Optimization of water quality index models using machine learning approaches

Fei Ding a,7, Wenjie Zhang a, Shaohua Cao b, Shilong Hao a, Liangyao Chen a, Xin Xie c, Wenpan Li c, Mingcen Jiang c,*

a Key Laboratory of Beijing for Water Quality Science and Water Environment Recovery Engineering, College of Architecture and Civil Engineering, Beijing University of Technology, Beijing 100124, China
b State Environmental Protection Key Laboratory of Soil Environmental Management and Pollution Control, Nanjing Institute of Environmental Sciences, Ministry of Ecology and Environment of China, Nanjing 210042, Jiangsu, China
c State Environmental Protection Key Laboratory of Quality Control in Environmental Monitoring, China National Environmental Monitoring Centre, Beijing 100012, China

ARTICLE INFO

Keywords:
Water quality assessment
Machine learning
LightGBM
Game theory
Combined weight
Chaobai River Basin

ABSTRACT

To optimize the water quality index (WQI) assessment model, this study upgraded the parameter weight values and aggregation functions. We determined the combined weights based on machine learning and game theory to improve the accuracy of the models, and proposed new aggregation functions to reduce the uncertainty of the model. A new water quality assessment system was established, and took the Chaobai River Basin as a case study. To optimize the weight, two combined weights were established based on game theory. The weight $CW_{AE}$ was combined by the Analytic Hierarchy Process (AHP) and Entropy Weight Method (EWM). The weight $CW_{AL}$ was combined by AHP and machine learning (LightGBM). $CW_{AL}$ was judged to be an optimal composite weight by comparing the coefficient of variation (CV) values and the Kaiser-Meyer-Olkin (KMO) extracted values. To reduce the uncertainty of the model, we proposed two aggregation functions, the Sinusoidal Weighted Mean (SWM) and the Log-weighted Quadratic Mean (LQM). The three water quality assessment models (WQI$_S$, WQI$_L$ and WQI$_W$) were established based on the optimal weights besides. All three models had good reliability. Both WQI$_S$ and WQI$_W$ models had low eclipsing problems (25.49% and 18.63%). The accuracy of the models was ranked as WQI$_S >$ WQI$_W >$ WQI$_L$. The uncertainty of WQI$_S$ was low, and so was WQI$_W$ (0.259) in assessing poor water quality. Overall, the WQI$_S$ model was recommended for assessing poor water quality and the WQI$_W$ model was recommended for assessing good water quality. The assessment results can provide a scientific basis for the protection of the regional water environment.

1. Introduction

Rivers are one of the most important water resources for ecological restoration and socioeconomic development (Pan et al., 2022). In recent years, human activities such as industrialization, urbanization, and rapid population growth have greatly affected the hydrological cycle, leading to changes in river water quality (Dai et al., 2022). Changes in river water quality may pose a threat to aquatic organisms and may also compromise groundwater quality, thereby affecting human health (Maansi et al., 2022). Assessment works can effectively identify the water quality of the river, as well as provide scientific recommendations for the development, utilization, and protection of the river.

There are various methods for water quality assessment, including single-factor evaluation (SF) (He et al., 2021), fuzzy comprehensive evaluation (FCE) (Yao et al., 2021), principal component analysis (PCA) (Obiri et al., 2021), and water quality index (WQI) (Sim et al., 2015). Water quality assessment methods have their own limitations (Ding et al., 2022a, 2022b). Among these, WQI is a more efficient tool in water
quality assessment methods (Rangeti et al., 2015), which can reflect the water quality condition by representing the water quality class with a single value (Dunnette, 1979; Hurley et al., 2012). The WQI model mainly consists of five main components: water quality indicators, classification scheme, weight values, sub-indices, and aggregation functions (Lumb et al., 2011; Uddin et al., 2021a, 2023a). The WQI study focuses on optimizing the WQI model by improving the parameter weight values (Ding et al., 2022a, 2022b; Ji et al., 2016a, 2016b) and aggregation functions (Gao et al., 2022a, 2022b; Pan et al., 2022).

