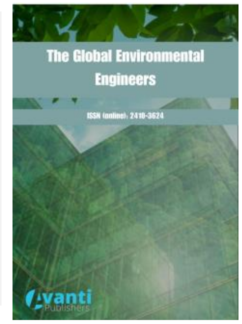
The Global Environmental Engineers, 2026, 47-67

Published by Avanti Publishers

The Global Environmental Engineers

ISSN (online): 2410-3624



Predicting and Managing Water Quality in Complex River Networks Using a Machine Learning Approach: A Case Study of the Qinhuai River Basin

Sheng Zhang[1,#], Shaoyan Feng[1,#] and Xin Qian[1,2,*]

[1]State Key Laboratory of Water Pollution Control and Green Resource Recycling, School of Environment, Nanjing University, Nanjing 210023, China

[2]Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science & Technology, Nanjing, 210044, China

ARTICLE INFO

Article Type: Research Article

Academic Editor: Nannan Wang

Keywords:
Machine learning
Scenario simulation
Water quality prediction
River basin management
Non-point source pollution

Timeline:
Received: May 05, 2026
Accepted: June 12, 2026
Published: June 29, 2026

Citation: Zhang S, Feng S, Qian X. Predicting and managing water quality in complex river networks using a machine learning approach: A case study of the Qinhuai River Basin. Glob Environ Eng. 2026; 13(1): 47-67.

DOI: https://doi.org/10.15377/2410-3624.2026.13.5

[#]The two authors contributed equally to this work.

*Corresponding Author
Email: xqian@nju.edu.cn
Tel: +(86) 1377 0668 211

ABSTRACT

As water pollution intensifies, traditional monitoring and modeling methods struggle to address the complexities of dynamic water environments, particularly in complex river networks. Machine learning, a powerful data-driven model, has become a key research focus in water resource management due to its accurate predictive capabilities. This study examines the Qinhuai River Basin (QRB) in the lower Yangtze River, using water quality data from 10 monitoring stations (2021-2023) for dissolved oxygen (DO), permanganate index (CODMn), ammonia nitrogen (NH_3-N), and total phosphorus (TP). The study first analyzes the spatiotemporal distribution of water quality, then develops machine learning models to identify key factors affecting water quality at monitoring points. Finally, scenario simulations and management strategies are proposed. The results show that upstream DO and CODMn concentrations are higher, while NH_3-N and TP are lower compared to downstream. Summer is the primary period for water quality exceedance, with DO as the key indicator. XGBoost outperforms other models with R^2 values of 0.882–0.947 for CODMn and 0.824–0.890 for NH_3-N. Key factors influencing water quality include upstream water quality, 48-hour cumulative rainfall runoff, basin inflow and outflow, and wastewater treatment plant (WWTP) effluent. Scenario simulations reveal that improving upstream water quality to Class III standards leads to downstream improvements, while rainfall runoff interception, ecological water replenishment, and reducing WWTP discharge can effectively reduce water quality indicator concentrations. This study provides useful insights into water quality management and demonstrates the potential of the proposed framework as a transferable tool for supporting adaptive water quality management in other data-limited and complex river basins.

1. Introduction

Rivers, as one of the most important freshwater resources, are essential for both human societies and natural ecosystems. They play a vital role in drinking water supply, hydropower generation, transportation, and agricultural irrigation [1-3]. A healthy river environment is a fundamental prerequisite for human well-being, ecosystem stability, and sustainable socio-economic development. However, river water pollution has become one of the most pressing environmental challenges of the twenty-first century. Over the past decades, rapid industrialization and urbanization in China have resulted in the direct discharge of large amounts of industrial effluents and municipal wastewater into rivers, leading to a sharp increase in pollutant loads [4-6]. This has severely degraded riverine environments and ecosystems, thereby exacerbating the problem of freshwater scarcity.

At present, water quality models can be broadly classified into process-based (mechanistic) models and data-driven machine learning models. Widely used process-based models include EFDC, WASP, and MIKE [7-10]. For example, Jiang *et al*. [11] employed the EFDC model to simulate nutrient dynamics in Lake Taihu. Li *et al*. [12] developed the MIKE21 model to assess the impacts of water diversion on lake water quality. However, conventional process-based water quality models generally require detailed hydrodynamic information, numerous model parameters, and well-defined boundary conditions, many of which are difficult to obtain in practice. Their performance also relies heavily on extensive parameter calibration and an accurate representation of complex physical and biogeochemical processes. These limitations become particularly pronounced in complex river networks, where hydraulic structures (e.g., sluices and gates) create highly dynamic flow conditions, intensive human activities generate multiple and spatially heterogeneous pollution sources, and monitoring data are often insufficient to support process-based model development and calibration. Consequently, the applicability of conventional process-based models for water quality prediction and management in such systems is often constrained [13-17]. In recent years, data-driven machine learning (ML) approaches have emerged as a promising alternative because of their strong nonlinear mapping capability and data-learning ability. Unlike process-based models, ML methods can directly learn complex relationships between environmental variables and water quality responses from historical observations without explicitly describing the underlying physical processes. Their flexibility, high predictive performance, and relatively low dependence on detailed process information make them particularly suitable for water quality prediction and management in complex, data-limited river basins [18-21]. With the continuous improvement of environmental monitoring systems in China, automatic monitoring stations can now provide high-frequency and large-volume datasets, offering robust data support for the development and application of machine learning models.

The Yangtze River Basin is the largest river basin in China and plays a critical role in the country's socio-economic development by providing abundant water resources for industrial and agricultural production as well as domestic use [22]. In the context of increasing emphasis on ecological and environmental protection, water environment conservation in the Yangtze River Basin has become a key component of the national "Yangtze River Protection" strategy. Benefiting from the comprehensive implementation of basin-wide protection policies, the overall quality of the aquatic ecosystem in the Yangtze River Basin has improved substantially. However, rivers in certain sub-regions of the basin still suffer from serious non-point source pollution, deterioration of water quality during rainfall events, and direct discharge of untreated wastewater [23-26]. To achieve the sustainable development of water resources, water environment, and aquatic ecosystems in the Yangtze River Basin, it is essential to focus on tributary watersheds as fundamental management units. A thorough analysis of water quality evolution patterns in individual tributary basins is required to develop targeted and effective water quality management strategies, thereby ensuring compliance with water quality standards, particularly the long-term and stable attainment of standards at river sections discharging into the main stem of the Yangtze River.

Therefore, this study selected the Qinhuai River Basin, a first-order tributary located on the southern bank of the lower reaches of the Yangtze River, as the study area. As an important economic and cultural hub, the water environmental quality of the Qinhuai River Basin is of strategic significance for regional sustainable development. From the perspective of watershed water environment management, long-term time-series water quality data were utilized in this study. First, the spatiotemporal distribution characteristics of water quality in the basin were comprehensively analyzed and evaluated to accurately characterize the overall water environmental status. Second, multiple machine learning models were developed and systematically compared, and the model that best captures the complex water quality dynamics of river network systems was selected to improve the accuracy and reliability of water quality prediction. Finally, considering the actual conditions of the basin, water quality improvement and pollution control strategies were integrative designed and evaluated, and targeted water environment management measures were proposed. The findings of this study provide a scientific basis for sustainable and refined watershed management and offer practical insights into water environment governance in typical tributary basins of the Yangtze River.

2. Materials and Methods

2.1. Study Area

The Qinhuai River Basin (QRB) (118°39'E–119°19'E, 31°34'N–32°10'N) is a first-order tributary located on the southern bank of the lower reaches of the Yangtze River. It flows through the main urban area of Nanjing, Jiangning District, Lishui District, and Jurong City [27, 28], as shown in Fig. (**1**). The river network in the basin is well developed, with numerous tributaries. The main stem has a total length of approximately 110 km, and the basin covers an area of about 2,631 km² [29, 30]. The basin has an average annual temperature of 15.4°C and an average annual precipitation of 1048 mm, with pronounced intra-annual variability [31]. Precipitation is mainly concentrated during the flood season from June to September [32], accounting for approximately 60% of the annual total, which makes the basin prone to flooding events [33, 34]. In addition, stormwater overflows during rainfall often lead to the deterioration of river water quality. A hydraulic control structure, the New Qinhuai River hydro-junction, is constructed on the Qinhuai River at the confluence of the Qinhuai New River and the Yangtze River and serves as one of the key hydraulic hubs in the basin. The gate plays an important role in water diversion for pollution flushing as well as flood control and drainage. The water diversion and discharge conditions of the control gate from 2021 to 2023 are illustrated in Fig. (**S1**). The daily water diversion volume ranged from 0 to 5.97 × 10^6 m³ d^{-1}, while the daily discharge volume ranged from 0 to 5.574 × 10^7m³ d^{-1}. The gate operated predominantly in diversion mode, with an average of 272 days per year under water diversion, whereas drainage operations were

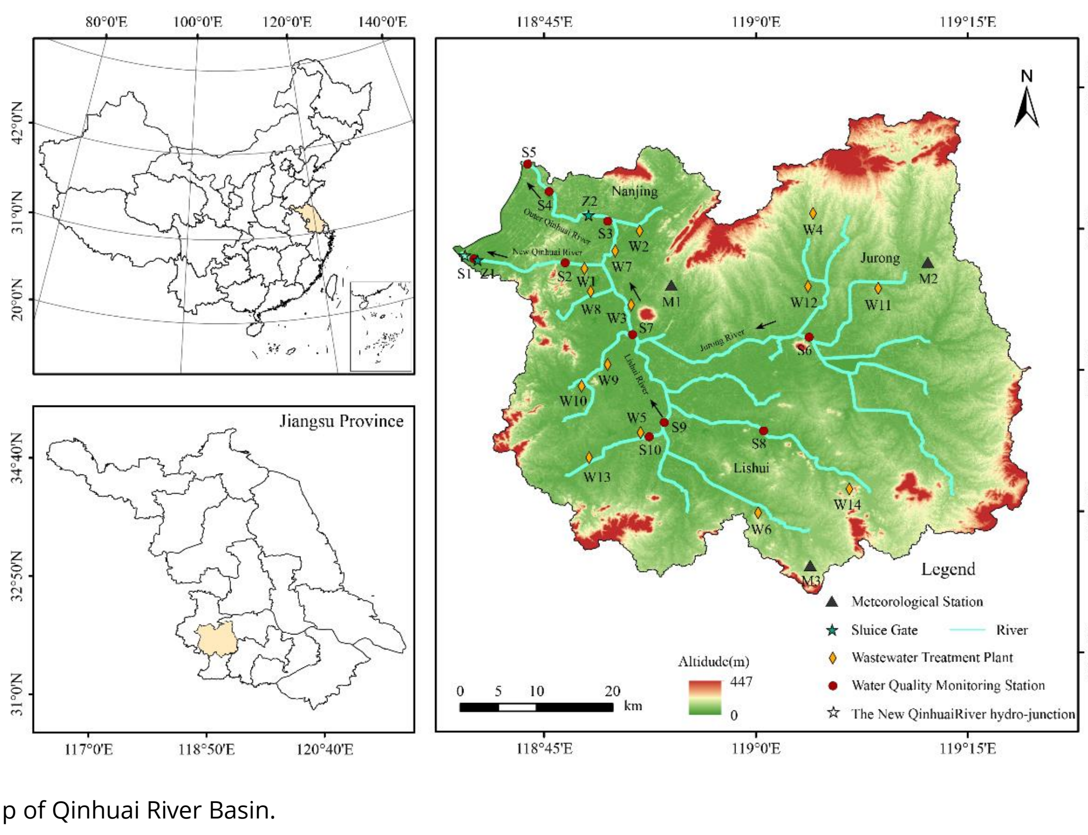


Figure 1: Map of Qinhuai River Basin.

mainly concentrated during the flood season from June to August. At present, studies on the spatiotemporal distribution of water quality in the Qinhuai River Basin have mainly focused on the downstream urban reaches and the upstream Jurong section, whereas basin-wide investigations of water quality evolution remain limited [35-37]. Moreover, the relatively low frequency of water quality monitoring in previous studies has introduced a degree of randomness and uncertainty into the results, and research on comprehensive pollution control strategies at the basin scale is still insufficient.

2.2. Data Collection and Processing

The datasets used in this study mainly included water quality data, meteorological data, hydrological data, and wastewater treatment plant (WWTP) data. A total of ten automatic river water quality monitoring stations were deployed in the Qinhuai River Basin (Table **S1** and Fig. **1**), namely Jiezhizha (S1), Tiexinqiao (S2), Qiqiaoweng (S3), Shitoucheng (S4), Sanchahekou (S5), Tuqiao (S6), Yangqiao (S7), Kaitaiqiao (S8), Wuchaqiao (S9), and Huangqiao (S10). Four key water quality parameters were collected, including dissolved oxygen (DO), permanganate index (CODMn), ammonia nitrogen (NH_3–N), and total phosphorus (TP), which are priority indicators for water quality management in the Qinhuai River Basin. Among them, DO was monitored at an hourly interval, whereas the other parameters were monitored every four hours.Three meteorological monitoring stations were located within the basin (Fig. **1**), including Jiangning (M1), Jurong (M2), and Lishui (M3). Meteorological variables comprised sea-level pressure, wind direction, wind speed, air temperature, relative humidity, and precipitation, all recorded at an hourly frequency. Hydrological data included the water diversion and discharge volumes of the New Qinhuai River hydro-junction and the discharge volume of the Wudingmen Gate, with a daily temporal resolution. In addition, fourteen major wastewater treatment plants operate within the Qinhuai River Basin, and their basic information is provided in Table **S2**. The collected WWTP data included chemical oxygen demand (COD), NH_3–N, TP, and daily wastewater discharge volume, with a daily sampling frequency. All datasets covered the period from 2021 to 2023 and were obtained from local environmental protection agencies or meteorological authorities. Due to the influence of extreme weather events and occasional equipment failures, missing values and outliers were present in the raw monitoring data. Prior to analysis, data preprocessing was performed, in which missing values were filled using linear interpolation, and outliers were identified and removed using the boxplot method.

