17	L11	116.1126137	30.53653796
18	L12	116.1476326	30.5056318
19	L13	116.1605072	30.49594387
20	L14	116.1911488	30.48950938
21	L15	116.2025642	30.49897608
22	L16	116.2229919	30.50296956
23	L17	116.209259	30.47693499
24	L18	116.2262535	30.48337032
25	L19	116.2317467	30.4906188
26	L20	116.2397289	30.48115129
27	L21	116.2370682	30.47423498
28	L22	116.2430763	30.47700896
29	L23	116.241703	30.4706102
30	L24	116.2403297	30.46569063
31	L25	116.2453079	30.46635645
32	L26	116.2384415	30.46121472
33	L27	116.2484944	30.46674485
34	L28	116.2502754	30.46508029

Table S2 Parameters used to build machine learning models.

Category of parameters	Parameters
Physical and chemical parameters	T (℃), pH, v (m/s), SD (m), EC (μS/cm), ORP (mV), TN (mg/L), TP (mg/L), TN/TP, NH_4-N (mg/L), AP (mg/L), COD_{Mn} (mg/L), DO (mg/L), and MC (ng/L)
Atmosphere/Scene parameters	High temperature (HT) (℃), Low temperature (LT) (℃), Air visibility (AV) (m), Wind speed (WS) (m/s), and Total rainfall (TR) (mm)
Phytoplankton parameters	Shannon index (H'), Margalef index (H), Pielou index (J), Simpson index (D), Invsimpson index (1/D), Top three dominant species (1st, 2nd, 3rd), and Algae cell density (ACD) (cells/L)
Total parameters	**28**

Table S3 Parameters TN/TP, TP, and TN thresholds division standard.

Parameter	Thresholds	Nutrient limiting element
TN/TP	≥ 22.6	P limitation
	≤ 9	N limitation
	9 < &< 22.6	N-P co-limitation
TP	> 0.029 mg/L	Eutrophication
TN	> 0.58 mg/L	

Table S4 Parameters Margalef index (H) threshold division standard.

Parameters	Cleanness	Oligotrophic	β-mesotrophic	α-mesotrophic	eutrophic
H	>5	4 ≤ & < 5	3 ≤ & < 4	≤ 3	/
H′	/	> 3	2 ≤ & < 3	1 ≤ & < 2	≤ 1

Table S5 The optimal hyperparameter values for Decision Tree (DT).

Class-ifier	Hyperparameters	Optimal model hyperparameters					
		H'	**J**	**D**	**H**	**TN/TP**	**MC**
DT	Average computation time (s)	0.028	0.018	0.025	0.432	0.236	0.034
	Data Split Ratio	0.7	0.7	0.7	0.7	0.7	0.7
	Data Shuffle	Yes	Yes	Yes	Yes	Yes	Yes
	Cross-validation Folds	5	5	5	5	5	5
	Node Splitting Criterion	friedman_mse	friedman_mse	friedman_mse	friedman_mse	friedman_mse	friedman_mse
	Feature Splitting Criterion	best	best	best	best	best	best
	Max Features Considered for Split	None	None	None	None	None	None
	Min Samples Required for Internal Node Split	2	2	2	2	2	2
	Min Samples Required for Leaf Node	1	1	1	1	1	1
	Min Sample Weight in Leaf Node	0	0	0	0	0	0
	Max Depth of Tree	10	10	10	10	10	10
	Max Number of Leaf Nodes	50	50	50	50	50	50
	Impurity Threshold for Splitting Nodes	0	0	0	0	0	0

Table S6 The optimal hyperparameter values for Decision Tree (RF).

Classifier	Optimal model hyperparameters						
	Hyperparameters	**H'**	**J**	**D**	**H**	**TN/TP**	**MC**
RF	Average computation time (s)	0.836	0.796	0.932	0.680	1.135	0.859
	Data Split Ratio	0.7	0.7	0.7	0.7	0.7	0.7
	Data Shuffle	Yes	Yes	Yes	Yes	Yes	Yes
	Cross-validation Folds	5	5	5	5	5	5
	Node Splitting Criterion	mse	mse	mse	mse	mse	mse
	Max Features Considered for Split	None	None	None	None	None	None
	Min Samples Required for Internal Node Split	2	2	2	2	2	2
	Min Samples Required for Leaf Node	1	1	1	1	1	1
	Min Sample Weight in Leaf Node	0	0	0	0	0	0
	Max Depth of Tree	20	20	20	20	20	20
	Max Number of Leaf Nodes	50	50	50	50	50	50
	Impurity Threshold for Splitting Nodes	0	0	0	0	0	0
	Number of Trees in the Forest	100	100	100	100	100	100
	Bootstrap Sampling	true	true	true	true	true	true
	Out-of-Bag (OOB) Data Testing	false	false	false	false	false	false

Table S7 The optimal hyperparameter values for Decision Tree (CatBoost).

Classifier	Optimal model hyperparameters						
	Hyperparameters	**H'**	**J**	**D**	**H**	**TN/TP**	**MC**
CatBoost	Average computation time (s)	1.043	0.972	1.135	1.362	1.683	1.034
	Data Split Ratio	0.7	0.7	0.7	0.7	0.7	0.7
	Data Shuffle	Yes	Yes	Yes	Yes	Yes	Yes
	Cross-validation	No	No	No	No	No	No
	Iterations	100	100	100	100	100	100
	Learning Rate	0.1	0.1	0.1	0.1	0.1	0.1
	L2 Regularization Term	1	1	1	1	1	1.02
	Max Depth of Tree	20	20	20	20	20	20
	Overfitting Detection Threshold	0	0	0	0	0	0
	Iterations After Optimization	20	20	20	20	20	20

Table S8 The optimal hyperparameter values for Decision Tree (XGBoost).