The weight values reflect the proportion of importance of each parameter and are one of the important structures of WQI (Zhe et al., 2021). Parameter weight values are mainly divided into subjective weight values (Hamilton and Parparov, 2010; Yi et al., 2019), objective weight values (Liu et al., 2021; Yang et al., 2018), and combined weight values (Mladenovic-Ranisavljevic et al., 2018; Yan et al., 2016). Combined weights can play the advantages of many different weighting methods. When combined weights are coupled with subjective and objective weights not only can decision-makers attach importance to different indicators but also have a certain degree of objectivity (Yang et al., 2022). The combined weights based on game theory can not only combine multiple weights but also take advantage of determining more reasonable weight values (Ding et al., 2022a, 2022b; Liu et al., 2021). Among the many methods of assigning weights, because machine learning can handle large amounts of data and high-dimensional features, it is often introduced into the weights as an objective weight for water quality assessment (Shah et al., 2021; Taromideh et al., 2022).

Uddin et al. (2023b) comprehensively evaluated feature selection techniques and found that machine learning could be effective to select the crucial indicators from the given data set accurately. Ke et al. (2017) improved the strategy of XGBoost, optimized the computation of machine learning, and proposed the Light Gradient Boosting Machine (LightGBM) machine learning approach. Li et al. (2022) compared the results of five water quality models, including Categorical Boosting (CatBoost), XGBoost, and LightGBM, and concluded that "LightGBM performed the best overall". Zhou et al. (2022) used LightGBM for surface water quality prediction in the Beijing area and achieved good prediction results. To determine a more reasonable combination of weights, the combination of machine learning and subjective weights using game theory can not only balance the single disadvantage of weights but also take advantage of machine learning to deal with the importance of features. Therefore, the combination of game theory and machine learning weights needs to be studied in depth.

The aggregation function is the most important structure in the WQI model, and it’s the source of model uncertainty that determines the results of the WQI model (Sutadian et al., 2016; Uddin et al., 2022a, 2023c). The aggregation function of the WQI is divided into an un-weighted aggregation function and a weighted aggregation function (Gazzaz et al., 2012; Uddin et al., 2021a, 2022a). The two types of aggregation functions are mainly computed in the form of additive functions, multiplicative functions (Gupta and Gupta, 2021; Sutadian et al., 2018), and summation averaging (Bordalo et al., 2006; Pham et al., 2011). Most studies improved model accuracy by reducing model uncertainty. In 1973, Brown proposed a WQI aggregation function incorporating the NSF model, which improved the WQI model (Lumb et al., 2011). Dosjido et al. (1995) proposed a modified aggregation function for the Oregon WQI model (Cade, 2001). Said et al. (2004) proposed a unique logarithmic function that aggregates parameter weights with sub-logarithms. Uddin et al. (2022a) proposed two un-weighted aggregation functions and the weighted quadratic mean (WQM) aggregation function based on previous studies, which reduced the uncertainty of the model. To reduce model uncertainty and improve the accuracy of model calculation results, further research on the aggregation functions is needed.

The Chaobai River is one of the five major rivers in the Haihe River system in China and the second largest river in Beijing, and its water quality is related to the health of people in Beijing, Tianjin, and Hebei Province (Qiao et al., 2020). Water quality studies in the Chaobai River basin have focused on "PH" (He et al., 2017), "total nitrogen" (Zhaot et al., 2021), "total phosphorus" (YU et al., 2020), "antibiotics" (Zhang et al., 2016), "PFASs" (Cai et al., 2022), etc. After the completion of the South-North Water Transfer Central Project in 2014, the ecological replenishment of the Chaobai River basin was increased and the water quality of the Chaobai River basin was changed (Zhang et al., 2021). To better protect the water ecological environment of the Chaobai River Basin, it is necessary to conduct a basin-wide, multi-parameter water quality assessment study of the Chaobai River.

Based on the above discussion, this study improved the WQI model by integrating two aspects of parameter weight values and aggregation functions. By introducing machine learning, establishing combined weights, and proposing new aggregation functions to improve the WQI model to form a new water quality assessment system. We took Chaobai River Basin as a case to carry out the water quality assessment.

Based on the above considerations, the main works of this study are as follows:

1. Objective weights are determined by LightGBM and subjective weights are determined by Analytic Hierarchy Process (AHP). Next, objective weights and subjective weights are combined by game theory to obtain the combined weights. KMO Measure and coefficient of variation are used for parameter weighting meritocracy.

