Appendix A. Supplementary material

1 Text S1: Traditional Machine Learning Models

1.1 KNN

KNN is a very simple and understandable ML algorithm that does not have complex learning parameters. The core of this is the fact that when we input a new or unknown sample, the new input sample is estimated according to its Euclidean distance from a known or similar sample value. In this process, the shortest distance result shows that this unknown sample is closest to a corresponding trained sample to realize the classification or regression (Guo et al., 2003).

1.2 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 (Lewis, 2000).

1.3 SVM

The basic idea of SVM learning is to find a separation hyperplane that can correctly divide the training dataset and have the largest geometric interval. That is, the model should find the most appropriate hyperplane or line in space to divide samples into its sides and ensure that all of the sample points have the furthest distance to this hyperplane or line (Cherkassky and Ma, 2004).

1.4 Bagging

Bagging uses bootstrap sampling technology to randomly generate multiple samples to form a subset. Each generated subset is used to build a base learner DT, which is then aggregated into the final model. The final prediction result is a weighted average of the results of all of the base

learners. Compared with DT, bagging effectively reduces the prediction variance and greatly improves the prediction performance and overfitting (Breiman, 1996).

1.5 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).

1.6 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).

Table S1 The coordinates of the estuary of the inflow rivers, the meteorological monitoring stations and Lake Taihu water quality monitoring stations.

Dataset	Name	Province	Longitude (°)	Latitude (°)
Inflow Rivers Estuary	Dagang River	Jiangsu	119.91	31.19
	Wuxi River	Jiangsu	119.89	31.23
	Liangxi River	Jiangsu	120.23	31.55
	Zhihu River	Jiangsu	120.13	31.51
	Wujin River	Jiangsu	120.12	31.50
	Wangyu River	Jiangsu	120.40	31.45
	Daxi River	Jiangsu	120.35	31.46
	Yincun River	Jiangsu	120.01	31.45
	Taige Canal	Jiangsu	120.04	31.48
	Baidu River	Jiangsu	120.08	31.45
	Guandu River	Jiangsu	119.96	31.35
	Hongxiang River	Jiangsu	119.94	31.32
	Chendong River	Jiangsu	119.94	31.32
	Shedu River	Jiangsu	119.97	31.37
	Dapu River	Jiangsu	119.93	31.31
	Jiapu River	Zhejiang	119.94	31.11
	Hexi River	Zhejiang	119.96	31.08
	Yangjiapu River	Zhejiang	120.01	31.02
	Changxing River	Zhejiang	119.99	31.06
	Daqian River	Zhejiang	120.19	30.93
	Shaoxi River	Zhejiang	120.13	30.94
	Maoer River	Zhejiang	120.11	30.96
Weather Station	Yixing	Jiangsu	119.82	31.33
	Wuxi	Jiangsu	120.35	31.62
	Suzhou	Jiangsu	120.43	31.07
	Huzhou	Zhejiang	120.04	30.87
Lake Taihu Water Quality Monitoring Station	THL00		120.22	31.54
	THL01		120.19	31.51
	THL03		120.19	31.48
	THL04		120.19	31.44
	THL05		120.19	31.41
	THL06		120.13	31.50
	THL07		120.18	31.34
	THL08		120.17	31.25

Table S2 Descriptive statistics of variables.