Classifier	Optimal model hyperparameters						
	Hyperparameters	**H'**	**J**	**D**	**H**	**TN/TP**	**MC**
XGBoost	Average computation time (s)	1.236	1.573	1.379	1.936	2.031	1.962
	Data Split Ratio	0.7	0.7	0.7	0.7	0.7	0.7
	Data Shuffle	Yes	Yes	Yes	Yes	Yes	Yes
	Cross-validation Folds	5	5	5	5	5	5
	Base Learner	gbtree	gbtree	gbtree	gbtree	gbtree	gbtree
	Number of Base Learners	100	100	100	100	100	100
	Learning Rate	0.1	0.1	0.1	0.1	0.1	0.1
	L1 Regularization Term	0.1	0	0	0	0.01	0
	L2 Regularization Term	1	1.1	1	1	1	1
	Subsample Rate of Training Data	1	1	1	1	1	1
	Subsample Rate of Columns for Trees	1	1	1	1	1	1
	Subsample Rate of Columns for Splitting Nodes	1	1	1	1	1	1
	Min Sample Weight in Leaf Node	0	0	0	0	0	0
	Max Depth of Tree	20	20	20	20	20	20

Table S9 The optimal hyperparameter values for Decision Tree (LightGBM).

Classifier	Optimal model hyperparameters						
	Hyperparameters	**H'**	**J**	**D**	**H**	**TN/TP**	**MC**
LightGBM	Average computation time (s)	0.86	1.023	1.352	1.094	1.962	1.263
	Data Split Ratio	0.7	0.7	0.7	0.7	0.7	0.7
	Data Shuffle	Yes	Yes	Yes	Yes	Yes	Yes
	Cross-validation Folds	5	5	5	5	5	5
	Base Learner	gbdt	gbdt	gbdt	gbdt	gbdt	gbdt
	Number of Base Learners	100	100	100	100	100	100
	Learning Rate	0.1	0.1	0.1	0.1	0.1	0.1
	L1 Regularization Term	0.01	0	0.01	0	0	0
	L2 Regularization Term	1	1	1.1	1	1	1
	Subsample Rate of Training Data	1	1	1	1	1	1
	Subsample Rate of Columns for Trees	1	1	1	1	1	1
	Splitting Node Threshold	0	0	0	0	0	0
	Min Sample Weight in Leaf Node	0	0	0	0	0	0
	Max Depth of Tree	20	20	20	20	20	20
	Min Samples in Leaf Node	10	10	10	10	10	10

Table S10 Statistics of various parameters in the lake area (average, maximum, and minimum).

Parameters/Groups	LS-1			LS-2			LS-3			LS		
\	Average	Mix	Min	Average	Mix	Min	Average	Mix	Min	Average	Mix	Min
T	27.0	34.2	15.9	26.5	33.7	15.7	26.6	33.5	15.0	26.7	34.2	15.0
pH	8.0	9.8	4.7	8.0	9.8	4.9	7.8	9.8	4.7	7.9	9.8	4.7
v (m/s)	0.2	0.5	0.0	0.2	0.5	0.0	0.3	0.6	0.0	0.3	0.6	0.0
SD (m)	2.2	4.0	1.0	1.0	2.7	0.1	2.5	4.5	1.0	1.9	4.5	0.1
EC (μS/cm)	596.3	764.4	91.8	603.8	765	104.1	813.0	2550	89.6	807.2	2550	89.6
ORP (mV)	-42.9	91.2	-95.9	-44.2	86.3	-99.9	-57.0	61.9	-95.9	-51.7	91.2	-99.9
TP (mg/L)	0.0	0.1	0.0	0.1	0.1	0.0	0.0	0.1	0.0	0.1	0.1	0.0
TN (mg/L)	0.7	1.6	0.1	1.1	3.3	0.1	0.6	1.6	0.0	0.9	3.3	0.0
TN/TP	69.1	453.3	1.3	52.2	294	1.9	59.1	391.8	0.2	68.2	453.3	0.2
NH_4-N (mg/L)	0.1	0.3	0.0	0.1	0.3	0.0	0.0	0.2	0.0	0.1	0.3	0.0
AP (mg/L)	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0
COD_{Mn} (mg/L)	3.4	5.3	2.1	4.2	7.0	2.3	3.2	4.5	2.0	3.9	7.0	2.0
DO (mg/L)	9.1	10.7	5.3	10.1	14.2	5.0	8.8	10.5	5.3	9.7	14.2	5.0
MC（ng/L）	324.2	619.5	133	342.5	641	148.7	245.6	542.2	64.8	310.4	641.2	64.8
H'	1.4	2.3	0.6	1.7	2.3	0.9	1.4	2.2	0.5	1.5	2.3	0.5
H	0.8	1.4	0.2	1.1	1.8	0.3	0.7	1.3	0.3	0.8	1.8	0.2
J	0.5	0.9	0.2	0.6	0.9	0.2	0.5	0.9	0.2	0.5	0.9	0.2
D	0.6	0.9	0.2	0.7	0.9	0.3	0.6	0.8	0.2	0.6	0.9	0.2
1/D	3.1	7.0	1.2	4.2	7.2	1.5	3.0	6.5	1.2	3.3	7.2	1.2
ACD (cells/L)	6.09 E+06	3.09 E+07	2.22 E+05	2.57 E+07	2.73 E+08	5.25 E+05	4.35 E+06	1.70 E+07	2.32 E+05	1.01 E+07	2.73 E+08	2.22 E+05

Table S11 Statistics of various parameters of rivers entering the lake (average value, maximum value, and minimum value).