2. Two new aggregation functions (Log-weighted Quadratic Mean and Sinusoidal Weighted Mean) are proposed based on the existing aggregation function in this study. The optimal weights are combined with the new aggregation functions to form an improved WQI water quality assessment model and establish a new water quality assessment system.

3. The Chaobai River Basin has been taken as a case study for water quality assessment research. By comparing and analyzing the reliability, uncertainty, eclipsing problems and accuracy of the models, the optimal aggregation function and the optimal assessment model are determined to verify the rationality of the new water quality assessment system.

4. Descriptive statistical analysis of indicators in the Chaobai River Basin to analyze the spatial-temporal distribution characteristics of water quality and the causes of pollutants.

The paper is structured as follows: Section 2 describes materials and methods of study, Section 3 shows the results of parameters, weights, aggregation functions, and assessment, Section 4 presents discussions of the models’ benefits and spatiotemporal analysis of typical pollutants, Section 5 summarizes the main conclusions.

2. Materials and methods

2.1. Study area

The Chaobai River Basin is one of the five major rivers of the Haihe River Basin in China, and the basin runs through the provinces of Hebei, Beijing, and Tianjin. The study area is located at 115°20′00″E–118°00′00″E, 39°00′00″N–41°00′00″N. The region has a typical north-temperate subhumid continental monsoon climate. The Chaobai River Basin is dry and cold in winter while hot and humid in summer. In the past 50 years, the average annual temperature is 9–10 °C, and the annual rainfall is 300–700 mm. The mean annual rainfall in 2021 is 431.97 mm, and the annual evaporation is 1175 mm (Gao et al., 2022a, 2022b). Data from the National Meteorological Science Data Center (http://data.cma.cn). Precipitation during the rainy season (June to September) accounts for more than 80% (Xu et al., 2009), and the precipitation in July accounts for 32% of the annual precipitation (Zhaot et al., 2021). Chaobai River has a total length of 467 km and a drainage area of 19,354 square kilometers. The water area is...
14.8 square kilometers, the average river width is 500 m, the widest water surface is 800 m, the average river depth is 2.5 m, and the average river velocity is 0.015 m/s (Su et al., 2021). A map of sampling site locations in the study area is shown in Fig. 1.

2.2. Sampling sites arrangement and monitoring

In this study, 17 sampling sites in the Chaobai River Basin were studied. The geographical information of the sampling sites was shown in Fig. 1. The water quality data were collected from the monthly sampling data of the state-controlled sections of surface water by the China National Environmental Monitoring Center in 2021.

The pollutant indicators of the state-controlled sections include organic pollutants, inorganic pollutants, metal pollutants, and others, a total of 19 indicators. Organic pollutant indicators: dissolved oxygen (DO), permanganate index (COD\textsubscript{Mn}), chemical oxygen demand (COD\textsubscript{Cr}), and five-day biochemical oxygen demand (BOD\textsubscript{5}); inorganic pollutant indicators: ammonia nitrogen (NH\textsubscript{3}–N), total phosphorus (TP), total nitrogen (TN), fluoride (F\textsuperscript{−}), prussiate (CN), sulfide (S), selenium (Se), arsenic (As); metal pollutant indicators: copper (Cu), zinc (Zn), hydrargyrum (Hg), cadmium (Cd), hexavalent chromium (Cr\textsuperscript{6+}), lead (Pb); other indicator: potential of hydrogen (pH). Monitoring methods based on technical specifications for surface water environmental quality monitoring (HJ 91.2—2022), and the specific method of water sample detection were shown in Table S1.