2.3. Water Quality Assessment Method

In this study, the single-factor pollution index method was used to evaluate water quality [38]. This method assesses each water quality parameter by comparing its measured concentration with the corresponding standard limit. It is straightforward to calculate and provides an intuitive representation of the pollution level for individual water quality indicators, facilitating the identification of key pollutants [39]. In this study, the standard limits for each water quality parameter were based on Class III of the *Environmental Quality Standards for Surface Water (GB 3838–2002)* [40] (Table **S3**). Single-factor pollution indices were calculated for four key parameters:DO, COD_{Mn}, NH_3–N, and TP. A pollution index greater than 1 indicates that the water quality parameter exceeds the standard, with higher values corresponding to more severe pollution. The exceedance rate was further used to quantify the frequency of standard violations for each parameter. Detailed calculation procedures are provided in Text **S1**.

2.4. Development of Machine Learning Models for Water Quality Prediction

2.4.1. Model Selection

Machine learning models can exhibit varying performance across different datasets; therefore, multiple models were selected for comparative evaluation. In this study, five models were chosen: Elastic Net (EN), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Detailed descriptions of these five models are provided in Text **S2**.

2.4.2. Variable Selection

In this study, three monitoring stations were selected as the study targets: Qiqiaoweng (S3) and Yangqiao (S7), which are national assessment stations, and Sanchahekou (S5), representing the basin outlet. Water quality models for COD_{Mn} and NH_3–N were constructed for these stations.

The initial input variables were selected based on the major factors influencing river water quality in the Qinhuai River Basin, including upstream inflow water quality, water diversion and drainage operations, wastewater treatment plant (WWTP) effluent discharge, and meteorological variables (precipitation-runoff, wind speed, wind direction, air temperature, humidity, and atmospheric pressure). These variables comprehensively represent the hydrological conditions, anthropogenic activities, and meteorological processes that are expected to affect water quality dynamics in the study area. According to the nearest-distance principle, meteorological data from Jiangning were used as input for all three models. Precipitation-runoff was estimated using the composite runoff coefficient method. Based on the *Code for Design of Building Water Supply and Drainage* (GB50015–2019), a 2 km buffer of built-up land and farmland was used as weights to calculate the composite runoff coefficients for each section. The resulting coefficients for Sanchakou, Qiqiaoweng, and Yangqiao were 0.871, 0.777, and 0.373, respectively. Furthermore, Since the effects of upstream inflow, water diversion and drainage operations, WWTP effluent discharge, and rainfall–runoff are not instantaneous but propagate through the river network over time, their influences on downstream water quality were considered using time-lagged variables. Candidate lag times of 4, 8, 24, and 48 h were selected to represent short-, medium-, and relatively long-term response periods under different hydrological conditions. Initially, the numbers of input variables for the Sanchahekou, Qiqiaoweng, and Yangqiao models were 32, 25, and 24, respectively, with an average sample size of approximately 6,300 after data preprocessing (Table **1**). To identify the optimal lag time for each variable, variable importance analysis was performed for each prediction model, and only the lag time with the highest contribution was retained as the final model input. Therefore, candidate lag times were intentionally designed to span multiple temporal scales rather than assuming a fixed response time, thereby enabling the models to identify the most informative lag for each predictor according to the data.

Table 1: Input factors and sample size of the model.

Target		Water Quality of Upstream Inflow	Water Diversion and Discharge Volume	Discharge from WWTP	Meteorological Factors	Sample Size
Sanchahekou	COD_{Mn}	The current, 4-hour lagged, and 8-hour lagged concentrations of COD_{Mn} and NH_3–N at the Qiqiaoweng (S3) station.	The current, 8-hour, 24-hour, and 48-hour lagged water diversion and discharge volumes at the New Qinhuai River hydro-junction, and the Wudingmen.	The current, 4-hour lagged, and 8-hour lagged effluent COD, NH_3–N concentrations, and flow rates from the Chengdong and Kaifaqu WWTP	Precipitation-runoff: cumulative runoff for the previous 4, 8, 24, and 48 hours. Other meteorological factors: current wind speed, wind direction, air temperature, humidity, and atmospheric pressure.	6494
	NH_3–N					6494
Qiqiaoweng	COD_{Mn}	The current, 4-hour lagged, and 8-hour lagged concentrations of COD_{Mn} and NH_3–N at the Tiexinqiao(S2) and Yangqiao ((S7) station.	The current, 8-hour, 24-hour, and 48-hour lagged water diversion and discharge volumes at the New Qinhuai River hydro-junction	The current, 4-hour lagged, and 8-hour lagged effluent COD, NH_3–N concentrations, and flow rates from the Kaifaqu WWTP		6529
	NH_3–N					6534
Yangqiao	COD_{Mn}	The current, 4-hour lagged, and 8-hour lagged concentrations of COD_{Mn} and NH_3–N at the Tuqiao (S6), Wushaqiao (S9) and Kaitaiqiao (S8) station.	-	The current, 4-hour lagged, and 8-hour lagged effluent COD, NH_3–N concentrations, and flow rates from the Lukou		5832

2.4.3. Model Training

Prior to model training, the cleaned data were standardized to eliminate the influence of differing units among feature variables, ensuring comparability across all variables. As shown in Fig. (**2**), model training involved three main steps: dataset partitioning, hyperparameter tuning, and model performance evaluation. The final performance of the models largely depended on the optimization of hyperparameters. For dataset partitioning, two-thirds of the total samples were randomly selected as the training set, with the remaining one-third used as the test set. During hyperparameter tuning, the search space for each model was first defined, followed by random search combined with ten-fold cross-validation to determine the optimal hyperparameter combination. The selected hyperparameters for the five models and their descriptions are provided in Text **S3**. Model performance was evaluated using the coefficient of determination (R^2) and the root mean square error (RMSE) [41]. R^2 reflects the proportion of variance in the dependent variable explained by the independent variables, ranging from 0 to 1, while RMSE indicates the deviation between observed and predicted values, ranging from 0 to +∞ [42]. Higher R^2 and lower RMSE values indicate better model fitting [43]. Moreover, smaller differences in R^2 and RMSE between the training and test sets indicate stronger model generalization ability. In addition, the optimal lag time for each input variable was determined based on variable importance. Variable importance was quantified by calculating the increase in prediction error after permuting each factor [44]. Using mean squared error as the metric, the relative impact of each input variable on model performance was assessed, and the final models were constructed based on these results.

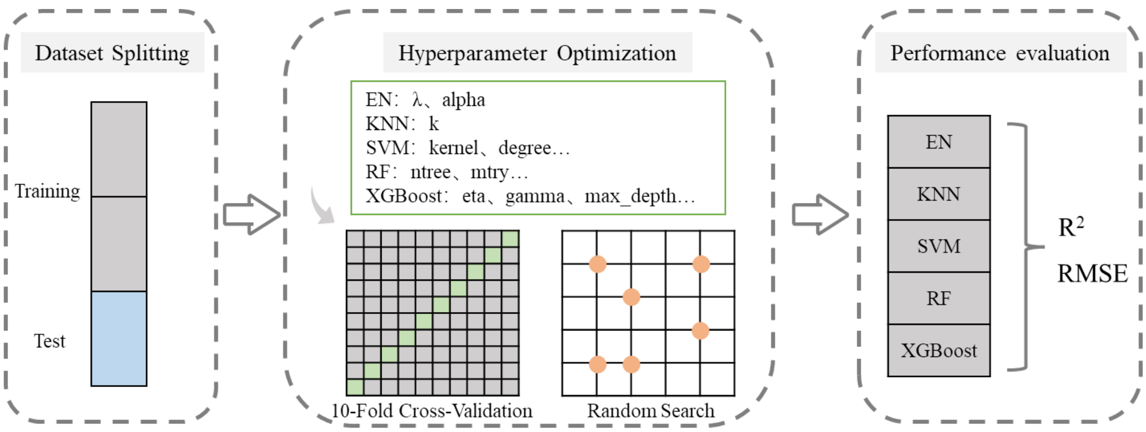


Figure 2: Training process of model.

2.4.4. Model Interpretation

SHAP (Shapley Additive Explanations) analysis and partial dependence plots (PDPs) were used to interpret the models. SHAP analysis is a game-theoretic method for explaining machine learning models [45, 46], which calculates the SHAP value for each sample and reveals how input variables influence the output. A positive SHAP value indicates a positive contribution of the input variable to the prediction, whereas a negative value indicates a negative contribution. The magnitude of the SHAP value represents the strength of the input variable's impact. In this study, the mean absolute SHAP value was used to quantify the importance of each input variable. Partial dependence plots illustrate the marginal effect of one or two target variables on the predicted outcome [47]. Specifically, while holding other variables constant, the target variable is varied from high to low, and the partial dependence function calculates the corresponding average predicted outcome. Line plots are then drawn to visualize the relationship between the target variable and the prediction.

2.5. Scenario Design for Water Pollution Control

Based on the previously constructed machine learning models and the actual conditions of the basin, five major categories of scenarios were designed to investigate the effects of different measures on water quality at the monitoring stations: upstream inflow pollution reduction (S1), precipitation-runoff interception (S2), ecological water supplementation (S3), wastewater treatment plant effluent reduction (S4), and a combined scenario (S5). Specifically, in scenario S1, the concentrations of all water quality parameters in upstream inflow were set to meet Class III surface water standards, with COD_{Mn} at 6 mg/L and NH_3–N at 1 mg/L. In scenario S2, precipitation-runoff was intercepted within a 2 km buffer zone, achieved by converting built-up land to green space, increasing the green area by 10% (S2_1), 20% (S2_2), and 30% (S2_3). Scenario S3 involved ecological water supplementation via the New Qinhuai River hydro-junction, with diversion volumes set at 0 m^3 (S3_1), 0.9 × 10^6 m^3 (S3_2), 2.0 × 10^6 m^3 (S3_3), 3.5 × 10^6 m^3 (S3_4), and 5.0 × 10^6 m^3 (S3_5). The scenario with 0 m^3 diversion (S3_1) was defined as the baseline because it represents the reference condition without additional ecological water supplementation through the New Qinhuai River hydro-junction. This scenario reflects the water quality conditions under the existing pollution loads and hydrological conditions without external water diversion, thereby providing a practical benchmark for evaluating the effectiveness of ecological water supplementation. Using S3_1 as the baseline also enables a quantitative assessment of the incremental benefits associated with different diversion volumes by distinguishing the contribution of ecological water diversion to pollutant dilution and water quality improvement from other influencing factors. In scenario S4, the effluent from wastewater treatment plants was reduced through measures such as reclaimed water reuse and regulation tanks, with reduction rates of 10% (S4_1), 20% (S4_2), and 30% (S4_3). Scenario S5 represented a comprehensive scenario, combining individual measures to assess the overall improvement in water quality. For the Sanchahekou and Qiqiaoweng stations, the combined scenario included “upstream inflow pollution reduction (S1) + precipitation-runoff interception (S2_1) + ecological water supplementation (S3_2) + WWTP effluent reduction (S4_1)”; for the Yangqiao station, the combination was “upstream inflow pollution reduction (S1) + precipitation-runoff interception (S2_1) + WWTP effluent reduction (S4_1)”.

2.6. Data Analysis and Visualization

ArcGIS 10.2 was used to create maps of the basin, and R v4.3.2 was employed for data analysis and processing, enabling data visualization, including spatiotemporal analysis of water quality and the construction and evaluation of machine learning models. Model building and evaluation were performed using the mlr package, while plots were generated using the ggplot2 package. Python 3.10 was used for model interpretation and related analyses.

3. Results and Discussion

3.1. Spatial Characteristics of Water Quality

Based on the calculations, the mean values and standard deviations (SD) of the single-factor pollution indices for DO, COD_{Mn}, NH_3–N, and TP at the ten automatic water quality monitoring sections in the Qinhuai River Basin are presented in Table **2**. The average pollution indices of DO at the Jiezhizha (S1), Tiexinqiao (S2), and Wuchaqiao (S9) sections were 0.63 ± 1.19, 0.78 ± 1.20, and 0.81 ± 1.03, respectively, indicating overall compliance with the water quality standard. In contrast, the mean DO pollution indices at the remaining sections exceeded 1, with the highest values observed at the Shitoucheng (S4) and Sanchahekou (S5) sections, suggesting severe exceedance of DO at these locations. Across the entire basin, the mean pollution indices of COD_{Mn}, NH_3–N, and TP were all below 1. Among them, the highest COD_{Mn} pollution index was observed at the Tuqiao section (S6), with an average value of 0.92 ± 0.22. The Sanchahekou section (S5) exhibited the highest NH_3–N pollution index, with a mean value of 0.77 ± 0.49, while the Shitoucheng section (S4) showed the highest TP pollution index, averaging 0.80 ± 0.25.

Table 2: Mean values and standard deviation of single factor pollution index values.