Parameters/Groups	RS-2			RS-3			RS			Wan River		
\	Average	Mix	Min	Average	Mix	Min	Average	Mix	Min	Average	Mix	Min
T	27.2	32.8	19.5	27.3	34.2	17.3	27.3	33.1	19.1	26.0	32.6	15.5
pH	7.7	9.1	4.7	7.9	9.5	4.7	7.7	9.2	4.7	8.0	9.7	4.8
v (m/s)	0.3	0.4	0.2	0.2	0.4	0.0	0.3	0.4	0.1	0.2	0.4	0.1
SD (m)	0.7	0.9	0.5	1.0	1.1	0.9	0.7	0.9	0.6	2.1	2.5	1.2
EC (μS/cm)	603.7	764.0	131.6	598.5	764.0	110.3	602.6	764	127.3	754.5	762.7	738.0
ORP (mV)	-27.4	128.2	-95.8	-42.3	85.1	-95.4	-30.4	119.6	-95.7	-59.8	6.5	-92.1
TP (mg/L)	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.1	0.0	0.0	0.1	0.0
TN (mg/L)	1.2	1.5	0.8	0.9	1.5	0.2	1.1	1.4	0.8	0.6	1.6	0.1
TN/TP	47.0	145.2	12.2	33.6	97.4	1.8	44.4	136	10.4	45.2	157.6	3.1
NH_4-N (mg/L)	0.1	0.2	0.0	0.1	0.3	0.0	0.1	0.2	0.0	0.0	0.1	0.0
AP (mg/L)	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0
COD_{Mn} (mg/L)	3.5	3.9	3.1	3.5	3.9	2.9	3.5	3.8	3.1	3.0	3.9	2.6
DO (mg/L)	9.7	10.7	7.8	9.6	10.5	7.6	9.7	10.6	7.8	8.9	10.1	6.4
MC（ng/L）	243.2	405.7	123.6	265.2	466.2	82.6	247.6	418	115.4	293.5	459.8	153.5
H'	1.8	2.9	1.2	1.6	2.0	1.1	1.7	2.9	1.0	1.4	1.7	1.0
H	1.2	2.2	0.5	1.1	1.2	0.9	1.1	2.2	0.5	0.7	0.8	0.5
J	0.6	0.8	0.3	0.5	0.7	0.4	0.6	0.9	0.3	0.5	0.9	0.4
D	0.7	0.9	0.5	0.7	0.8	0.4	0.7	0.9	0.4	0.6	0.8	0.4
1/D	4.4	8.0	2.1	3.9	5.2	1.8	4.1	8.0	1.6	2.9	4.3	1.6
ACD (cells/L)	1.32 E+07	7.58 E+07	4.34 E+05	1.13 E+07	1.88 E+07	1.90 E+06	1.12 E+07	7.58 E+07	4.34 E+05	3.37 E+06	7.58 E+06	6.16 E+05

Table S12 The input parameters of different target parameters for predictions.

Target parameter	H'	J	D	H	TN/TP	MC
Input parameters	T	T	J	v	T	T
	pH	v	H'	SD	pH	pH
	SD	TP	SD	TP	EC	EC
	ORP	AP	T	TN	AP	ORP
	TN	MC	3rd	NH_4-N	MC	3rd
	H	H'	ORP	DO	TP	TN/TP
	J	H	H	H'	TN	DO
	D	D	DO	J	1/D	1st
	1/D	ACD	pH	ACD	\	AV
	1st	WS	TN	AV	\	\

Table S13 Test set performance of five models (DT, RF, CatBoost, XGBoost, and LightGBM) used to predict target parameters (TN/TP, H', J, MC, D, and H).

Parameters	Models	MAE	MAPE (%)	MSE	R^2	RMSE
TN/TP	DT	18.27	44.67	1556.30	0.73	39.45
	RF	11.50	26.64	689.43	0.86	26.26
	CatBoost	17.32	38.89	1544.73	0.76	39.30
	XGBoost	10.98	26.14	605.01	0.88	24.60
	LightGBM	24.38	569.17	2704.00	0.60	52.00
H'	DT	0.11	7.60	0.03	0.86	0.16
	RF	0.08	5.93	0.02	0.92	0.12
	CatBoost	0.10	6.93	0.03	0.85	0.16
	LightGBM	0.08	4.95	0.01	0.92	0.11
	XGBoost	0.08	5.75	0.01	0.91	0.12
J	DT	0.05	10.38	0.01	0.76	0.08
	RF	0.04	7.07	0.00	0.91	0.05
	CatBoost	0.06	10.93	0.01	0.79	0.08
	XGBoost	0.04	8.34	0.00	0.88	0.06
	LightGBM	0.04	8.01	0.00	0.89	0.06
MC	DT	76.33	33.27	14351.02	0.31	117.84
	RF	69.80	24.59	10190.00	0.50	100.14
	CatBoost	63.49	22.60	9114.12	0.55	94.47
	XGBoost	67.10	23.94	11864.32	0.38	107.67
	LightGBM	68.50	25.52	10050.81	0.51	99.60
D	DT	0.03	5.32	0.00	0.86	0.04
	RF	0.03	4.61	0.00	0.90	0.04
	CatBoost	0.05	7.66	0.00	0.78	0.07
	XGBoost	0.03	5.02	0.00	0.89	0.04
	LightGBM	0.03	5.41	0.00	0.88	0.04
H	DT	0.20	25.85	0.09	0.29	0.29
	RF	0.16	17.76	0.05	0.58	0.22
	CatBoost	0.18	20.62	0.06	0.48	0.24
	XGBoost	0.16	18.98	0.05	0.49	0.23
	LightGBM	0.15	17.57	0.05	0.59	0.22

Table S14 Predict the cross-validation set performance of models (XGBoost and LightGBM) for the target parameters (TN/TP and H').