2.3. Materials and methods

Firstly, collecting water quality information and conducting statistical analysis of water quality data in the basin. Secondly, the WQI model is improved from two aspects of parameter weight values and aggregation functions, and the optimized WQI model is established. The optimized WQI model is divided into five components: water quality indicators, classification scheme, weight values, sub-indices, and aggregation functions. The determination of the five components is taken through four steps. (i) Water quality indicators and classification scheme are determined by the available water quality data and the specification. (ii) Calculation methods of weight include: subjective weight (Analytic Hierarchy Process (AHP)), objective weights (Entropy Weight Method (EWM) and LightGBM), the combined weight of AHP and EWM (CW\textsubscript{AE}), and the combined weight of AHP and LightGBM (CW\textsubscript{AL}). The optimal weights are determined by calculating the Kaiser-Meyer-Olkin (KMO) Measure and Coefficient of Variation (CV). (iii) The parameter sub-indices are the scales of the WQI model and controlled the upper and lower limits of the WQI results. In this study, developed linear interpolation rescaling function (DLIR), exponential function (EF), and inverse sine function (ISF) are selected for the calculation of the parameter sub-indices. (iv) The aggregation functions of Log-weighted Quadratic Mean (LQM) and Sinusoid Weighted Mean (SWM) are proposed and compare with the weighted quadratic mean Weighted Quadratic Mean (WQM). The WQI\textsubscript{L}, WQI\textsubscript{S}, and WQI\textsubscript{W} models are built based on the aggregation functions, while the reliability, accuracy, and uncertainty of the models are judged to be excellent. Finally, the optimized WQI model is used for water quality assessment.

We propose an improved WQI model to establish a new water quality assessment system and take the Chaobai River Basin as a case study to explore the optimal model. The flowchart of the research method is shown in Fig. 2.

2.3.1. Water quality indicators and classification scheme

1. Indicators selection: WQI selects the appropriate indicators according to the availability of data and the purpose of the study. There is no uniform standard for the number of indicators selected. Uddin et al. (2021a) considered that 4--26 indicators are reasonable. The selection of indicators will be based on the following aspects:

![Fig. 1. Map of sampling sites in the Chaobai River Basin.](image-url)
(1) Completeness of monitoring data. A total of 19 monitoring indicators are collected in 2021, and some data are lost or not detected.

(2) Typical pollutants in the region. Nitrogen is the key pollutant in the Chaobai River basin (Zhao et al., 2021), and NH$_3$–N, TN, COD$_{Cr}$, BOD$_5$, and DO indicators are often used in water quality studies in the basin (Chen et al., 2022; Zhang et al., 2021; Zhao et al., 2021). Ji et al. (2016a, 2016b) found that the basin had Cd contamination. The analysis of our monitored data find that the basin had water samples with concentrations of Cu, Zn, and Cd exceeding Class I standards in May, July, August, and October.

(3) Common water quality assessment indicators. Common water quality assessment indicators include DO, COD$_{Mn}$, COD$_{Cr}$, BOD$_5$, NH$_3$–N, and TN (Ding et al., 2022a, 2022b; Yang et al., 2023). Indicators of pH, DO, TP, F$^-$, COD$_{Mn}$, COD$_{Cr}$, and BOD$_5$ are often used in WQI for surface water quality assessment studies (Brunner et al., 2023; Georgescu et al., 2023; Uddin et al., 2022b).

In this study, the feature selection of LightGBM is taken for water quality indicators. The final water quality indicators are obtained by combining the above considerations. The results of the selected indicators include nitrogen communities (NH$_3$–N, TN), metals and metalloids (Cu, Zn, Hg, Cd, Pb, As), and other indicators (pH, DO, COD$_{Mn}$, COD$_{Cr}$, BOD$_5$, TP, F$^-$, S), excluding CN, Se, and Cr$^{6+}$, and a total of 16 water quality indicators are identified.

2. Classification scheme: In this study, according to the National Environmental Quality Standard for Surface Water (GB3838-2002) in China, the water quality is divided into five categories, Class I, Class II, Class III, Class IV, and Class V. The indicator concentrations of each grade of water quality are shown in Table S2.

2.3.2. Weight values

The weight value is an important part of WQI (Uddin et al., 2021b), and affects the results of water quality assessment (Gai and Guo, 2023). Seifi et al. (2020) found that differences in weight selection directly affect the uncertainty of WQI, which subsequently affects the calculation results of the model.

Objective weights cannot highlight the characteristics of the key pollutants, subjective weights tend to overlook the data information, and combined weights can avoid the limitations of single-weight values (Ding et al., 2022a, 2022b; Li et al., 2023). Liu et al. (2021) found that the combined weights can be calculated more rationally using game theory. Ding et al. (2022a, 2022b) combined objective weights (entropy weight method (EWM)) and subjective weights (AHP) by game theory, and found that the best results of water quality assessment based on combined weights (CW$_{AE}$).