Dataset	Variable	Unit	Min.	Mean	Median	Max.	SD
Taihu Water Quality Monitoring Dataset	DO_LT	mg/L	3.35	5.26	5.11	10.05	1.30
	COD_LT	mg/L	5.20	9.08	9.02	13.44	1.71
	SS_LT	mg/L	6.39	45.33	42.77	96.99	19.81
	NH_3-N_LT	mg/L	0.02	0.41	0.31	1.37	0.30
	TN_LT	mg/L	1.21	2.82	2.59	7.39	1.13
	TP_LT	mg/L	0.05	0.14	0.12	0.33	0.06
Inflow Rivers Water Quality Monitoring Dataset	DO_JSIR	mg/L	3.15	6.40	6.23	10.30	1.49
	COD_JSIR	mg/L	3.71	5.06	5.08	7.13	0.56
	BOD_JSIR	mg/L	2.40	3.77	3.88	5.53	0.68
	NH_3-N_JSIR	mg/L	0.22	1.08	1.03	3.24	0.58
	TN_JSIR	mg/L	1.97	3.38	3.30	6.10	0.89
	TP_JSIR	mg/L	0.11	0.17	0.16	0.31	0.04
	DO_ZJIR	mg/L	4.70	8.00	7.89	12.46	1.90
	COD_ZJIR	mg/L	2.63	4.33	4.41	6.05	0.64
	BOD_ZJIR	mg/L	1.61	2.77	2.86	4.65	0.50
	NH_3-N_ZJIR	mg/L	0.02	0.36	0.34	1.05	0.20
	TN_ZJIR	mg/L	0.62	1.95	1.88	4.29	0.75
	TP_ZJIR	mg/L	0.03	0.10	0.10	0.21	0.03
WWTPs' Pollutant Discharge Dataset	COD_WD	kt	1.38	2.29	2.41	3.24	0.39
	BOD_WD	t	226.76	349.77	356.27	638.11	58.32
	SS_WD	t	390.16	539.20	547.72	684.16	59.54
	NH_3-N_WD	t	47.85	83.45	77.51	170.91	24.47
	TN_WD	t	232.02	705.89	787.12	964.90	215.62
	TP_WD	t	10.56	17.13	17.09	25.39	3.06
Output Dataset of Industrial Products	CNA_JSIP	kt	1.15	3.51	3.30	8.27	1.29
	SA_JSIP	kt	14.66	27.97	26.52	74.11	7.94
	CF_JSIP	kt	6.29	20.01	19.67	46.43	7.16
	CP_JSIP	kt	3.83	6.68	6.50	12.69	1.45
	SD_JSIP	t	90.14	1530.59	1142.61	3722.48	872.43
	CNA_ZJIP	t	49.20	566.34	559.40	1051.90	222.28
	SA_ZJIP	t	987.12	2117.49	2062.21	3368.10	546.80
	NF_ZJIP	t	419.74	1077.49	939.83	2382.70	425.46
	PF_ZJIP	t	0.61	50.02	27.15	224.14	51.21
	CP_ZJIP	t	369.81	882.64	861.60	1596.74	216.20
	SD_ZJIP	t	626.29	2444.89	2309.97	5269.21	798.84
Meteorology Dataset	RH_M	%	60.13	75.54	76.04	87.94	5.67
	AP_M	hPa	1002.12	1015.14	1016.53	1029.90	8.19
	T_M	℃	0.21	16.87	17.92	31.53	8.74
	E_M	mm	27.00	86.83	81.85	236.20	39.80
	P_M	mm	6.73	112.23	96.06	401.35	78.52
	WS_M	m/s	1.60	2.28	2.27	3.33	0.34

Table S3 All features were preserved after two-step MIC.

Prediction task	Dataset	Retained Features	Deleted Features
TN prediction	Taihu Water Quality	DO_LT, COD_LT, SS_LT, NH_3-N_LT	TP_LT
	Inflow Rivers Water Quality	COD_JSIR, BOD_JSIR, NH_3-N_JSIR, TN_JSIR, TP_JSIR, DO_ZJIR, COD_ZJIR, BOD_ZJIR, NH_3-N_ZJIR, TN_ZJIR, TP_ZJIR	DO_JSIR
	WWTPs' Pollutant Discharge	SS_WD, NH_3-N_WD, TN_WD, TP_WD	COD_WD BOD_WD
	Industrial Products	CAN_JSIP, CF_JSIP, CP_JSIP, CAN_ZJIP, SA_ZJIP, NF_ZJIP, PF_ZJIP, CP_ZJIP	SA_JSIP, SD_JSIP, SD_ZJIP
	Meteorology	RH_M, T_M, P_M, WS_M	AP_M, E_M
TP prediction	Taihu Water Quality	DO_LT, COD_LT, SS_LT, NH_3-N_LT, TN_LT	
	Inflow Rivers Water Quality	DO_JSIR, COD_JSIR, BOD_JSIR, NH_3-N_JSIR, TN_JSIR, TP_JSIR, DO_ZJIR, COD_ZJIR, BOD_ZJIR, NH3-N_ZJIR, TN_ZJIR, TP_ZJIR	
	WWTPs' Pollutant Discharge	BOD_WD, SS_WD, NH_3-N_WD, TN_WD, TP_WD	COD_WD
	Industrial Products	CAN_JSIP, CF_JSIP, CP_JSIP, CAN_ZJIP, SA_ZJIP, NF_ZJIP, PF_ZJIP, CP_ZJIP	SA_JSIP, SD_JSIP, SD_ZJIP
	Meteorology	RH_M, E_M, P_M, WS_M	AP_M, T_M