Section	DO Mean ±SD	COD_{Mn} Mean ±SD	NH_3–N Mean ±SD	TP Mean ±SD
S1	0.63±1.19	0.32±0.17	0.09±0.15	0.40±0.11
S2	0.78±1.20	0.41±0.20	0.25±0.29	0.46±0.17
S3	1.19±1.51	0.47±0.16	0.38±0.31	0.51±0.13
S4	2.51±2.55	0.70±0.20	0.62±0.40	0.80±0.25
S5	2.48±2.61	0.62±0.18	0.77±0.49	0.76±0.26
S6	1.21±1.54	0.92±0.22	0.47±0.54	0.62±0.20
S7	1.12±1.42	0.84±0.21	0.39±0.27	0.53±0.19
S8	1.18±1.64	0.76±0.13	0.26±0.29	0.28±0.12
S9	0.81±1.03	0.73±0.20	0.40±0.43	0.51±0.23
S10	1.13±1.50	0.83±0.18	0.37±0.44	0.47±0.21

To more intuitively illustrate the spatial variation patterns of different water quality indicators, box plots and spatial distribution maps of water quality concentrations at each monitoring section were produced, as shown in Fig. (**3**) and Fig. (**4**). The results of the Kruskal–Wallis test indicated that significant differences in water quality indicator concentrations existed among the monitoring sections (P < 0.05). In the downstream Qinhuai River, from the Jiezhizha section (S1) to the Sanchahekou section (S5), the concentrations of the COD_{Mn}, NH_3–N, and TP exhibited an increasing trend with increasing water diversion distance, whereas DO show a decreasing trend. This pattern suggests that water diversion from the New Qinhuai River hydro-junction effectively improves water quality, although the improvement effect gradually weakens with increasing distance diversion. In the middle and

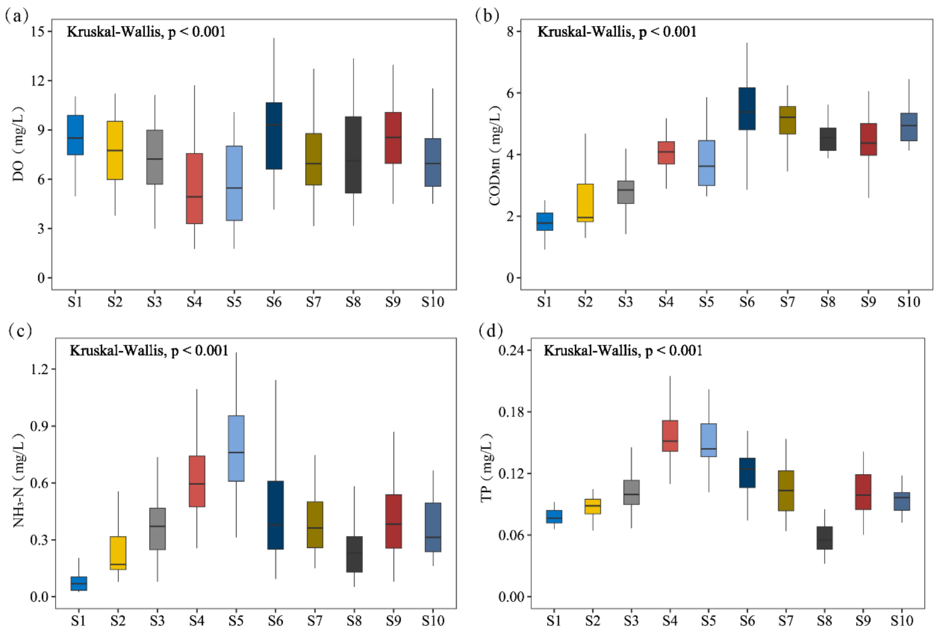


Figure 3: Box plots of water quality concentrations at different monitoring stations during 2021–2023.

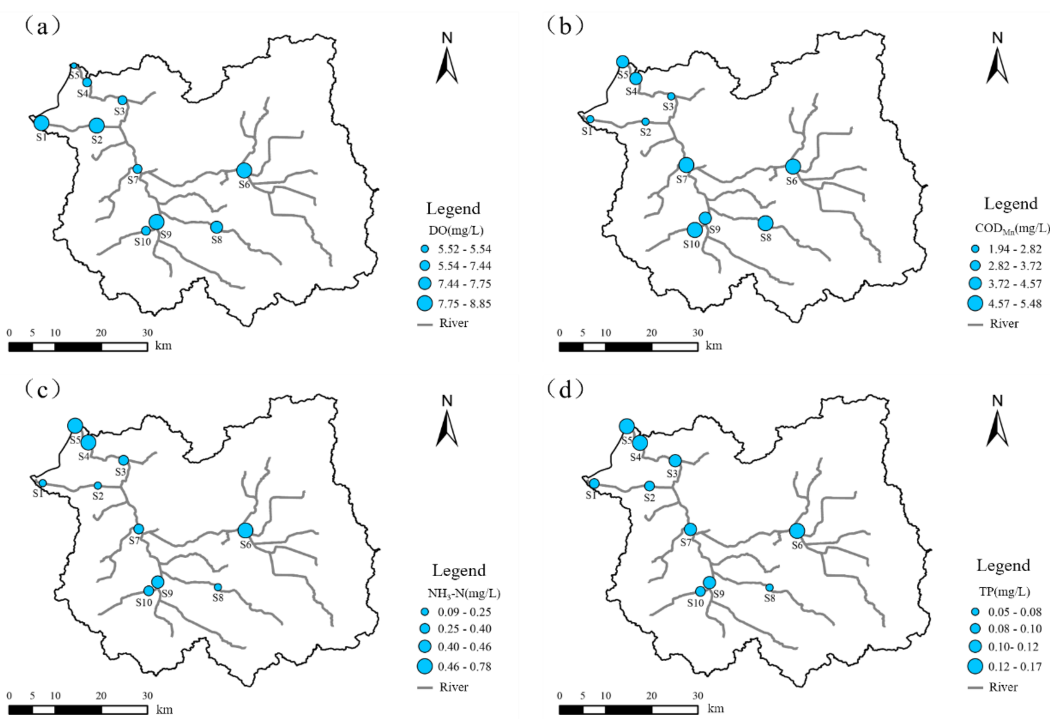


Figure 4: Spatial distribution of the mean concentrations of water quality indicators at different monitoring stations during 2021–2023.

upper reaches of the Qinhuai River Basin, the Kaitaiqiao section (S8), located in Lishui District, exhibited relatively good water quality, whereas the Tuqiao section (S6) in Jurong City showed comparatively poor water quality. The Yangqiao section (S7), which is located at the confluence of the Jurong River and the Lishui River, receives pollutants from upstream inflows; however, its overall water quality was better than that at the Tuqiao section. This improvement can be attributed to the self-purification and dilution effects of the river system, which reduces pollutant concentrations. Overall, DO and COD_{Mn} concentrations in the middle and upper reaches of the Qinhuai River Basin were higher than those in the downstream Qinhuai River, whereas NH_3–N and TP concentrations were lower.

3.2. Temporal Characteristics of Water Quality

To investigate the seasonal variation characteristics of water quality indicators, the year was divided into four seasons: spring (March–May), summer (June–August), autumn (September–November), and winter (December–February of the following year). After evaluating DO, COD_{Mn}, NH_3–N, and TP using the single-factor pollution index method, the exceedance rates of each indicator in different seasons were calculated, as shown in Fig. (**5**). DO exhibited higher exceedance rates in summer and autumn. Although the mean DO pollution indices at the Jiezhizha (S1), Tiexinqiao (S2), and Wuchaqiao (S9) sections generally met the standard (Table **1**), their exceedance rates in summer reached 25.72%, 45.29%, and 35.14%, respectively. In contrast, Sanchahekou (S5) and Shitoucheng (S4), which showed the most severe DO exceedance, exhibited summer exceedance rates as high as 93.88% and 98.21%, indicating that DO remained persistently below the standard during summer. For COD_{Mn},

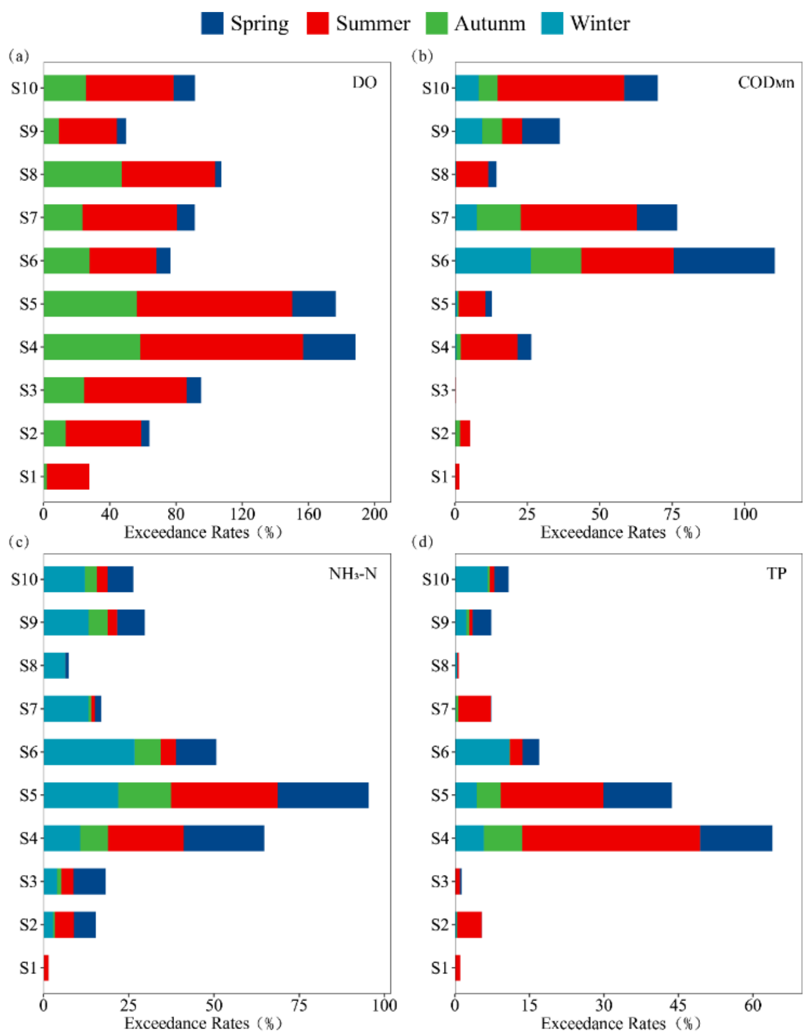


Figure 5: Exceedance rates of water quality indicators at different seasons from 2021 to 2023.

significant differences were observed between summer and both autumn and winter ($P < 0.001$). The Huangqiao (S10) and Yangqiao (S7) sections showed relatively high exceedance rates in summer, reaching 43.63% and 40.00%, respectively, followed by the Tuqiao (S6) section with a summer exceedance rate of 31.97%. The exceedance rates of NH_3–N at the Sanchahekou (S5) and Shitoucheng (S4) sections in summer were 31.35% and 22.13%, respectively. TP concentrations were also relatively higher in summer, and significant differences were detected between summer and the other three seasons ($P < 0.01$). In particular, the summer exceedance rates of TP at the Sanchahekou (S5) and Shitoucheng (S4) sections reached 20.72% and 35.84%, respectively. Overall, water quality exceedance rates at all monitoring sections increased during summer.

3.3. Machine Learning Modeling Results

After hyperparameter optimization, the modeling results for the COD_{Mn} and NH_3–N at the two national control sections, Qiqiaoweng (S3) and Yangqiao (S7), as well as the basin outlet section Sanchahekou (S5), are presented in Fig. (**6**). The results indicate that, for all water quality indicators, the EN model exhibited the poorest performance, suggesting its limited ability to capture nonlinear relationships in the data. The performance of the KNN model was slightly better than that of the SVM. In contrast, the RF and XGBoost models showed clear advantages in

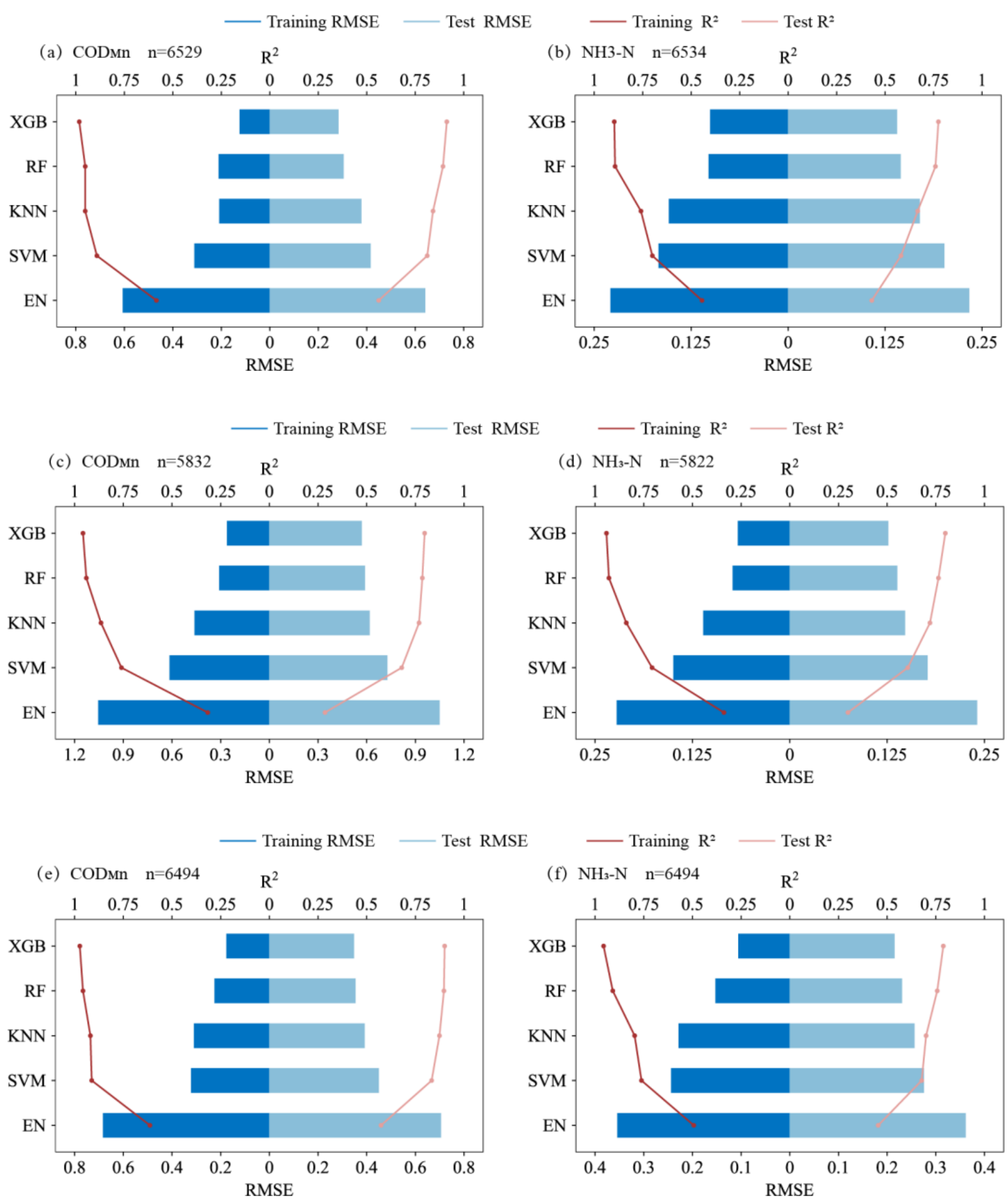


Figure 6: Performance of different models at Qiqiaoweng (**a**) and (**b**), Yangqiao (**c**) and (**d**), Sanchahekou (**e**) and (**f**) section.

water quality prediction. Compared with EN, KNN, and SVM, the ensemble learning models reduced the RMSE of the testing set by 29.11%–59.62% and increased the testing set R^2 by 36.14%–179.02%. The superior performance of RF and XGBoost can be attributed to their ensemble learning mechanisms: RF benefits from random feature selection and bootstrap sampling, while XGBoost is built on a gradient boosting framework and incorporates regularization terms. These characteristics enable ensemble models to effectively handle large and complex datasets, mitigate overfitting, and enhance generalization capability. A further comparison between RF and XGBoost revealed that XGBoost achieved slightly better predictive performance. Specifically, for COD_{Mn} prediction, the testing set R^2 of XGBoost ranged from 0.798 to 0.919, while for NH_3–N prediction, the testing set R^2 ranged from 0.732 to 0.824. Therefore, XGBoost was selected for subsequent analyses, including factor selection, model interpretation, and scenario simulation.