Machine learning can handle multidimensional features (Wang et al., 2023). To improve the weight part and explore reasonable weight, machine learning (LightGBM) is used to determine the objective weights, and AHP is used to determine the subjective weights. Then combined weights (CW$_{AL}$) are determined by game theory in this study.

1. Combined weight of AHP and EWM: Ding et al. (2022a, 2022b) described in detail the calculation steps regarding the game theoretic combination of AHP and EWM in their study. Among them, the objective weight value based on EWM is denoted as $W'_o$, the subjective weight based on AHP is denoted as $W'_s$, and the combined weight CW$_{AE}$ through game theory is denoted as $W$. Therefore, the specific steps of the game theoretic combination of AHP and EWM are not duplicated in this paper.

2. Combined weight of AHP and lightGBM: Combined weights using game theory can be calculated to reach Nash equilibrium and determine the optimal weights (Ding et al., 2022a, 2022b; Liu et al., 2021). The introduction of machine learning into the game theory to combine
weights can take advantage of processing large amounts of data and high-dimensional features and compensate for the shortcomings of single weights. Three of the main steps are as follows:

(1) Subjective weight AHP

AHP is selected as the calculation method of subjective weight, and the important calculation steps are as follows.

a. Establish the judgment matrix A, denoted as:

\[
A = \begin{bmatrix}
    a_{11} & a_{12} & \cdots & a_{1n} \\
    a_{21} & a_{22} & \cdots & a_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{n1} & a_{n2} & \cdots & a_{nn}
\end{bmatrix},
\]

\( a_{ij} = 1, \) \( j \neq i \)

(1)

Where \( a_{ij} \) is the relative importance,

b. The judgment matrix is normalized and summed by row to obtain the eigenvectors of each row and normalized by column to obtain the subjective weights \( W \):

\[
W = \begin{bmatrix}
    \sum_{j=1}^{n} \left( a_{1j} / \sum_{j=1}^{n} a_{1j} \right) \\
    \vdots \\
    \sum_{j=1}^{n} \left( a_{nj} / \sum_{j=1}^{n} a_{nj} \right)
\end{bmatrix},
\]

(2)

(2) Objective weight LightGBM

The method of LightGBM (source code address: https://github.com/Microsoft/LightGBM) was selected to rank the feature importance of indicators, and the objective weight \( W_{o} \) was calculated by Rank Order Centroid (ROC) based on the ranking results.

a. Rationale of LightGBM

Gradient Boosting Decision Tree (GBDT), which is the foundation of LightGBM, divides data to each node by optimal cutting point (maximum information gain point) and measures information gain by variance after segmentation (Friedman, 2001). LightGBM improves GBDT-based information gain estimation and the number of features determination (Ke et al., 2017). LightGBM improves GBDT by Gradient-based One-Side Sampling (GOSS) algorithm. GOSS is an algorithm that balances reducing the amount of data and ensuring accuracy. First, the training instances are sorted in descending order according to their absolute values of gradients. Second, the algorithm retains the head \( a \times 100\% \) data instances with larger gradients to obtain a subset of instances \( A \). For the remaining set \( A' \) consisting of \( (1-a) \times 100\% \) data instances with smaller gradients, a further subset \( B \) of data with size \( b \times |A'| \) is randomly selected. Finally, the subset \( A \cup B \) is partitioned according to the estimated variance gain \( \bar{V}_{j}(d) \):

\[
\bar{V}_{j}(d) = \frac{1}{n} \left( \frac{\left( \sum_{i \in A} a + \frac{1-a}{b} \sum_{i \in A} a \right)^{2}}{n_{A}'(d)} + \frac{\left( \sum_{i \in A} a + \frac{1-a}{b} \sum_{i \in A} a \right)^{2}}{n_{B}'(d)} \right).
\]

(3)

Among them, \( A_{l} = \{ x_{i} \in A : x_{i} \leq d \}, \) \( A_{r} = \{ x_{i} \in A : x_{i} > d \}, \) \( B_{l} = \{ x_{i} \in B : x_{i} \leq d \}, \) \( B_{r} = \{ x_{i} \in B : x_{i} > d \} \). The coefficient \( \frac{1-a}{b} \) is used to normalize the sum of the gradients on \( B \) to the size of \( A' \).

b. Feature importance ranking of LightGBM

In this study, the tree model of LightGBM established above is combined with the indicators selected in Section 2.3.1 to rank the importance of regional water quality features. The implementation of the model is based on open source code and Python 3.9, and the determined indicators are characterized as \( i = 1, 2, \ldots, N \).