Table S4 Model performance for time series using key features.

Prediction task	Features	RMSE (mg/L)	MAPE (%)	R^2
TN time series forecast	TN	0.39	16.35	0.29
	TN, NH_3-N_LT	0.37	15.71	0.36
	TN, NH_3-N_LT, NH_3-N_JSIR	0.21	8.8	0.8
	TN, NH_3-N_LT, NH_3-N_JSIR, NH_3-N_WD	0.18	7.04	0.85
	TN, NH_3-N_LT, NH_3-N_JSIR, NH_3-N_WD, COD_LT	0.20	8.65	0.81
TP time series forecast	TP	0.027	15.27	0.81
	TP, COD_LT	0.021	8.89	0.89
	TP, COD_LT, DO_JSIR	0.015	6.89	0.94
	TP, COD_LT, DO_JSIR, SS_LT	0.014	6.06	0.95
	TP, COD_LT, DO_JSIR, SS_LT, TN_JSIR	0.020	10.49	0.9

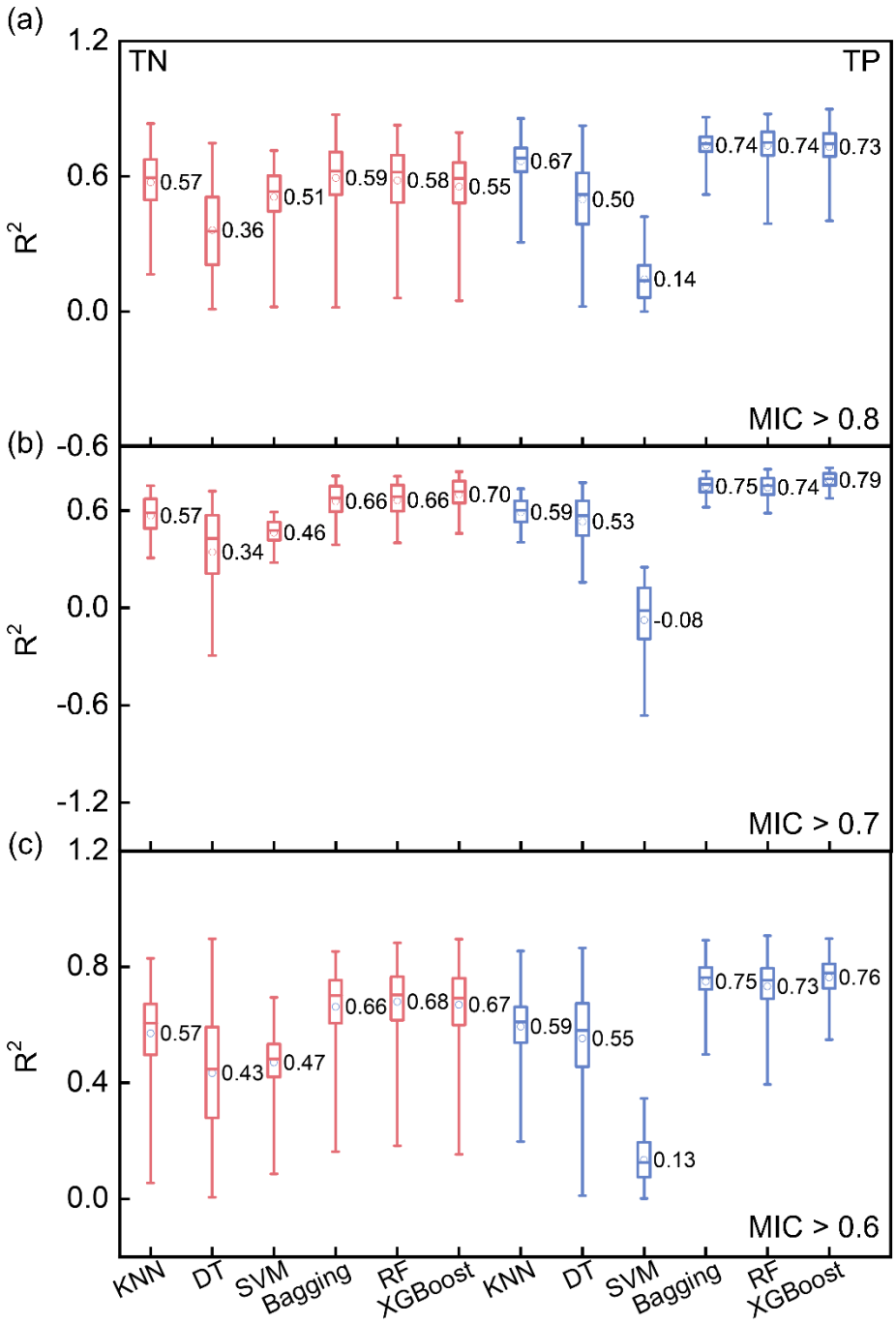

Fig. S1 Performance of 6 models for nitrogen and phosphorus prediction with different MIC values. Red represents the TN prediction task, and blue represents the TP prediction task. The two outermost horizontal lines represent the outer limit, and the box represents the range between the upper quartile and the lower quartile. The horizontal line inside the box represents the median, the center circle represents the mean, and the cross represents outliers. (a) Models R^2 for the TN and TP prediction tasks when