To determine the optimal lag time for each input variable, variable importance analyses were conducted for all water quality models. The ranking results of the input variables are presented in Table **S4**–**S6**, and for each variable, the lag time corresponding to the highest importance was selected as the input for the optimized model. After this selection, the number of input variables for the Qiqiaoweng (S3), Yangqiao (S7), and Sanchahekou (S5) models were 13, 11, and 11, respectively (Text **S4**). Based on the optimized input variables, the XGBoost models were retrained, and the results are shown in Fig. (**7**). Following variable selection, model performance improved further: for COD_{Mn} prediction, the testing set R^2 ranged from 0.882 to 0.947, while for NH_3–N prediction, the testing set R^2 ranged from 0.824 to 0.890. These results indicate that the models achieved satisfactory predictive performance and were able to effectively capture the inherent patterns in the data. The final optimal hyperparameters for all six prediction models are shown in Table **S7**.

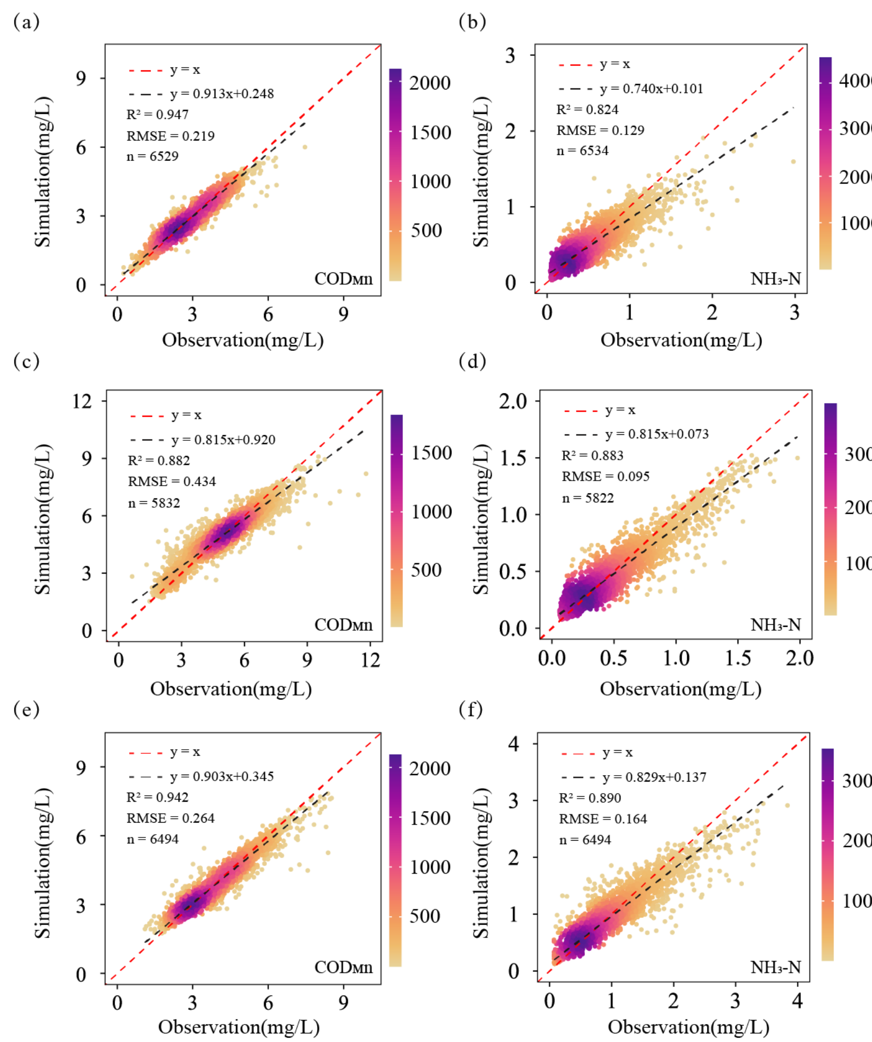


Figure 7: Optimal performance of model at Qiaoqiaoweng (**a**) and (**b**), Yangqiao (**c**) and (**d**), Sanchahekou (**e**) and (**f**) section.

3.4. Identification of Key Determinants of Water Quality

In this study, the optimal XGBoost models were interpreted to explore the effects of input variables on the predicted water quality. Variable importance was quantified using the mean SHAP values, reflecting the relative influence of each input variable on the model. Partial dependence plots (PDPs) were employed to illustrate the relationships between individual input variables and the output variables. The interpretation results for the water quality models at the three monitoring sections are shown in Fig. (**8**), where key variables were selected for PDP visualization, the abbreviations of the variables are listed in Table **S8**.

As shown in Fig. (**8**), the quality of upstream inflow water had a relatively strong influence on the water quality at the studied sections, and this influence was positively correlated. Specifically, for the COD_{Mn} and NH_3–N models at all three sections, upstream COD_{Mn} and NH_3–N were among the most critical variables, with the water quality at the target sections deteriorating as the upstream concentrations increased. The dominant influence of upstream water quality reflects the strong hydrological connectivity within the Qinhuai River network. Pollutants transported from upstream constitute the background water quality entering downstream reaches and continuously affect local water quality through advection and mixing processes. Because the residence time of pollutants in the river is relatively short compared with the transport time between monitoring stations, fluctuations in upstream CODMn and NH_3-N concentrations can rapidly propagate downstream, making upstream water quality one of the most influential predictors. This result also indicates that downstream water quality management cannot rely solely on local pollution control, but requires coordinated watershed-scale management to reduce pollutant inputs from upstream reaches.

The results also indicate that NH_3–N was more strongly affected by precipitation-runoff than COD_{Mn}, and the Sanchahekou (S5) section was more influenced by rainfall-runoff than the Qiqiaoweng (S3) and Yangqiao (S7) sections, suggesting that urban impervious surfaces exacerbate the negative effects of rainfall on water quality. Specifically, cumulative rainfall-runoff over the previous 48 h had a negative impact on river water quality. Its relative importance in the NH_3–N models at the Qiqiaoweng, Yangqiao, and Sanchahekou sections was 8.15%, 3.76%, and 23.35%, respectively. This finding suggests that the impacts of rainfall on nitrogen transport are characterized by a delayed response rather than an immediate effect. Rainfall first mobilizes nitrogen accumulated on urban impervious surfaces and within drainage systems, after which pollutants are gradually conveyed to the river network through stormwater pipelines and tributaries. The 48-h cumulative rainfall therefore better captures the integrated effects of pollutant mobilization, transport, and mixing than short-term rainfall. Moreover, Sanchahekou is located in a highly urbanized area with a high proportion of impervious surfaces, where rapid surface runoff generation reduces infiltration and enhances the delivery of nitrogen-rich runoff into receiving waters.

The inflow and outflow at the New Qinhuai River hydro-junction also significantly influenced water quality at the study sections. When the hydro-junction was in discharge mode (jjzDiv48 < 0), water quality at the Sanchahekou and Qiqiaoweng sections remained relatively stable but at high concentrations. Conversely, when the hydro-junction was in diversion mode (jjzDiv48 > 0), water quality concentrations at these sections decreased with increasing inflow. This finding suggests that Water diversion from the Qinhuai New River sluice also showed a significant influence on downstream water quality. When the sluice operated in diversion mode, increasing inflow introduced relatively cleaner water into the river network, thereby enhancing hydraulic exchange and increasing river discharge. This process promoted pollutant dilution, accelerated water renewal, and reduced pollutant residence time, ultimately decreasing COD_{Mn} and NH_3-N concentrations. In contrast, under drainage conditions, the river network experienced weaker hydraulic flushing, allowing pollutants to accumulate and resulting in relatively stable but elevated pollutant concentrations. These findings demonstrate that ecological water diversion can effectively improve water quality by strengthening hydrodynamic conditions in complex river networks.

Additionally, effluent from wastewater treatment plants influenced the water quality at the study sections to some extent, with discharge flow being more important than pollutant concentration. At the Sanchahekou and Qiqiaoweng sections, increases in the Kaifaqu WWTP effluent flow corresponded to higher COD_{Mn} and NH_3–N concentrations, while at the Yangqiao section, increases in the Lukou WWTP effluent flow led to higher NH_3–N

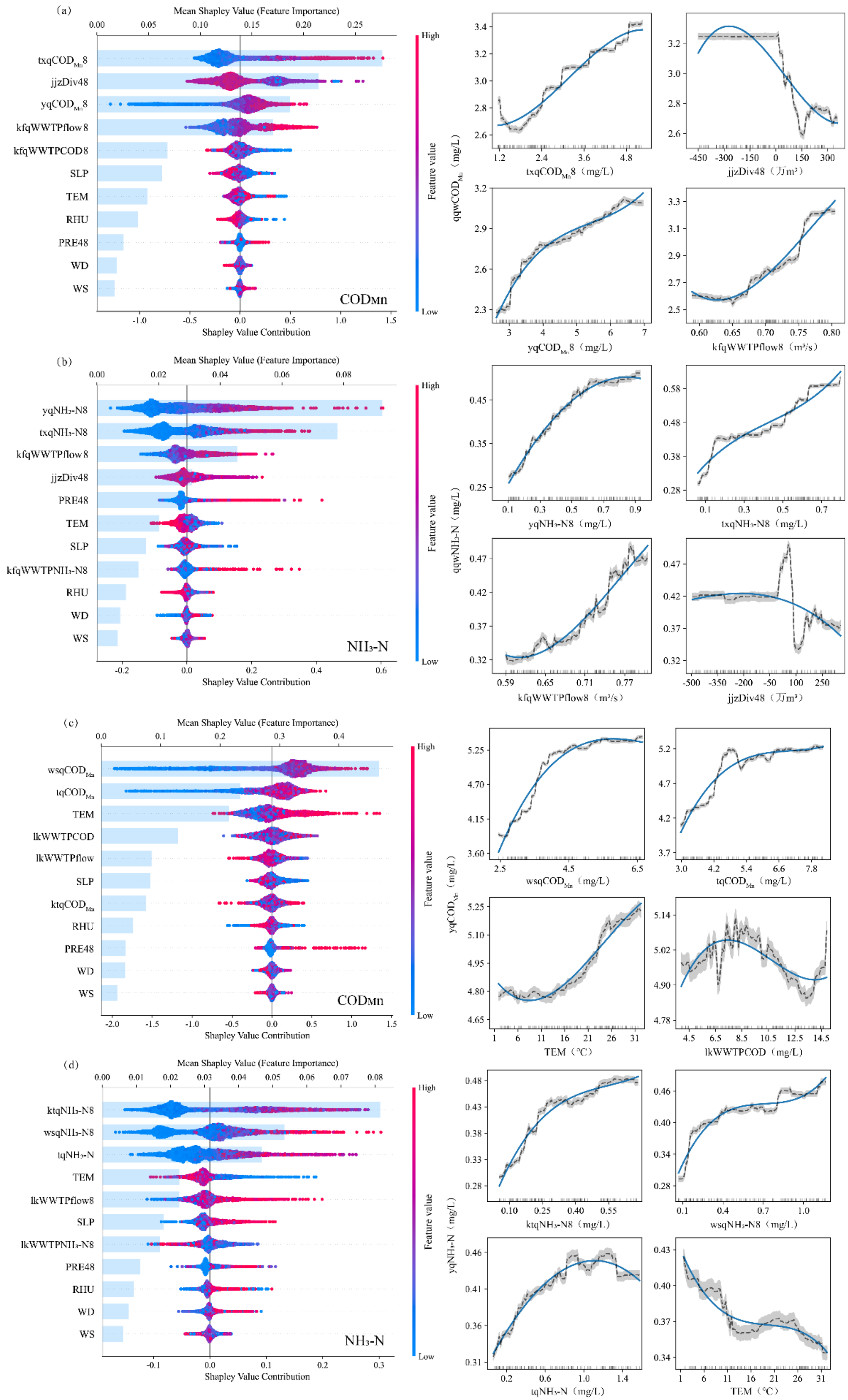

(a)
Mean Shapley Value (Feature Importance)
txqCODMn8
jjzDiv48
yqCODMn8
kfqWWTPflow8
kfqWWTPCOD8
SLP
TEM
RHU
PRE48
WD
WS
CODMn
Shapley Value Contribution
High
Feature value
Low
qqwCODMn (mg/L)
txqCODMn8 (mg/L)
jjzDiv48 (万m³)
yqCODMn8 (mg/L)
kfqWWTPflow8 (m³/s)
(b)
Mean Shapley Value (Feature Importance)
yqNH3-N8
txqNH3-N8
kfqWWTPflow8
jjzDiv48
PRE48
TEM
SLP
kfqWWTPNH3-N8
RHU
WD
WS
NH3-N
Shapley Value Contribution
qqwNH3-N (mg/L)
yqNH3-N8 (mg/L)
txqNH3-N8 (mg/L)
kfqWWTPflow8 (m³/s)
jjzDiv48 (万m³)
(c)
Mean Shapley Value (Feature Importance)
wsqCODMn
tqCODMn
TEM
lkWWTPCOD
lkWWTPflow
SLP
ktqCODMn
RHU
PRE48
WD
WS
CODMn
Shapley Value Contribution
yqCODMn (mg/L)
wsqCODMn (mg/L)
tqCODMn (mg/L)
TEM (℃)
lkWWTPCOD (mg/L)
(d)
Mean Shapley Value (Feature Importance)
ktqNH3-N8
wsqNH3-N8
tqNH3-N
TEM
lkWWTPflow8
SLP
lkWWTPNH3-N8
PRE48
RHU
WD
WS
NH3-N
Shapley Value Contribution
yqNH3-N (mg/L)
ktqNH3-N8 (mg/L)
wsqNH3-N8 (mg/L)
tqNH3-N (mg/L)
TEM (℃)

Figure 8: contd....