(i) The more times the indicator feature makes a critical decision (split) in the decision tree, the higher its relative importance. The feature importance score is calculated based on the number of times each indicator feature splits the training data in all trees. The importance of indicator features obtained by LightGBM operation is noted as \( F[i] \) (Ke et al., 2017).

(ii) The numerical size ranking is performed on the indicator feature importance \( F[i] \). The expression of the size ranking is \( \text{RANK}(X) \), so the order of the indicators after the size ranking is \( \text{RANK}(F[i]) \), and the ranking result reflects the importance of the indicators.

c. Determination of objective weights \( W_{o} \)

This study determines the objective weights of LightGBM based on the Rank Order Centroid (ROC). The ROC method maintains the ranking order and minimizes the maximum error in the weights, which can reduce the risk of overlap due to inappropriate weights (Roszkowska, 2013). The ROC is calculated as follows:

\[
W_{o} = \begin{bmatrix}
    \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\text{RANK}(F[k])} \\
    \vdots \\
    \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\text{RANK}(F[k])}
\end{bmatrix}
\]

(4)
Where \( N \) is the number of indicators, \( RANK(F(i)) \) is the ranking result of feature importance, and \( W_o \) is the objective weight.

(3) Combined weight \( CW_{AL} \)

Game theory determines the optimal weights by optimizing the weight coefficients \( a_i \) in the equation so that the deviation between \( w_j \) and each \( w_o \) is minimized.

\[
\min \left\| \sum_{j=1}^{n} a_j w^*_j - w_j \right\|_2, \quad (k = 1, 2, \ldots, n) \tag{5}
\]

According to the differential properties of the matrix, the conditions for the optimal first order derivative of the equation are as follows:

\[
\sum_{j=1}^{n} a_j w^*_j = w_j, \quad (j = 1, 2, \ldots, n) \tag{6}
\]

The above subjective weights \( W_s \), objective weights \( W_o \) as the two weights adopted in this case, and their corresponding linear equation systems are written in matrix form as follows:

\[
\begin{bmatrix}
W^T \cdot W & W^T \cdot W_o \\
W^T \cdot W_o & W^T \cdot W_o
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2
\end{bmatrix}
= \begin{bmatrix}
W^T \cdot W \\
W^T \cdot W_o
\end{bmatrix}
\begin{bmatrix}
\alpha_s \\
\alpha_o
\end{bmatrix} \tag{7}
\]

The weight coefficients \( a_1, a_2 \) are normalized by solving for:

\[
a'_1 = \frac{a_1}{a_1 + a_2}, \quad a'_2 = \frac{a_2}{a_1 + a_2} \tag{8}
\]

Finally, the portfolio weights are calculated as follows:

\[
W = a'_1 W_s + a'_2 W_o \tag{9}
\]

Where the final combined weight matrix is \( W \) and the final weight of each water quality indicator is \( W_i \).

3. Excellent selection: An excellent selection is performed for the parameter weight values. There are many methods of merit selection, and in this paper, the calculation of the coefficient of variation (CV) with the Kaiser-Meyer-Olkin (KMO) Measure is selected. The steps for the excellent selection of weights are as follows:

1. Coefficient of variation reflects the degree of dispersion of the data, the smaller the CV value, the more stable the data are, and the better the weights are.
2. KMO Measure can be used to measure the overall information overlap level of the indicator set (Ma et al., 2022). We used KMO Measure to determine the optimal weights, and the larger the KMO value the more reasonable the weights. The formula for KMO calculation is as follows:

\[
KMO = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij}^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij}^2 + \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij}^2} \tag{10}
\]

Where \( r_{ij} \) denotes the Pearson correlation coefficient of weight \( W_i \) with another weight \( W_j \), and \( p_0 \) denotes the n-2nd order deviation correlation coefficient of \( W_i \) with \( W_j \).

2.3.3. Sub-indices

A sub-index of the indicator is one of the means of converting the actual measured values of a water quality indicator into dimensionless values (Uddin et al., 2022a, 2023a, c). Sub-indices refer to data transformations within a range of values for a common scale (Georgescu et al., 2023; Sutadian et al., 2016).