Parameters	Models	MAE	MAPE (%)	MSE	R^2	RMSE
H'	LightGBM	0.09	6.33	0.02	0.90	0.12
TN/TP	XGBoost	12.03	29.42	719.85	0.86	26.83

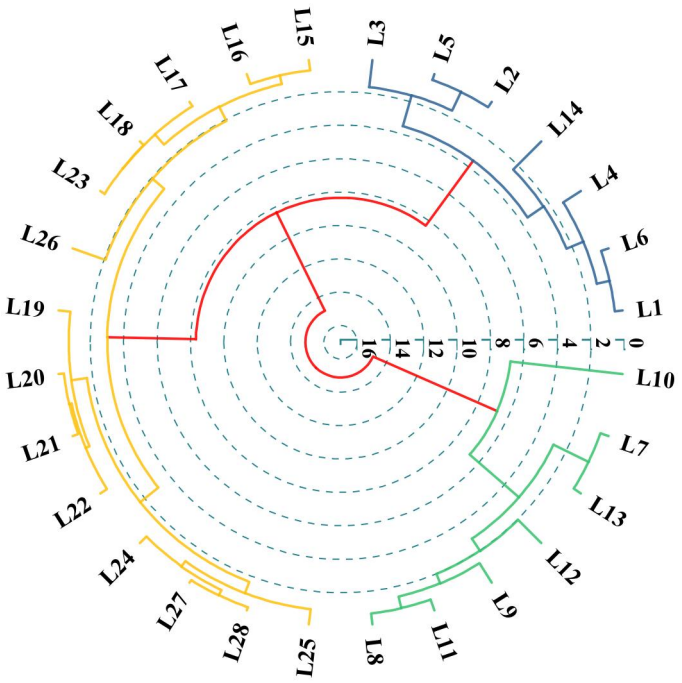


Fig. S1 Spatio-temporal heterogeneities in water quality in HuaTing Lake (China).

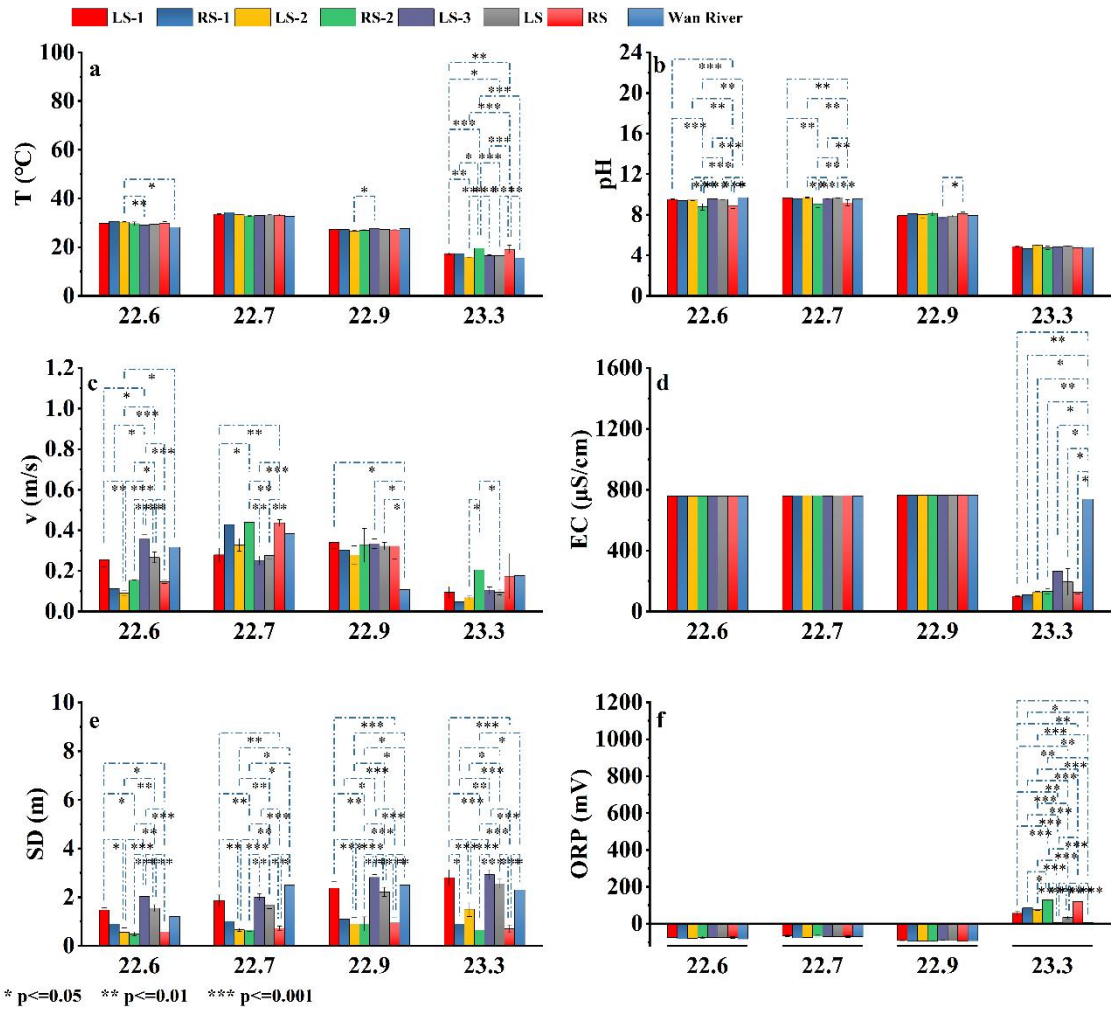


Fig. S2 Spatio-temporal variation of environmental physical parameters (T, Ph, v, EC, SD, and ORP) in Huating Lake for the months of 2022.6, 7, 9, and 2023.3: (a) T; (b) pH; (c) v; (d) EC; (e) SD; (f) ORP.