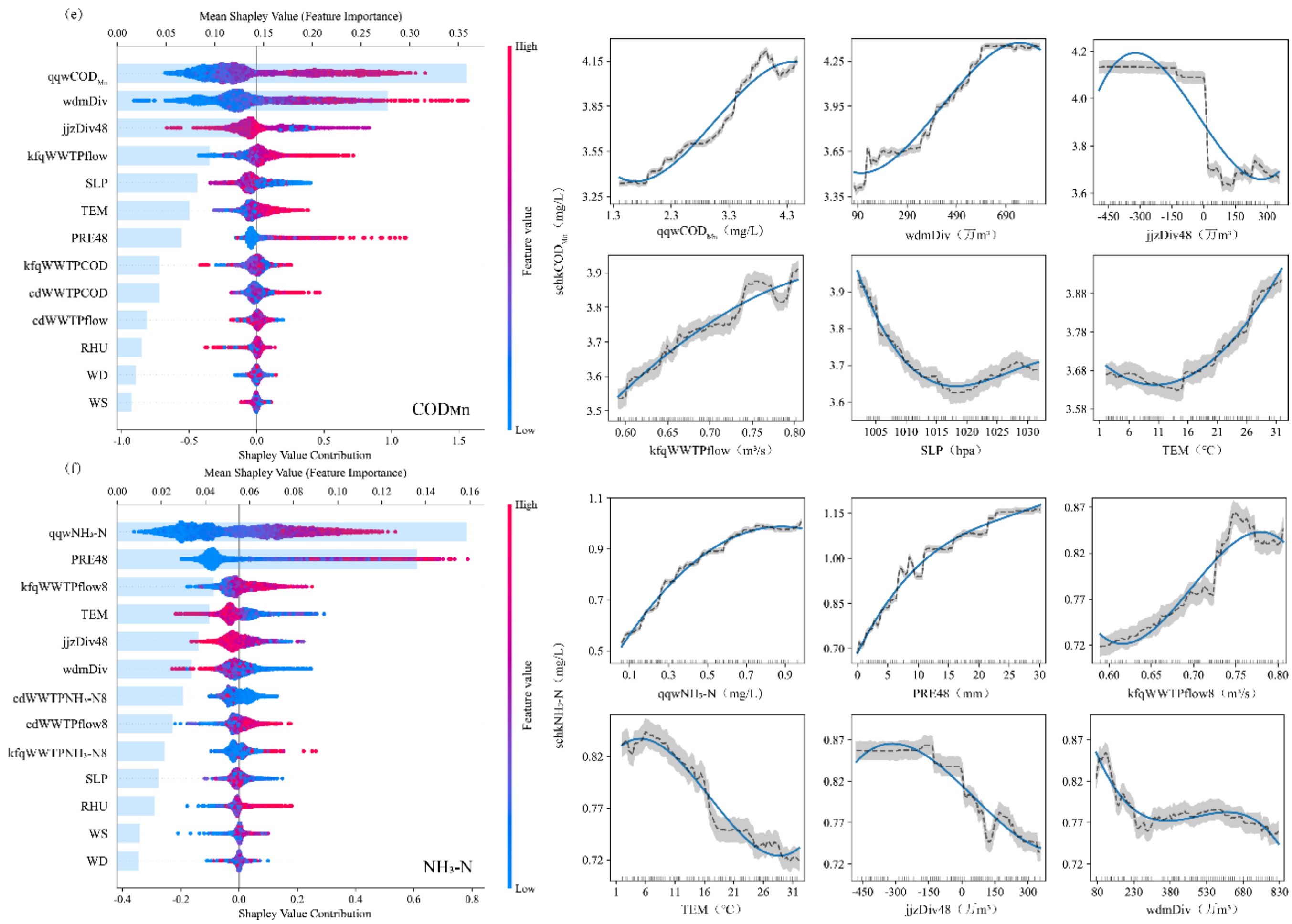


Figure 8: Variable importance analysis of the optimal model at Qiaoqiaoweng (**a**) and (**b**), Yangqiao (**c**) and (**d**), Sanchahekou (**e**) and (**f**) section.

concentrations. This finding indicates that pollutant loading, rather than concentration alone, is a more important determinant of downstream water quality. Even when effluent concentrations comply with discharge standards, increases in discharge volume substantially increase the total mass of pollutants entering the river, thereby elevating ambient pollutant concentrations. This effect is particularly evident during low-flow periods, when the river has limited dilution capacity and WWTP effluent constitutes a considerable proportion of streamflow. Therefore, effective water quality management should consider both effluent quality and discharge quantity, rather than focusing solely on concentration-based discharge standards.

3.5. Water Quality Prediction and Effectiveness Evaluation under Different Scenarios

The results of the scenario assessments for the monitoring sections at Qiqiaoweng, Yangqiao, and Sanchahekou are shown in Fig. (**9**). When the upstream water quality reaches Class III standards (S1), the reduction rates of COD_{Mn} and NH_3-N concentrations at Qiqiaoweng and Yangqiao are higher than those at Sanchahekou section, particularly at Yangqiao where NH_3-N concentration decreases from 1.191 mg/L to 0.965 mg/L, meeting Class III water standards. Rainfall runoff interception (S2) can effectively reduce the concentrations of water quality indicators, with reduction rates increasing as interception volume increases. Among these, the highest reduction rate in NH_3-N concentration is observed at Qiqiaoweng, ranging from 10.05% to 13.40%. Compared with the non-diversion scenario, increasing ecological water supplementation (S3) resulted in higher reduction rates of water quality indicator concentrations. When the diversion volume reaches 2.0 × 10^6 m^3 (S3_3), the NH_3-N concentration at Qiqiaoweng reaches 0.964 mg/L, meeting Class III standards, with a reduction rate of 21.76%.

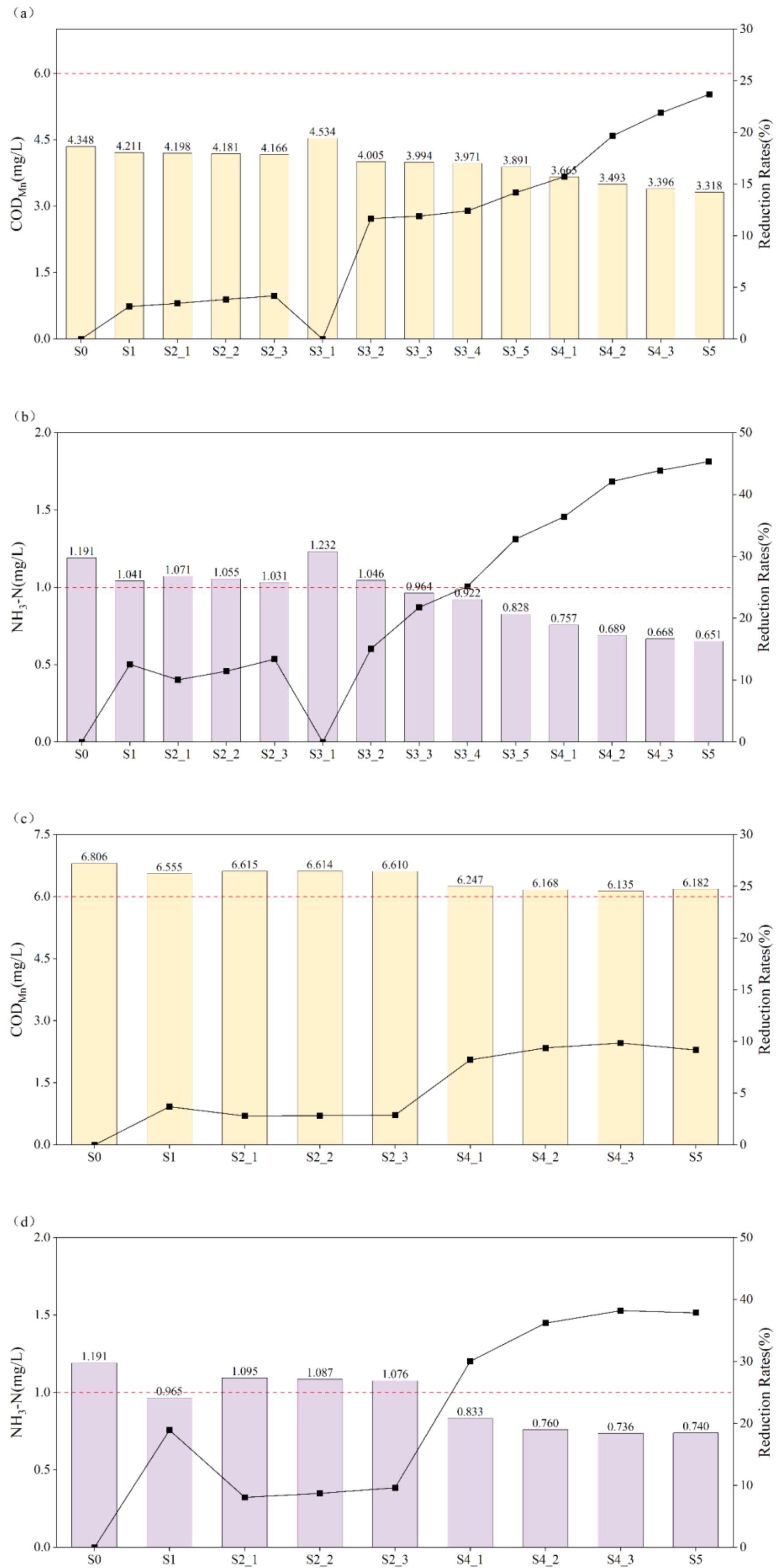

(a)
COD_{Mn}(mg/L)
Reduction Rates(%)
4.348
4.211
4.198
4.181
4.166
4.534
4.005
3.994
3.971
3.891
3.665
3.493
3.396
3.318
S0
S1
S2_1
S2_2
S2_3
S3_1
S3_2
S3_3
S3_4
S3_5
S4_1
S4_2
S4_3
S5
(b)
NH_3-N(mg/L)
Reduction Rates(%)
1.191
1.041
1.071
1.055
1.031
1.232
1.046
0.964
0.922
0.828
0.757
0.689
0.668
0.651
(c)
COD_{Mn}(mg/L)
Reduction Rates(%)
6.806
6.555
6.615
6.614
6.610
6.247
6.168
6.135
6.182
(d)
NH_3-N(mg/L)
Reduction Rates(%)
1.191
0.965
1.095
1.087
1.076
0.833
0.760
0.736
0.740

Figure 9: contd….

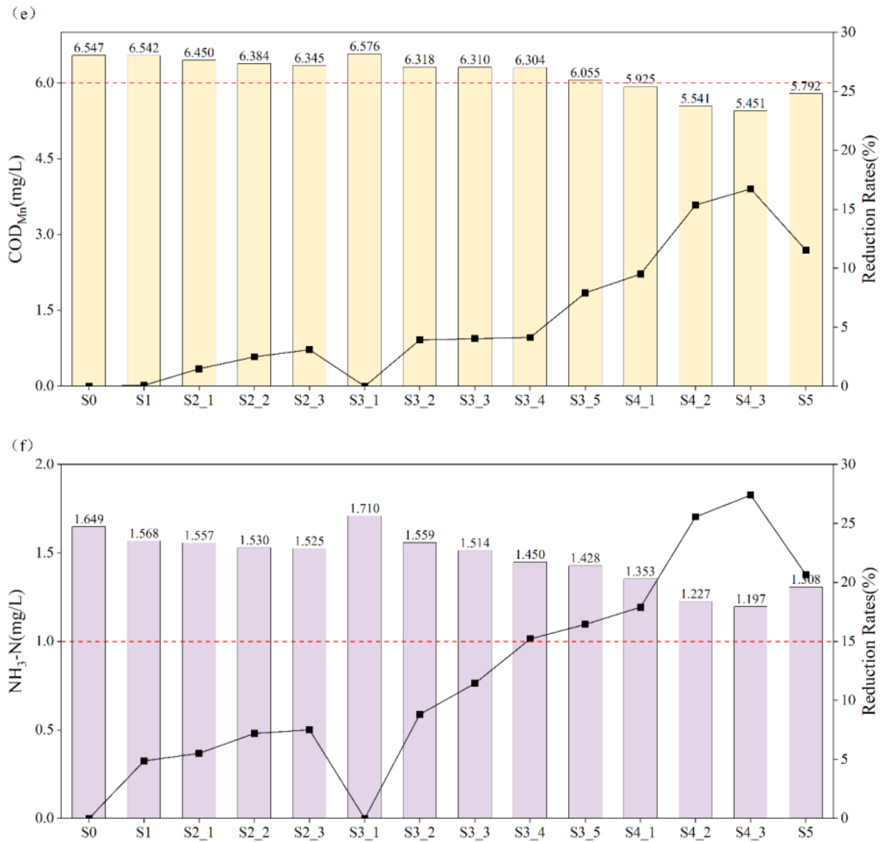


Figure 9: Evaluation of water quality improvement under different scenarios at Qiqiaoweng (**a**) and (**b**), Yangqiao (**c**) and (**d**), Sanchahekou (**e**) and (**f**) section. (The dashed line represents the Class III water quality standard, the solid line represents the reduction rate, and the bars represent the pollutant concentration).