MIC > 0.8 is used as the feature screening criterion; (b) Models R^2 for the TN and TP prediction tasks when MIC > 0.7 is used as the feature screening criterion; (c) Models R^2 for the TN and TP prediction tasks when MIC > 0.6 is used as the feature screening criterion.

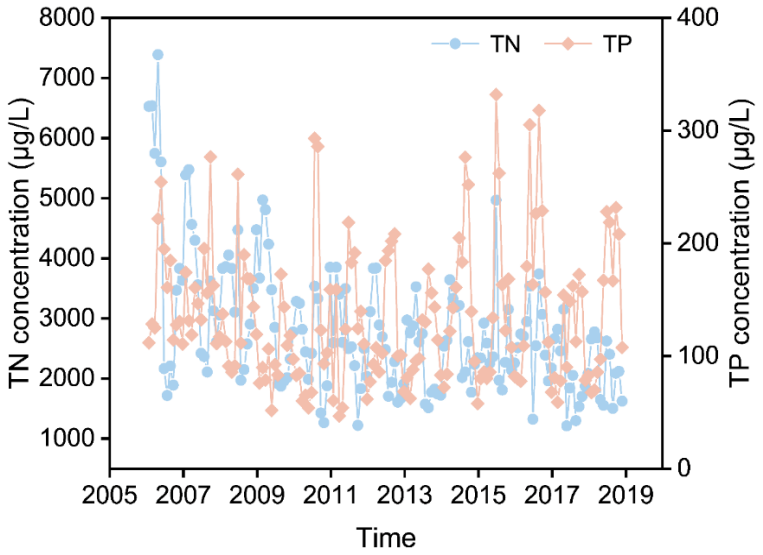

Fig. S2 Variations in TN and TP concentrations in Lake Taihu from 2007 to 2019.

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

Breiman L (1996). Bagging predictors. Machine learning, 24(2): 123–140

Cherkassky V, Ma Y (2004). Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks, 17(1): 113–126

Guo G, Wang H, Bell D, Bi Y, Greer K (2003). KNN model-based approach in classification. In: Robert M, Zahir T, Douglas C S, editors. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Berlin, Heidelberg: Springer, 986–996

Lewis R J (2000). An introduction to classification and regression tree (CART) analysis. In: Annual meeting of the society for academic emergency medicine in San Francisco, California. Princeton: Citeseer, 14

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