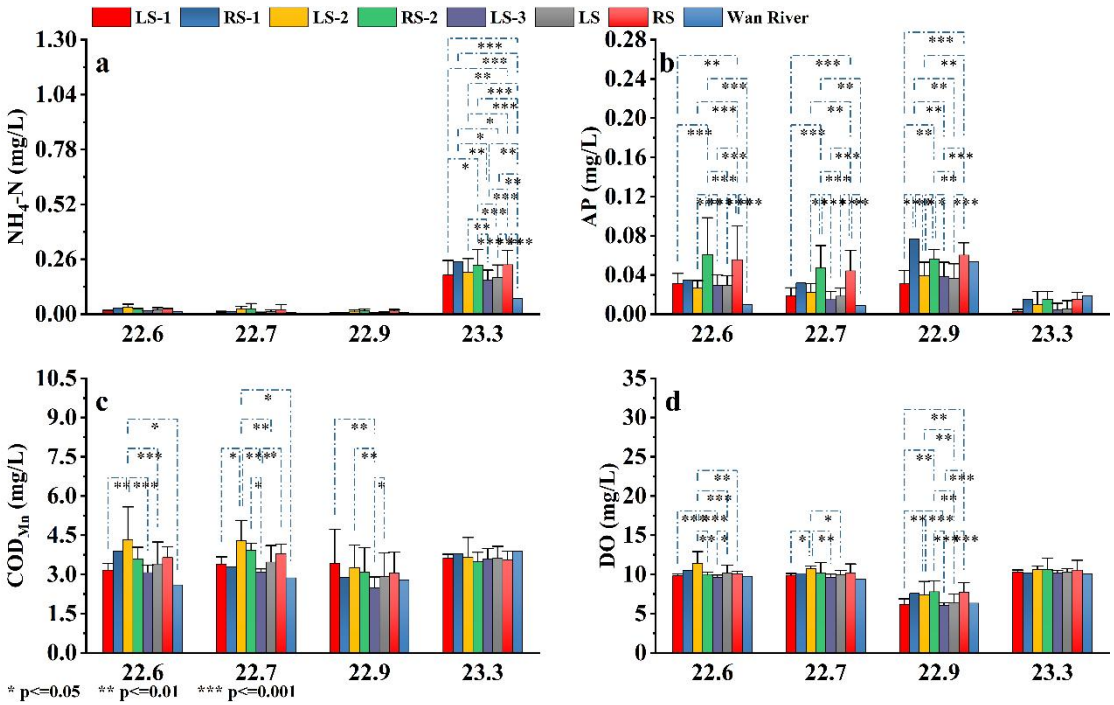
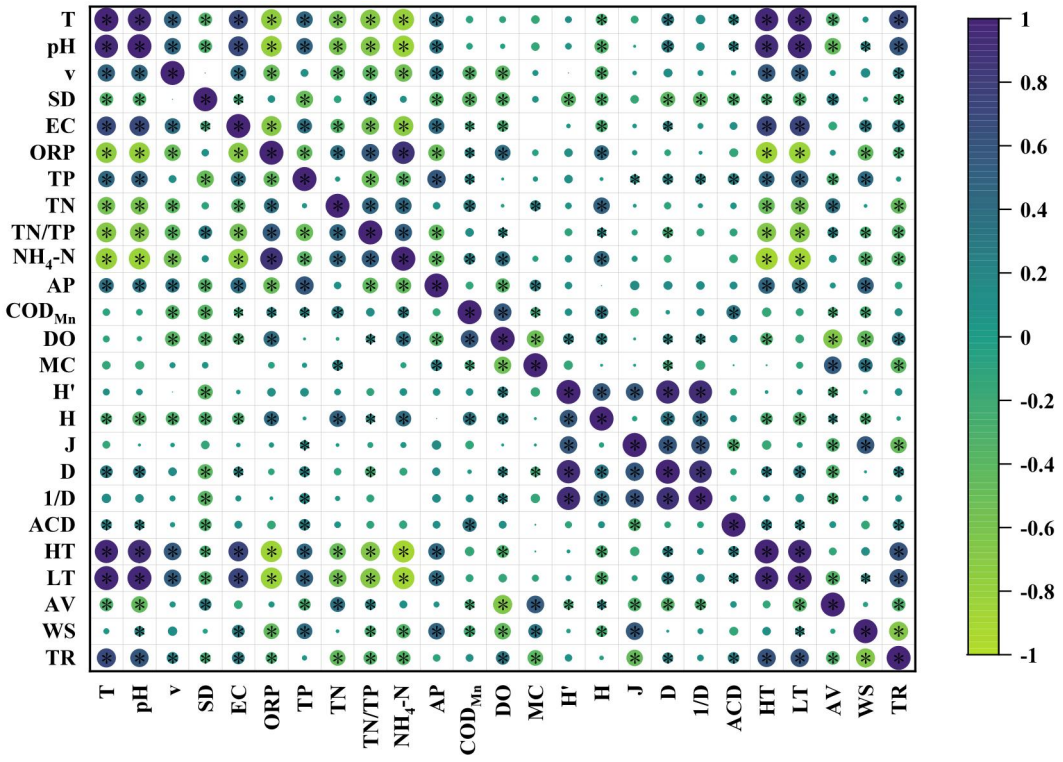


Fig. S3 Spatio-temporal variation of environmental partial chemical parameters (NH_4-N, AP, COD_{Mn}, and DO) in Huating Lake for the months of 2022.6, 7, 9, and 2023.3: (a) NH_4-N; (b) AP; (c) COD_{Mn}; (d) DO.