Moreover, under the same diversion conditions at the Jiezhizha section, the improvement effect for COD_{Mn} and NH_3-N at Qiqiaoweng section is better than at Sanchahekou section, which is related to the distance from the water diversion point. The reduction of effluent flow from wastewater treatment plants has the most significant impact on water quality (S4). When the effluent flow reduction rate is 10% (S4_1), the improvement effect on water quality indicators at the study sections is higher than in the scenario where green area is increased by 30% (S2_3), indicating that controlling the effluent discharge from wastewater treatment plants is an effective measure. In the comprehensive scenario (S5), the COD_{Mn} at Sanchahekou, NH_3-N concentration at Qiqiaoweng, and NH_3-N concentration at Yangqiao all meet Class III standards, with the reduction rates of water quality indicator concentrations at the three sections ranging from 8.52% to 45.31%. However, it was also observed that although the COD_{Mn} concentration at Yangqiao and the NH_3-N concentration at Sanchahekou showed some improvement under various scenarios, they could not meet the standards in the current hypothetical scenarios. This suggests that achieving water quality standards under the existing scenarios is unlikely in the near future. Further simulations revealed that a reduction of more than 70% in upstream pollutant concentrations could enable the standards to be met, indicating that the water quality of downstream monitoring points is primarily dependent on pollutant reduction in upstream water. This underscores the significant impact of upstream water quality on downstream water quality. Therefore, future regulatory measures should focus on addressing these issues.

3.6. Water Quality Management Strategies

The scenario simulations provide quantitative evidence for prioritizing water quality management strategies in the Qinhuai River Basin. Among the individual scenarios, upstream pollution control (S1) and WWTP effluent reduction (S4) showed greater potential for improving downstream water quality, while rainfall–runoff control (S2) and ecological water supplementation (S3) provided complementary benefits. Therefore, an integrated management framework involving source control, non-point source reduction, hydrodynamic regulation, and wastewater management is recommended.

I) Prioritize upstream pollution reduction through source control. The S1 scenario demonstrated that improving upstream inflow quality to Class III water quality standards could substantially reduce downstream COD_{Mn} and NH_3-N concentrations, particularly at Qiqiaoweng and Yangqiao stations. This indicates that pollutant transport from upstream reaches is a dominant pathway controlling downstream water quality. Therefore, upstream pollution reduction should be considered the highest priority strategy. Environmental authorities should strengthen supervision of upstream industrial and domestic pollution sources, while enterprises should improve wastewater treatment efficiency and ensure stable compliance discharge. Watershed-scale coordination mechanisms should also be established to prevent upstream pollution transfer to downstream areas.

II) Strengthen urban rainfall–runoff control. The S2 scenario showed that rainfall–runoff interception could effectively reduce pollutant concentrations, with NH_3-N reduction rates at Qiqiaoweng reaching 10.05%–13.40%. These results indicate that urban non-point source pollution is an important contributor to nitrogen loading, especially during rainfall events. Therefore, municipal governments should prioritize stormwater management strategies, including low-impact development (LID), permeable pavement, rain gardens, green roofs, and ecological drainage systems. These measures can reduce pollutant accumulation on impervious surfaces and mitigate short-term pollution pulses caused by storm runoff.

III) Optimize ecological water supplementation. The S3 scenario indicated that increasing ecological water supplementation could continuously improve water quality by enhancing dilution and hydraulic exchange. For example, when the diversion volume reached 2 million m^3 (S3_3), NH_3-N concentration at Qiqiaoweng decreased to 0.964 mg/L and met the Class III standard, with a reduction rate of 21.76%. However, ecological water supplementation should be regarded as a regulatory measure rather than a replacement for pollution control. River management authorities should optimize diversion timing and volume according to hydrological conditions and ecological requirements to maximize water quality benefits while minimizing water consumption.

IV) Improve WWTP effluent management. Among individual management scenarios, WWTP effluent reduction (S4) showed the strongest improvement potential, indicating that controlling point-source pollutant loads is highly effective for improving river water quality. The results suggest that reducing effluent discharge volume and pollutant loads should be prioritized. Wastewater treatment plants should further improve treatment efficiency and promote reclaimed water utilization for urban greening, road cleaning, industrial processes, and agricultural irrigation. Such measures can reduce direct discharge into the river network and improve water resource utilization.

V) Implement integrated management strategies. The combined scenario (S5) demonstrated that coordinated implementation of multiple measures achieved the greatest water quality improvement, with pollutant concentration reduction rates ranging from 8.52% to 45.31%. Therefore, long-term management of the Qinhuai River Basin should adopt an integrated strategy: upstream pollution control as the primary measure, supported by urban runoff management, ecological water supplementation, and WWTP effluent optimization

3.7. Model Uncertainty and Limitations

Although the proposed framework achieved satisfactory predictive performance, several limitations and uncertainties should be acknowledged. First, the relatively short data period (2021–2023) represents an important limitation of this study. The three-year dataset may not fully capture long-term climate variability, hydrological extremes, or gradual changes in watershed characteristics and pollution sources. Therefore, the relationships identified between environmental factors and water quality, as well as the scenario simulation results, may be influenced by interannual variability and should be interpreted with caution when extending the findings to long-term water quality management. Nevertheless, the high-frequency monitoring data used in this study covered different seasons and hydrological conditions, providing valuable information for model development and evaluation within the study period.

Beyond the limitations associated with data availability, uncertainties related to the modeling process should also be considered. First, uncertainties may arise from input data quality. The water quality observations, meteorological variables, and hydrological information used for model development inevitably contain measurement errors and uncertainties, which may propagate through the modeling process and influence predicted concentrations and scenario evaluation results. Second, machine learning models mainly establish statistical relationships between input variables and water quality responses based on historical observations. Although this data-driven approach provides strong nonlinear mapping capability, it may not fully represent all complex physical, hydrological, and biogeochemical processes governing water quality variations in urban river networks. Therefore, prediction uncertainty may increase when the model is applied under environmental conditions beyond the range of historical observations. Third, uncertainties associated with variable selection and model parameterization should also be considered, as the selection of input variables and optimization of XGBoost hyperparameters may affect model performance and prediction stability. In this study, variable importance analysis and SHAP interpretation were applied to identify key influencing factors and improve model interpretability. Moreover, the consistent predictive performance among different monitoring stations indicates that the proposed framework provides reliable predictions under the studied conditions.

Future studies should incorporate longer-term monitoring datasets, future climate scenarios, process-based hydrological information, and advanced uncertainty quantification approaches to further improve the robustness, transferability, and applicability of machine learning-based water quality prediction and management frameworks.

4. Conclusion

This study focuses on the Qinhuai River basin, a typical tributary of the Yangtze River, from the perspective of watershed water environment management. Utilizing long-term water quality data, the study first analyzes and evaluates the spatiotemporal distribution characteristics of water quality in the basin. Then, a machine learning model for complex river networks is developed to identify key factors influencing water quality at monitoring sites. Finally, based on the actual situation of the basin, scenario simulations are conducted, and corresponding management strategies are proposed. The main conclusions are as follows:

(1) Spatiotemporal characteristics: DO is the main factor exceeding the standard. The concentrations of COD_{Mn}, NH_3-N, and TP are generally within the standard range, but exhibit seasonal fluctuations, with summer being the peak season for water quality violations. In the upstream and midstream areas of the basin, the concentrations of DO and COD_{Mn} are higher than those in the downstream sections of the lower reaches of Qinhuai River, while the concentrations of NH_3-N and TP are lower.

(2) Model performance: Compared with EN, KNN, SVM, and RF, XGBoost demonstrates a significant advantage. After variable selection, the XGBoost model achieves a test R^2 of 0.882–0.947 for predicting COD_{Mn} and 0.824–0.890 for predicting NH_3-N. The key factors identified as influencing water quality include upstream water quality, the cumulative rainfall runoff over the past 48 h, the water intake and discharge volumes in the basin, and the effluent discharge from wastewater treatment plants.

(3) Scenario simulations demonstrated that improving upstream inflow quality to the Class III water quality standard could effectively improve downstream water quality. For example, the NH_3-N concentration at Yangqiao decreased from 1.191 to 0.965 mg/L and reached the Class III standard. Rainfall–runoff interception effectively reduced water quality indicator concentrations, with NH_3-N reduction rates at Qiqiaoweng ranging from 10.05% to 13.40%. The effectiveness of ecological water supplementation was dependent on diversion volume, and a diversion volume of 2 million m^3 reduced NH_3-N concentration at Qiqiaoweng by 21.76% (to 0.964 mg/L), achieving the Class III standard. WWTP effluent reduction also showed considerable potential for water quality improvement. Under the integrated management scenario, COD_{Mn} at Sanchahekou and NH_3-N at Qiqiaoweng and Yangqiao reached the Class III standard, with pollutant concentration reduction rates ranging from 8.52% to 45.31% across the three monitoring stations.

These results highlight that the proposed machine learning-based framework, integrating water quality prediction, driving factor identification, and scenario simulation, can serve as a transferable decision-support tool for water managers in other complex river networks, especially those with limited monitoring data and dynamic hydrological conditions.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This study was supported by the National Key R&D Program of China (grant no. 2022YFC3204400).

References

[1] Vega M, Pardo R, Barrado E, Debán L. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Res. 1998; 32(12): 3581-92. https://doi.org/10.1016/S0043-1354(98)00138-9

[2] Grill G, Lehner B, Thieme M, Geenen B, Tickner D, Antonelli F, *et al*. Mapping the world's free-flowing rivers. Nature. 2019; 569(7755): 215-21. https://doi.org/10.1038/s41586-019-1111-9

[3] Palmer M, Ruhi A. Linkages between flow regime, biota, and ecosystem processes: Implications for river restoration. Science. 2019; 365(6459): eaaw2087. https://doi.org/10.1126/science.aaw2087

[4] Chen SK, Jang CS, Chou CY. Assessment of spatiotemporal variations in river water quality for sustainable environmental and recreational management in the highly urbanized Danshui River basin. Environ Monit Assess. 2019; 191(2): 100. https://doi.org/10.1007/s10661-019-7246-1

[5] Strokal M, Ma L, Bai Z, Luan S, Kroeze C, Oenema O, *et al*. Alarming nutrient pollution of Chinese rivers as a result of agricultural transitions. Environ Res Lett. 2016; 11(2): 024014. https://doi.org/10.1088/1748-9326/11/2/024014

[6] Xiao Q, Hu Z, Hu C, Islam ARMT, Bian H, Chen S, *et al*. A highly agricultural river network in Jurong Reservoir watershed as significant CO_2 and CH_4 sources. Sci Total Environ. 2021; 769: 144558. https://doi.org/10.1016/j.scitotenv.2020.144558

[7] Kim J, Lee T, Seo D. Algal bloom prediction of the lower Han River, Korea using the EFDC hydrodynamic and water quality model. Ecol Model. 2017; 366: 27-36. https://doi.org/10.1016/j.ecolmodel.2017.10.015

[8] Shin S, Her Y, Muñoz-Carpena R, Yu X, Martinez C, Singh A. Climate change impacts on water quantity and quality of a watershed-lake system using a spatially integrated modeling framework in the Kissimmee River-Lake Okeechobee system. J Hydrol Reg Stud. 2023; 47: 101408. https://doi.org/10.1016/j.ejrh.2023.101408

[9] Yu J, Tian Y, Wang X, Sun T, Lancia M, Andrews CB, *et al*. Integrated modeling of flow, soil erosion, and nutrient dynamics in a regional watershed: Assessing natural and human-induced impacts. Water Resour Res. 2024; 60(9): e2024WR037531. https://doi.org/10.1029/2024WR037531

[10] Peng K, Li J, Zhou X, Li H, Xie W, Zhang K, *et al*. Simulation and control of non-point source pollution based on MIKE model: A case study of Danjiang River Basin, China. Ecohydrol Hydrobiol. 2023; 23(4): 554-68. https://doi.org/10.1016/j.ecohyd.2023.08.001

[11] Jiang L, Li Y, Zhao X, Tillotson MR, Wang W, Zhang S, *et al*. Parameter uncertainty and sensitivity analysis of water quality model in Lake Taihu, China. Ecol Model. 2018; 375: 1-12. https://doi.org/10.1016/j.ecolmodel.2018.02.014

[12] Li S, Feng J, Zhang S. Simulation analysis of water quality improvement effect of ecological water replenishment measures in Qilu Lake (In Chinese). Water Resour Power. 2020; 38(11): 35-9.

[13] Zanoni MG, Majone B, Bellin A. A catchment-scale model of river water quality by machine learning. Sci Total Environ. 2022; 838: 156377. https://doi.org/10.1016/j.scitotenv.2022.156377

[14] Gao Z, Chen J, Wang G, Ren S, Fang L, Yinglan A, *et al*. A novel multivariate time series prediction of crucial water quality parameters with long short-term memory (LSTM) networks. J Contam Hydrol. 2023; 259: 104262. https://doi.org/10.1016/j.jconhyd.2023.104262

[15] Liang Z, Zou R, Chen X, Ren T, Su H, Liu Y. Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach. J Hydrol. 2020; 581: 124432. https://doi.org/10.1016/j.jhydrol.2019.124432

[16] Hong SM, Abbas A, Kim S, Kwon DH, Yoon N, Yun D, *et al*. Autonomous calibration of EFDC for predicting chlorophyll-a using reinforcement learning and a real-time monitoring system. Environ Model Softw. 2023; 168: 105805. https://doi.org/10.1016/j.envsoft.2023.105805

[17] Keller AA, Garner K, Rao N, Knipping E, Thomas J. Hydrological models for climate-based assessments at the watershed scale: A critical review of existing hydrologic and water quality models. Sci Total Environ. 2023; 867: 161209. https://doi.org/10.1016/j.scitotenv.2022.161209

[18] Zhong S, Zhang K, Bagheri M, Burken JG, Gu A, Li B, *et al*. Machine learning: New ideas and tools in environmental science and engineering. Environ Sci Technol. 2021; 55(19): 12741-54. https://doi.org/10.1021/acs.est.1c01339

[19] Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, *et al*. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Res. 2020; 171: 115454. https://doi.org/10.1016/j.watres.2019.115454

[20] Kumar N, Sinha A, Pasupuleti S. Application of machine learning models for water quality prediction in a coal-mining region: Insights into fluoride and phenol contamination in Dhanbad District, Eastern India. Environ Pollut. 2026; 405: 128558. https://doi.org/10.1016/j.envpol.2026.128558

[21] Tong X, Zhang X, Xiao S, Chen J, Zhou C, Zhou J, *et al*. A machine learning-assisted integrated framework for linking water quality and phytoplankton carbon fixation potential in drinking water reservoirs. Water Res. 2026; 302: 126104. https://doi.org/10.1016/j.watres.2026.126104

[22] Yu L, Liu X, Hua Z. Occurrence, distribution, and risk assessment of perfluoroalkyl acids in drinking water sources from the lower Yangtze River. Chemosphere. 2022; 287: 132064. https://doi.org/10.1016/j.chemosphere.2021.132064

[23] Xu Z. Analysis of key elements in water environment governance of Yangtze River (In Chinese). Water Resour Dev Res. 1-13.

