Supplementary Materials

Table S1. Current status of collected data per process of the M water purification plant

Process	Water quality items	Period	Interval
Receiving well	pH, temperature, and flow	Jan.5, 2015–Dec. 30, 2019	Hour
	DOC and UV254		Day
Coagulation/ precipitation	pH, coagulant dosage, and retention time	Jan.5, 2015–Dec. 30, 2019	Hour
	DOC and UV254		Day
Clearwell	pH, temperature, flow, residual chlorine, and retention time	Jan.5, 2015–Dec. 30, 2019	Hour
	DOC, UV254, and Br		Day

Table S2. Explanation regarding the comparison of each ML model

ML Algorithms	Type	Pros	Cons
Decision Tree (DT)	Classification/ Regression	Simple, Interpretable, Visualizable	Prone to overfitting, unstable
Random Forest (RF)	Classification/ Regression	Handle large datasets, complex problems	Slow to train, hard to tune
Support Vector Machine (SVM)	Classification/ Regression	Work well with high-dimensional, nonlinear data	Sensitive to outliers, noise, kernel, parameters
K-nearest Neighbor (KNN)	Classification/ Regression	Simple, work well with small, clean datasets	Slow to predict, affected by irrelevant features
Gradient Boosting Regressor (GBR)	Regression	High accuracy, handle various types of data	Prone to overfitting, require tuning of learning rate
Multiple-Layer Perceptron (MLP)	Classification/ Regression	Learn complex, nonlinear patterns	Prone to overfitting, require a lot of data

Table S3. Hyperparameters for each AI algorithm

Algorithm	Main HPs		Optional HPs	
	Parameter	**Value**	**Parameter**	**Value**
DT	Criterion	squared_error	splitter	best
	max_depth	None	min_weight_fraction_leaf	0
	min_samples_split	2.0	max_leaf_nodes	None
	min_samples_leaf	10.0	-	-
RF	n_estimators	500	max_feature	1.0
	criterion	squared_error	min_weight_fraction_leaf	0
	max_depth	None	max_leaf_nodes	None
	min_samples_split	2.0	-	-
	min_samples_leaf	1.0	-	-
SVM	C	1.0	coef	0
	kernel	rbf	degree	3.0
	gamma	scale		
KNN	n_neighbors	5.0	weights	uniform
	-	-	p	2.0
	-	-	algorithm	auto
GBR	loss	squared_error	max_feature	None
	learning_rate	0.001	alpha	0.9
	n_estimators	10000	max_leaf_nodes	None
	max_depth	4.0	-	-
MLP	hidden_layer_size	(50, 100)	learning_rate_int	0.001
	activation	relu	max_iter	1000
	learining_rate	constant	random_state	None
	n_iter_no_change	50.0	-	-