* p<=0.05

Fig. S4 Pearson correlation values between parameters.

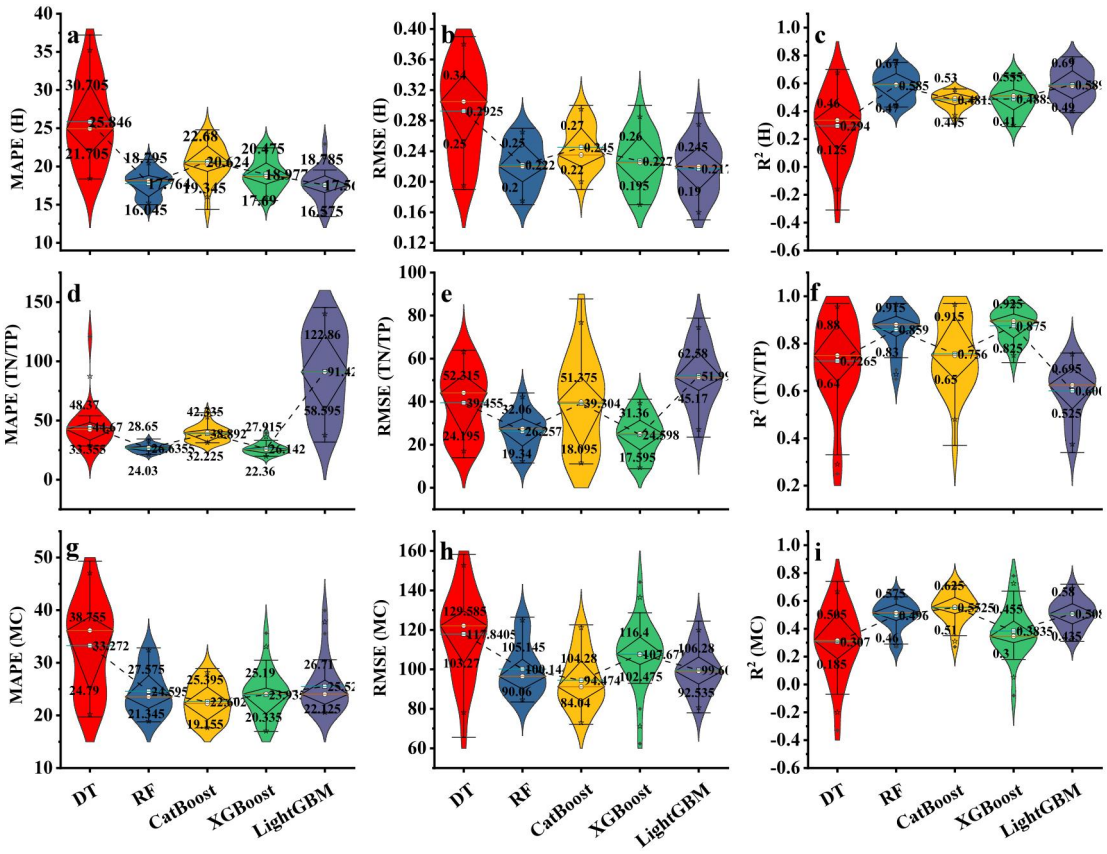


Fig. S5 Box line plots of evaluation metrics for 500 runs of 5 ML models with R^2, RMSE, and MAPE predicted for each target parameter (H, TN/TP, and MC). The two outermost horizontal lines are the outer limit, and the box is the range between the upper quartile and the lower quartile. The horizontal line inside the box is the median, the center circle is the mean, and the cross is the outliers: (a) MAPE(H); (b) RMSE(H); (c) R^2(H); (d) MAPE(TN/TP); (e) RMSE(TN/TP); (f) R^2(TN/TP); (g) MAPE(MC); (h) RMSE(MC); (i) R^2(MC).