[24] Ma Y, Chen X, Zhang B. Screening of characteristic pollutants and risk pollutants in different media in the Yangtze River Basin using comprehensive evaluation and risk quotient methods (In Chinese). Environ Sci. 2024; 45(9): 5214-26.

[25] Zou H, Li J, Cai Y, Wang X, Ma T, Duan X. Characterizing the spatial variations of the relationship between chemical industry and riparian water quality along the Yangtze River. Appl Geogr. 2026; 194: 104117. https://doi.org/10.1016/j.apgeog.2026.104117

[26] Zhang C, Li K, Fang Q, Zhang S. Construction and application of an aquatic ecological health assessment system based on multi-source data and multi-method integration: A case study of key water bodies in the Yangtze River. Ecol Indic. 2026; 188: 115038. https://doi.org/10.1016/j.ecolind.2026.115038

[27] Wang S, Pu Y, Li S, Li R, Li M. Spatio-temporal analysis of impervious surface expansion in the Qinhuai River Basin, China, 1988-2017. Remote Sens (Basel). 2021; 13(22): 4494. https://doi.org/10.3390/rs13224494

[28] Du J, Qian L, Rui H, Zuo T, Zheng D, Xu Y, *et al*. Assessing the effects of urbanization on annual runoff and flood events using an integrated hydrological modeling system for Qinhuai River Basin, China. J Hydrol. 2012; 464-465: 127-39. https://doi.org/10.1016/j.jhydrol.2012.06.057

[29] Li J, Li Y, Qian B, Niu L, Zhang W, Cai W, *et al*. Development and validation of a bacteria-based index of biotic integrity for assessing the ecological status of urban rivers: A case study of Qinhuai River Basin in Nanjing, China. J Environ Manage. 2017; 196: 161-7. https://doi.org/10.1016/j.jenvman.2017.03.003

[30] Gao Y, Chen J, Luo H, Wang H. Prediction of hydrological responses to land use change. Sci Total Environ. 2020; 708: 134998. https://doi.org/10.1016/j.scitotenv.2019.134998

[31] Liu M, Wang Y, Liu Y. Analysis of uneven distribution of precipitation during flood season in the Qinhuai River Basin (In Chinese). Water Resour Dev Manag. 2024; 10(5): 65-9.

[32] Bian G, Wang G, Chen J, Zhang J, Song M. Spatial and seasonal variations of hydrological responses to climate and land-use changes in a highly urbanized basin of Southeastern China. Hydrol Res. 2021; 52(2): 506-22. https://doi.org/10.2166/nh.2021.087

[33] Li R, Du J, Bian G, Wang Y, Chen C, Zhang X, *et al*. An integrated modelling approach for flood simulation in the urbanized Qinhuai River Basin, China. Water Resour Manag. 2020; 34(13): 3967-84. https://doi.org/10.1007/s11269-020-02614-4

[34] Abro MI, Zhu D, Wei M, Majidano AA, Khaskheli MA, Abideen ZU, *et al*. Hydrological appraisal of rainfall estimates from radar, satellite, raingauge and satellite-gauge combination on the Qinhuai River Basin, China. Hydrol Sci J. 2019; 64(16): 1957-71. https://doi.org/10.1080/02626667.2018.1557335

[35] Chen R, Hu Q, Shen W, Guo J, Yang L, Yuan Q, *et al*. Identification of nitrate sources of groundwater and rivers in complex urban environments based on isotopic and hydro-chemical evidence. Sci Total Environ. 2023; 871: 162026. https://doi.org/10.1016/j.scitotenv.2023.162026

[36] Xia Y, Ti C, She D, Yan X. Linking river nutrient concentrations to land use and rainfall in a paddy agriculture-urban area gradient watershed in southeast China. Sci Total Environ. 2016; 566-567: 1094-105. https://doi.org/10.1016/j.scitotenv.2016.05.134

[37] Yan X, Han H, Qiu J, Zhang L, Xia Y, Yan X. Suburban agriculture increased N levels but decreased indirect N_2O emissions in an agricultural-urban gradient river. Water Res. 2022; 220: 118639. https://doi.org/10.1016/j.watres.2022.118639

[38] Sang C, Wang J, Guo Y, Xie C, Zhou Q. Integrated evaluation of water quality in semi-arid Eurasian rivers using the methods of single-factor and water quality index. Ecol Indic. 2026; 183: 114674. https://doi.org/10.1016/j.ecolind.2026.114674

[39] Su K, Wang Q, Li L, Cao R, Xi Y. Water quality assessment of Lugu Lake based on Nemerow pollution index method. Sci Rep. 2022; 12(1): 13613. https://doi.org/10.1038/s41598-022-17874-w

[40] State Environmental Protection Administration. Environmental quality standards for surface water (PR China) (GB 3838-2002) [S]. Beijing: China Environmental Science Press; 2002.

[41] Wang Q, Li Z, Cai J, Zhang M, Liu Z, Xu Y, *et al*. Spatially adaptive machine learning models for predicting water quality in Hong Kong. J Hydrol. 2023; 622: 129649. https://doi.org/10.1016/j.jhydrol.2023.129649

[42] Zhang Q, Li Z, Zhu L, Zhang F, Sekerinski E, Han JC, *et al*. Real-time prediction of river chloride concentration using ensemble learning. Environ Pollut. 2021; 291: 118116. https://doi.org/10.1016/j.envpol.2021.118116

[43] Qiu R, Wang Y, Rhoads B, Wang D, Qiu W, Tao Y, *et al*. River water temperature forecasting using a deep learning method. J Hydrol. 2021;

595: 126016. https://doi.org/10.1016/j.jhydrol.2021.126016

[44] Ai H, Zhang K, Sun J, Zhang H. Short-term Lake Erie algal bloom prediction by classification and regression models. Water Res. 2023; 232: 119710. https://doi.org/10.1016/j.watres.2023.119710

[45] Wang R, Kim JH, Li MH. Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. Sci Total Environ. 2021; 761: 144057. https://doi.org/10.1016/j.scitotenv.2020.144057

[46] Li Z, Gu H, Kang Y, Zhang C, Niu Z, Shen D. Spatiotemporal assessment of water damage susceptibility in China's road infrastructure: A machine learning and SHAP approach using social media data. J Hydrol. 2026; 664: 134539. https://doi.org/10.1016/j.jhydrol.2025.134539

[47] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001; 29: 1189-232. https://doi.org/10.1214/aos/1013203451

TextS1 Water Quality Assessment Method

The calculation formulas for the single-factor pollution indices of COD_{Mn}, NH_3–N, and TP are as follows [1, 2]:

$$P_i = \frac{C_i}{C_{io}} \tag{1}$$

Where P_i is the pollution index of the water quality parameter, C_i is the measured concentration of the ith water quality parameter (mg/L), and C_{i0} is the standard limit of the ith parameter (mg/L). When $P_i \leq 1$, the water quality parameter meets the standard; when $P_i > 1$, the parameter exceeds the standard, and larger values of P_i indicate more severe pollution.

The calculation formula for the single-factor pollution index of DO is as follows.

$$P_{DO} = \frac{\left|DO_f - DO\right|}{DO_f - DO_o}(DO \geq DO_o) \tag{2}$$

$$P_{DO} = 10 - 9\frac{DO}{DO_o}(DO < DO_o) \tag{3}$$

$$DO_f = \frac{468}{31.6 + T} \tag{4}$$

Where DO is the measured concentration of dissolved oxygen (mg/L), DO_0 is the standard limit of dissolved oxygen (mg/L), DO_f is the saturated dissolved oxygen concentration (mg/L), and T is the water temperature (°C).

The calculation formula for the exceedance rate is as follows.

$$\text{Exceedance rate} = \frac{\text{Number of Exceedances}}{Total\ Number\ of\ Measurements} \times 100\% \tag{5}$$

TextS2 Model Description

In brief, Elastic Net (EN) is a linear regression algorithm that incorporates regularization into traditional linear models to prevent overfitting [3]. K-Nearest Neighbors (KNN) is a distance-based, non-parametric supervised learning method that is simple and intuitive and can be applied to both classification and regression problems [4]. Support Vector Machine (SVM) is a supervised learning algorithm that aims to construct a maximum-margin hyperplane, where the position of the hyperplane is determined by the support vectors [5]. Random Forest (RF) is an ensemble learning algorithm [6] that employs the bagging technique to randomly sample subsets of data with replacement, trains a decision tree on each subset, and generates the final prediction by averaging the outputs of all trees. This approach improves generalization and robustness, overcoming the overfitting and instability issues of individual decision trees [7]. Extreme Gradient Boosting (XGBoost) predicts the residuals of preceding decision trees by sequentially adding new trees and combines them to maximize prediction accuracy. As an improved version of Gradient Boosting Decision Trees (GBDT) [8], XGBoost introduces regularization into the objective function to reduce overfitting risk and accelerate convergence, while maintaining strong dependencies among the trees [9] .

TextS3 Hyperparameters of Model

Hyperparameter tuning was performed for each model as follows. For Elastic Net (EN), two hyperparameters were adjusted: λ, the regularization parameter, and α, the mixing proportion of the regularization terms. For K-Nearest Neighbors (KNN), the hyperparameter k, representing the number of neighbors, was tuned. For Support Vector Machine (SVM), three hyperparameters were optimized: degree, cost, and gamma, where degree is the polynomial kernel degree, cost is the penalty parameter, and gamma is the kernel coefficient that controls the shape of the kernel function. For Random Forest (RF), four hyperparameters were adjusted: ntree, nodesize, maxnodes, and mtry, where ntree is the number of trees in the forest, nodesize is the minimum number of samples per leaf node, maxnodes is the maximum number of nodes in each tree, and mtry is the number of features randomly selected for each split. For Extreme Gradient Boosting (XGBoost), eight hyperparameters were tuned: eta, gamma, max_depth, min_child_weight, subsample, colsample_bytree, nrounds, and lambda. Here, eta is the learning rate, gamma is the minimum loss reduction required to make a split, max_depth is the maximum depth of a tree, min_child_weight is the minimum sum of instance weights in a leaf node, subsample is the fraction of samples randomly selected for training each tree, colsample_bytree is the fraction of features randomly selected for training each tree, nrounds is the number of sequential trees constructed in the model, and lambda is the L2 regularization coefficient.

TextS4 Refined Variable Selection

At the Qiqiaoweng (S3) section, the input variables for the COD_{Mn} model included the COD_{Mn} concentrations 8 hours prior at the Yangqiao (S7) and Tiexinqiao (S2) sections, the New Qinhuai River hydro-junction inflow 48 hours prior, the COD and effluent flow 8 hours prior from the Kaifaqu WWTP, the cumulative rainfall runoff over the previous 48 hours, and meteorological factors including sea level pressure, wind direction, wind speed, relative humidity, and air temperature. For the NH_3–N model, the input variables included the NH_3–N concentrations 8 hours prior at the Yangqiao (S7) and Tiexinqiao (S2) sections, the New Qinhuai River hydro-junction inflow 48 hours prior, the NH_3–N concentrations and effluent flow 8 hours prior from the Kaifaqu WWTP, the cumulative rainfall runoff over the previous 48 hours, and the same meteorological factors.

At the Yangqiao (S7) section, the CODMn model inputs included the current COD_{Mn} concentrations at Kaitaiqiao (S8), Wuchaqiao (S9), and Tuqiao (S6) sections, the current COD and effluent flow from the Lukou WWTP, the cumulative rainfall runoff over the previous 48 hours, and meteorological factors (sea level pressure, wind direction, wind speed, relative humidity, and air temperature). For the NH_3–N model, the inputs included NH_3–N concentrations 8 hours prior at Kaitaiqiao (S8) and Wuchaqiao (S9), the current NH_3–N at Tuqiao (S6), the NH_3–N concentrations and effluent flow 8 hours prior from the Lukou WWTP, the cumulative rainfall runoff over the previous 48 hours, and the same meteorological factors.

At the Sanchakou (S5) section, the COD_{Mn} model inputs included the current COD_{Mn} at Qiqiaoweng (S3), the New Qinhuai River hydro-junction inflow 48 hours prior, the current outflow at Wudingmen secton, the current COD and effluent flow from the Chengdong and Kaifaqu WWTP, the cumulative rainfall runoff over the previous 48 hours, and meteorological factors (sea level pressure, wind direction, wind speed, relative humidity, and air temperature). For the NH_3–N model, the inputs included the current NH_3–N at Qiqiaoweng (S3), the New Qinhuai River hydro-junction inflow 48 hours prior, the current outflow at Wudingmen section, the NH_3–N concentrations and effluent flow 8 hours prior from the Chengdong and Kaifaqu WWTP, the cumulative rainfall runoff over the previous 48 hours, and the same meteorological factors.

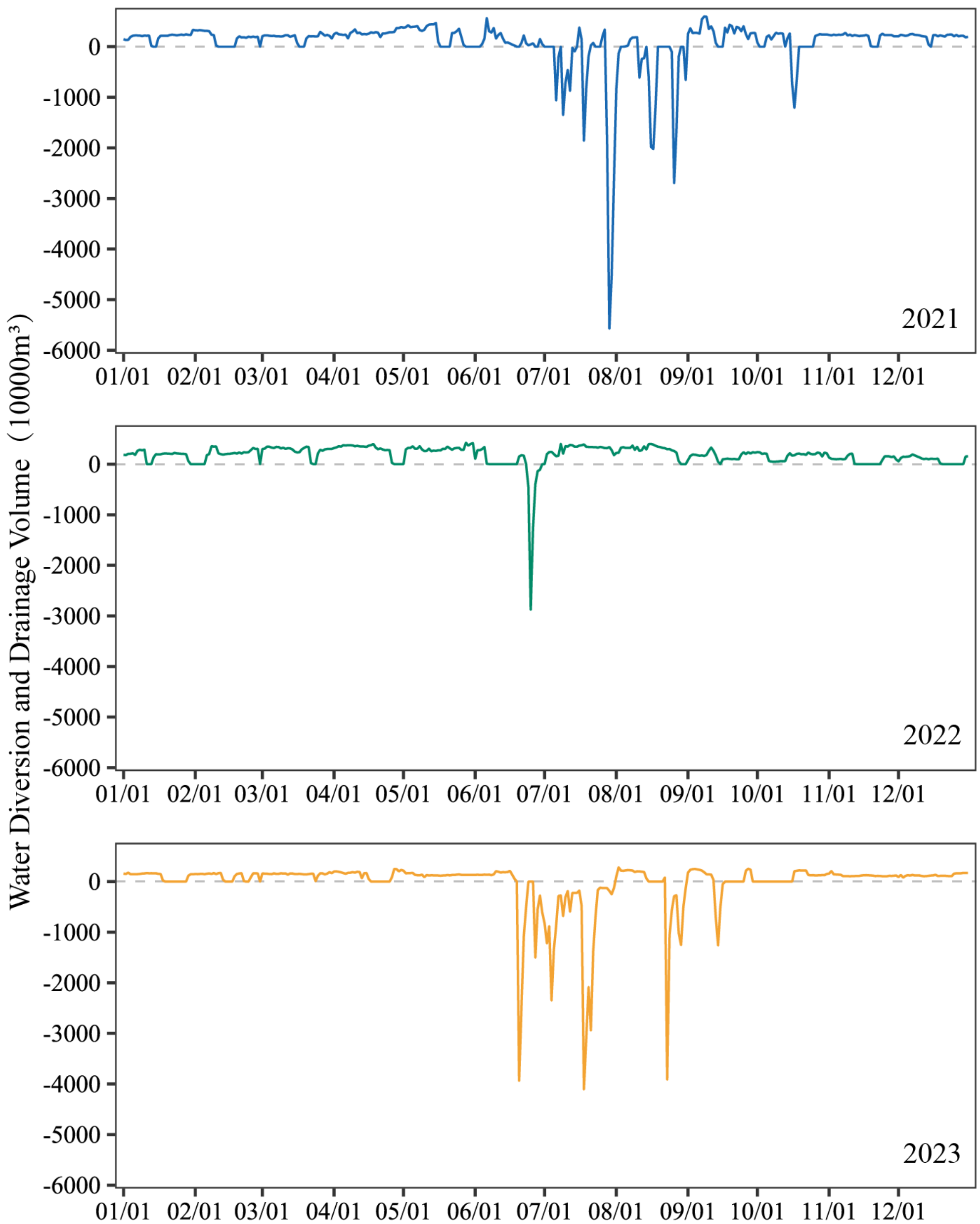


Figure S1: Water diversion and discharge volumes of the New Qinhuai River hydro-junction from 2021 to 2023 (positive values indicate water diversion; negative values indicate water discharge).

Table S1: Basic information of water quality monitoring section

No.	River Reach	Monitoring Stations	Longitude (°E)	Latitude (°N)
S1	New Qinhuai River	Jiezhizha	118.6669	31.9645
S2	New Qinhuai River	Tiexinqiao	118.7747	31.9596
S3	Qinhuai River	Qiqiaoweng	118.8249	32.0090
S4	Qinhuai River	Shitoucheng	118.7560	32.0438
S5	Qinhuai River	Sanchahekou	118.7302	32.0762
S6	Jurong River	Tuqiao	119.0627	31.8721
S7	Lishui River	Yangqiao	118.8544	31.8751
S8	Lishui River	Kaitaiqiao	119.0089	31.7616
S9	Lishui River	Wuchaqiao	118.8915	31.7715
S10	Lishui River	Huangqiao	118.8739	31.7545

Table S2: Basic information of wastewater treatment plants

No.	WWTP	Receptor Rivers	经度 (°E)	纬度 (°N)
W1	Kaifaqu	New Qianhuai River	118.7975	31.9531
W2	Chengdong	Qinhuai River	118.8625	31.9975
W3	Kexueyuan	Qinhuai River	118.8528	31.9101
W4	Tangshan	Jurong River	119.0672	32.0180
W5	Likou	Lishui River	118.8638	31.7595
W6	Qinyuan	Lishui River	119.0024	31.6650
W7	Chengbei	Qihuai River	118.8338	31.9736
W8	Zhongdian	Qinhuai River	118.8049	31.9260
W9	Nanqu	Qinhuai River	118.8250	31.8400
W10	Konggang	Qinhuai River	118.7941	31.8144
W11	Jurong	Jurong River	119.1441	31.9299
W12	Tuqiao	Jurong River	119.0613	31.9321
W13	Hengxi	Lishui River	118.8030	31.7301
W14	Dongping	Lishui River	119.1103	31.6931

Table S3: Environmental quality standards for surface water.

Unit：mg/L

Water Quality Parameter		I	II	III	IV	V
DO	≥	7.5	6	5	3	2
COD_{Mn}	<	2	4	6	10	15
NH_3–N	<	0.15	0.5	1	1.5	2.0
TP	<	0.02	0.1	0.2	0.3	0.4

Table S4: Factor rankings in Qiqiaoweng section.

Input Variables		Lag Time	COD_{Mn}	NH_3–N
Water Quality of Upstream Inflow	Water quality at the Yangqiao section	0h 4h 8h	6 11 4	6 3 2
	Water quality at the Tiexinqiao section	0h 4h 8h	5 8 2	4 8 1
Diversion and discharge volumes at the New Qinhuai River hydro-junction		0h 8h 24h 48h	9 3 7 1	13 16 10 7
Kaifaqu WWTP	Water quality	0h 4h 8h	17 19 15	18 20 17
	Discharge	0h 4h 8h	13 16 10	11 15 5
Meteorological factors	Precipitation-runoff	4h 8h 24h 48h	25 24 23 18	25 3 19 12
	Sea level pressure	0h	12	9
	Wind direction	0h	21	24
	Wind speed	0h	22	22
	Relative humidity	0h	20	21
	Air temperature	0h	14	14

Table S5: Factor rankings in Yanqiao section.

Input Variables		Lag Time	COD_{Mn}	NH_3–N
Water Quality of Upstream Inflow	Water quality at the Kaitaiqiao section	0h 4h 8h	10 16 12	6 5 3
	Water quality at the Wuchaqiao section	0h 4h 8h	1 3 4	7 11 1
	Water quality at the Tuqiao section	0h 4h 8h	5 6 8	4 8 12
Lukou WWTP	Water quality	0h 4h 8h	9 17 13	15 18 13
	Discharge	0h 4h 8h	15 20 18	16 14 10
Meteorological factors	Precipitation-runoff	4h 8h 24h 48h	24 23 22 14	24 23 22 17
	Sea level pressure	0h	7	9
	Wind direction	0h	19	19
	Wind speed	0h	21	21
	Relative humidity	0h	11	20
	Air temperature	0h	2	2

Table S6: Factor rankings in Sanchahekou section.

Input Variables		Lag Time	COD_{Mn}	NH_3–N
Water Quality of Upstream Inflow	Water quality at the Qiqiaoweng section	0h 4h 8h	1 9 5	2 4 3
Diversion and discharge volumes at the New Qinhuai River hydro-junction		0h 8h 24h 48h	13 17 18 2	19 15 12 8
Diversion and discharge volumes at the Wudingmen Section		0h 8h 24h 48h	3 16 15 4	9 11 26 16
Chengdong WWTP	Water quality	0h 4h 8h	20 29 22	30 24 22
	Discharge	0h 4h 8h	23 31 25	20 25 17
Kaifaqu WWTP	Water quality	0h 4h 8h	14 26 28	27 29 18
	Discharge	0h 4h 8h	11 19 12	10 21 7
Meteorological factors	Precipitation-runoff	4h 8h 24h 48h	32 30 10 8	28 23 5 1
	Sea level pressure	0h	6	13
	Wind direction	0h	24	32
	Wind speed	0h	27	31
	Relative humidity	0h	21	14
	Air temperature	0h	7	6

Table S7: Final optimized hyperparameters of the six XGBoost models developed for COD_{Mn} and NH3-N prediction at the Qiqiaoweng, Yangqiao, and Sanchahekou monitoring stations.

Monitoring Station	Water Quality Indicator	Eta	Gamma	Max_depth	Min_child_weight	Subsample	Colsample_bytree	Nrounds	Lambda
Qiqiaoweng	COD_{Mn}	0.0639	0.113	13	5.59	0.703	0.662	285	12.7
	NH_3-N	0.0804	0.133	16	2.31	0.866	0.712	118	13.4
Yangqiao	COD_{Mn}	0.0448	0.055	10	3.48	0.701	0.722	270	17.6
	NH_3-N	0.0931	0.0714	25	4.56	0.966	0.536	326	15.4
Sanchahekou	COD_{Mn}	0.0192	0.0536	15	9.79	0.838	0.997	460	10.4
	NH_3-N	0.0315	0.0788	13	2.52	0.92	0.59	252	10.1

Table S8: The abbreviations of the variables.

Abbreviations	Description
txqCODMn8	Tiexinqiao (S2) CODMn 8 h prior
jjzDiv48	New Qinhuai River hydro-junction inflow 48 h prior
yqCODMn8	Yangqiao (S7) COD_{Mn} 8 h prior
kfqWWTPflow8	Kaifaqu WWTP effluent flow 8 h prior
kfqWWTPCOD8	Kaifaqu WWTP WWTP COD 8 h prior
SLP	Atmospheric pressure
TEM	Air temperature
RHU	Relative humidity
PRE48	Cumulative rainfall over the previous 48 h
WD	Wind direction
WS	Wind speed
yqHN3-N8	Yangqiao (S7) NH_3–N 8 h prior
txqNH3-N8	Tiexinqiao (S2) NH_3–N 8 h prior
kfqWWTPflow8	Kaifaqu WWTP effluent flow 8 h prior
jjzDiv48	New Qinhuai River hydro-junction inflow 48 h prior
kfqWWTPNH3-N8	Kaifaqu WWTP NH_3–N 8 h prior
wsqCODMn	Wuchaqiao (S9) current COD_{Mn}
tqCODMn	Tuqiao (S6) current COD_{Mn}
lkWWTPCOD	Lukou WWTP current COD
lkWWTPflow	Lukou WWTP current effluent flow
ktqCODMn	Kaitaiqiao (S8) current COD_{Mn}
ktqNH3-N8	Kaitaiqiao (S8) NH_3–N 8 h prior
wsqNH3-N8	Wuchaqiao (S9) NH_3–N 8 h prior
tqNH3-N	Tuqiao (S6) current NH_3–N
lkWWTPflow8	Lukou WWTP NH_3–N 8 h prior
lkWWTPNH3-N8	Lukou WWTP effluent flow 8 h prior
qqwCODMn	Qiqiaoweng (S3) current COD_{Mn}
wdmDiv	Wudingmen section current outflow
jjzDiv48	New Qinhuai River hydro-junction inflow 48 h prior
kfqWWTPflow	Kaifaqu WWTP current effluent flow
kfqWWTPCOD	Kaifaqu WWTP current COD
cdWWTPCOD	Chengdong WWTP current COD
cdWWTPflow	Chengdong WWTP current effluent flow
qqwNH3-N	Qiqiaoweng (S3) current NH_3–N
kfqWWTPflow8	Kaifaqu WWTP NH_3–N 8 h prior
jjzDiv48	New Qinhuai River hydro-junction inflow 48 h prior
wdmDiv	Wudingmen section current outflow
cdWWTPNH3-N8	Chengdong WWTP NH_3–N 8 h prior
cdWWTPflow8	Chengdong WWTP effluent flow 8 h prior
kfqWWTPNH3-N8	Kaifaqu WWTP NH_3–N 8 h prior

References

[1] Li H, Hou L, Xu P. Application of various water quality index methods in water qualityevaluation of Xiajiang Reservoir (In Chinese) [J]. Yangtze River, 2020, 51(S2): 32-6+87.

[2] Hu Z, Hu L, Cheng H, et al. Application of Pollution Index Method Based on Dynamic Combination Weight to Water Quality Evaluation [J]. IOP Conference Series: Earth and Environmental Science, 2018, 153(6): 062008.

[3] McMillan P G, Feng Z Z, Deeth L E, et al. Improving monitoring of fish health in the oil sands region using regularization techniques and water quality variables [J]. Science of The Total Environment, 2022, 811: 152301.

[4] Cover T, Hart P. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-7.

[5] Li X, Sha J, Wang Z-L. Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake [J]. Environmental Science and Pollution Research, 2018, 25(20): 19488-98.

[6] Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1): 5-32.

[7] Wu C, Fang C, Wu X, et al. Health-Risk Assessment of Arsenic and Groundwater Quality Classification Using Random Forest in the Yanchi Region of Northwest China [J]. Exposure and Health, 2020, 12(4): 761-74.

[8] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM, 2016: 785–94.

[9] Deng X, Li M, Deng S, et al. Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification [J]. Medical & Biological Engineering & Computing, 2022, 60(3): 663-81.