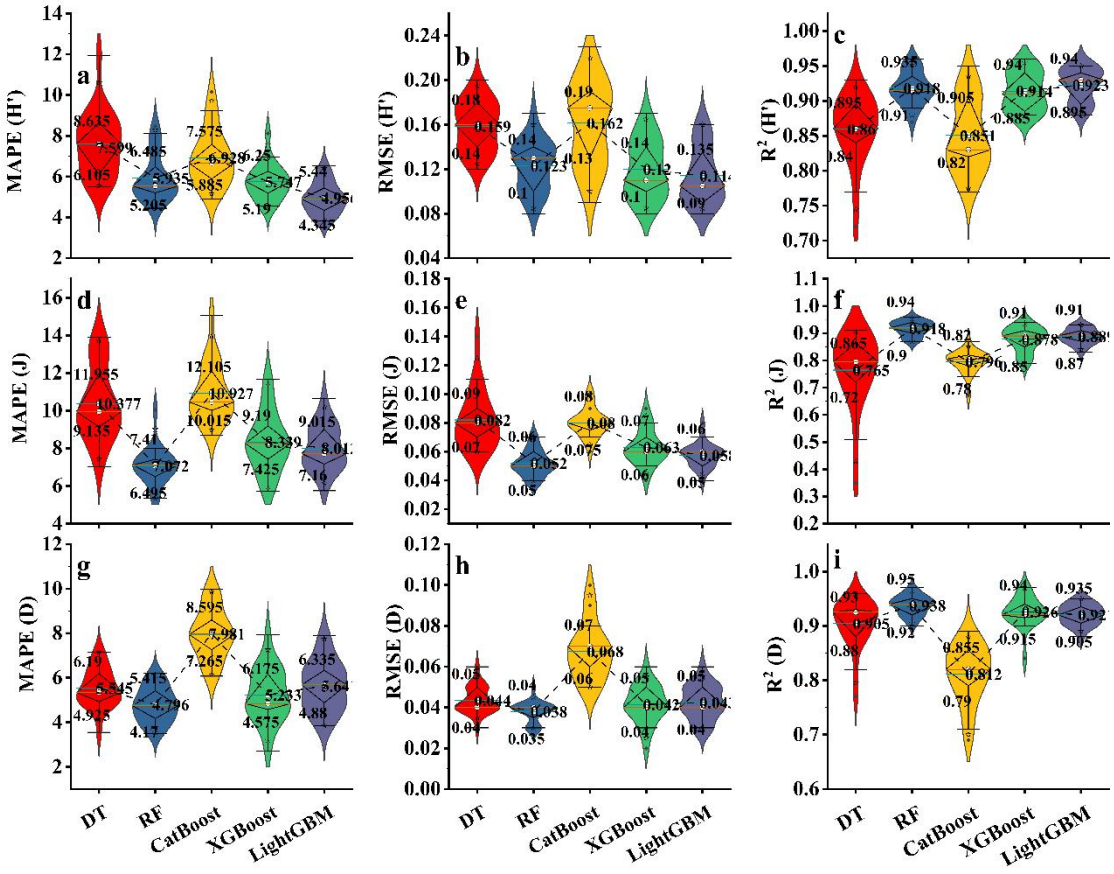


Fig. S6 Box line plots of evaluation metrics for 500 runs of 5 ML models with R^2, RMSE and MAPE predicted for each target parameter (H', J, and D). The two outermost horizontal lines are the outer limit, and the box is the range between the upper quartile and the lower quartile. The horizontal line inside the box is the median, the center circle is the mean, and the cross is outliers: (a) MAPE(H'); (b) RMSE(H'); (c) R^2(H'); (d) MAPE(J); (e) RMSE(J); (f) R^2(J); (g) MAPE(D); (h) RMSE(D); (i) R^2(D).

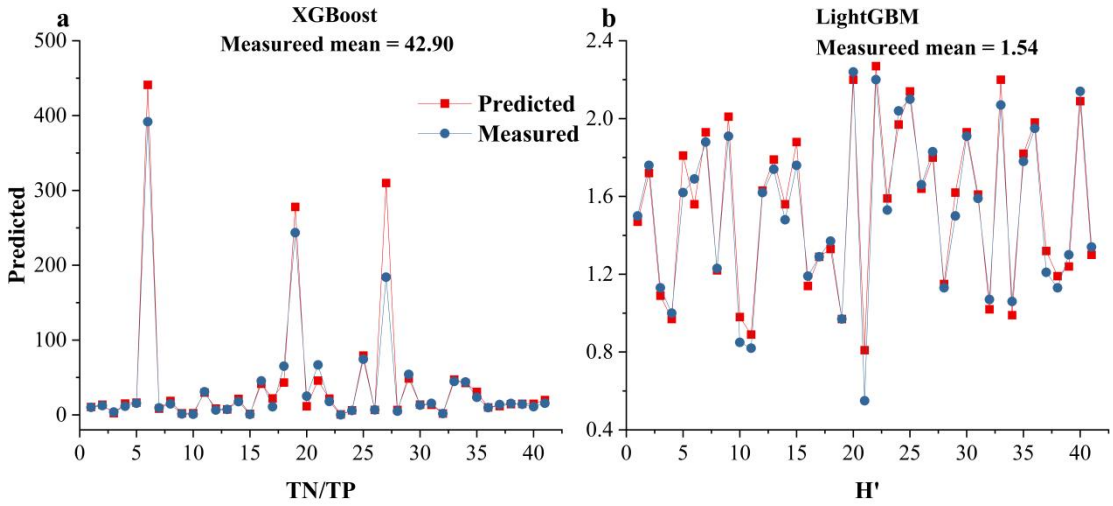


Fig. S7 Eighty percent of the 136 samples are randomly selected as the training set, and the remaining 30 % of the samples are used as the test set. (a) TN/TP prediction results of XGBoost model; (b) H' prediction result of LightGBM model.

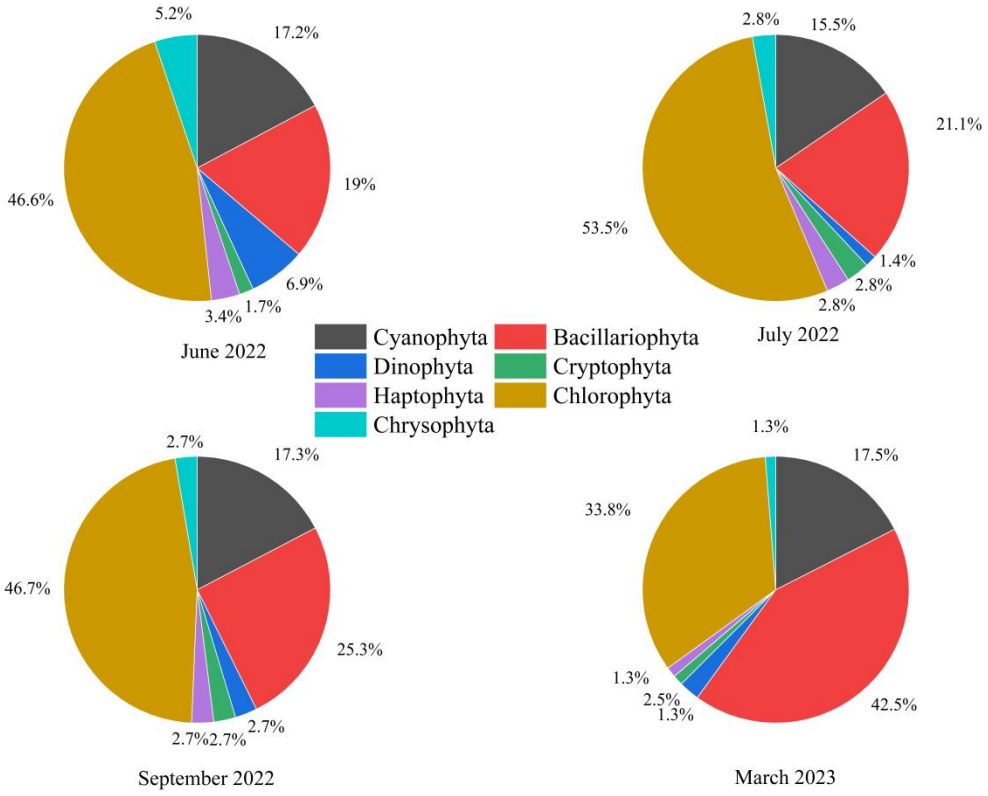


Fig. S8 Percentage of phytoplankton at different phylum levels in Huating Lake basin.

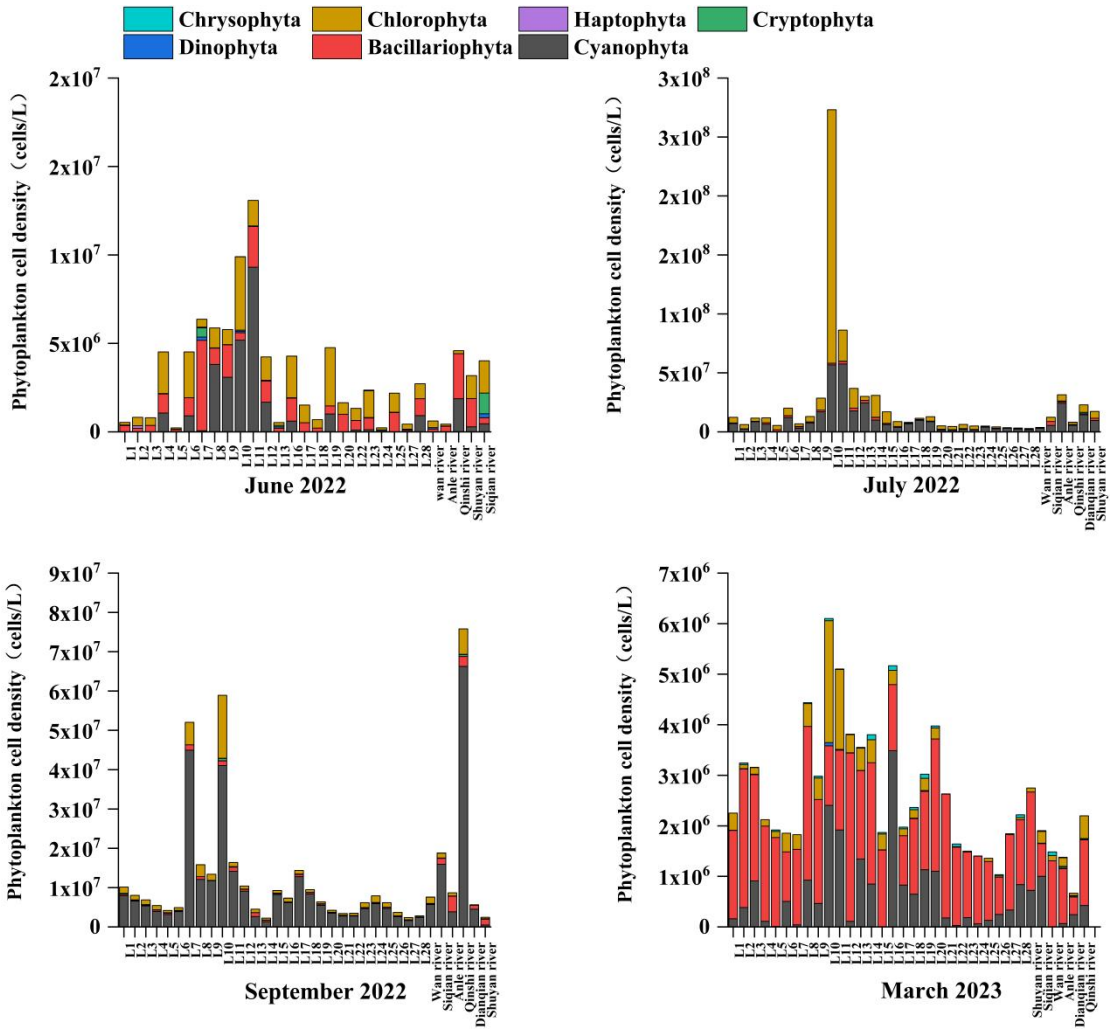


Fig. S9 Density of phytoplankton at different phyla levels in the Huating Lake Basin at various sites.

References

Belgiu M, Drăguţ L (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114: 24–31

Berk R A (2016). Classification and regression trees (CART). In: Berk R A, editor. Statistical Learning from a Regression Perspective. Cham: Springer International Publishing, 157–211

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T (2017). LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 3149–3157

Prokhorenkova L, Gusev G, Vorobev A, Dorogush A V, Gulin A (2018). CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 6639–6649

Sheridan R P, Wang W M, Liaw A, Ma J, Gifford E M (2016). Extreme gradient boosting as a method for quantitative structure–activity relationships. Journal of Chemical Information and Modeling, 56(12): 2353–2360