The calculations of SI values in existing WQI models are varies (Gupta and Gupta, 2021), such as the method for non-standardized variables (Said et al., 2004), comparison with the permissible limits (Liou et al., 2004), and linear interpolation rescaling (Misaghi et al., 2017), etc. Linear interpolation rescaling (LIR) can be used to compare indicators of different units easily, and it is well used in the WQI model (Akhtar et al., 2021; Uddin et al., 2023a). the range of SI values is usually 0–100, where 0 means ‘poor’ and 100 means ‘good’ water quality (Uddin et al., 2021a). LIR is chosen for the calculation of SI values in this study. The formula for calculating the sub-indices is as follows:

\[
SI = S_i - \left[ (S_i - S_j) \frac{x_i - x_j}{x_2 - x_1} \right] \tag{11}
\]

\[
SI = S_i - \left[ (S_i - S_j) \frac{x_1 - x_i}{x_2 - x_1} \right] \tag{12}
\]

Where \( SI \) is the sub-index value, \( S_1 \) and \( S_2 \) are the scale values of the upper and lower limits of the common scale, mostly from 100 to 0 (Sutadian et al., 2018; Zhang, 2019), \( x_1 \) and \( x_2 \) are the allowed values of the lower and upper limits of the indicator. \( x_i \) is the actual measured value of the parameter. The SI value of each indicator is recorded as \( S_i \).

In this study, a non-linear scale will be used to apply to water quality assessment for better adaptation of the aggregation function. The scales of \( S_1 \) and \( S_2 \) are set as nonlinear functions such as exponential function (EF) and inverse sine function (ISF).

1. Developed linear interpolation rescaling functions: Uddin et al., (2022a) improved the calculation method of LIR and proposed developed linear interpolation rescaling functions (DLIR). The functions of DLIR are as follows:

\[
SI = \left( S_i - S_j \right) \frac{x_i - x_j}{x_2 - x_1} \tag{13}
\]

\[
SI = \frac{x_i - x_j}{x_2 - x_1} \times S_i \tag{14}
\]

\[
SI = \left( S_i - S_j \right) - \frac{x_i - x_j}{x_2 - x_1} \tag{15}
\]

\( S_1 \) and \( S_2 \) are the scale values of the upper and lower limits of the common scale, from 100 to 0. The calculation formula applied to the indicators were shown in Table S3.

2. Exponential function: Based on the SI values calculated by EF using Eq. (13) to Eq. (15), the scales of \( S_1 \) and \( S_2 \) are calculated in this study by taking an exponential function with a base of 10, which is calculated as follows:

\[
S_{1,2} = 10^{0.6/50} - 1 \tag{16}
\]

Where the value of \( S_i \) is taken as the grade scale in the DLIR model, the calculated \( S_1 \) and \( S_2 \) range from 99 to 0.

3. Inverse sine function: Similarly, the SI values calculated based on ISF are used in Eq. (13) to Eq. (15) to calculate the scales of \( S_1 \) and \( S_2 \), which are calculated as follows:

\[
S_{1,2} = \arcsin \left( \frac{S_i}{100} \right) \tag{17}
\]

Where the value of \( S_i \) is taken as the grade scale in the DLIR model, the calculated \( S_1 \) and \( S_2 \) range from 1.571 to 0.

2.3.4. Aggregation functions

The aggregation function is the most important part of WQI, it is an important presentation tool for water quality condition data.

1. Establishment of aggregation functions: This study introduces logarithmic functions and trigonometric functions to establish new aggregation functions. And we proposed the WQI models of Log-weighted Quadratic Mean (LQM) and Sinusoidal Weighted Mean (SWM). The Weighted Quadratic Mean (WQM) is used as a parallel aggregation
function to compare with the aggregation function proposed in this study, and the specific WQI aggregation function expressions are shown in Table 1.

2. Excellent selection: The aggregation function is obtained by aggregating the weight values and sub-indices to obtain the WQI values, and the optimal aggregation function is determined by WQI model selection. The main steps of model selection are as follows:

(i) Reliability of the model. Ding et al. (2022a, 2022b) compared and analyzed different water quality assessment results of the same basin to determine whether the model calculation results are reliable. A similar approach is used for model reliability analysis in this study.

(ii) Eclipsing problem and accuracy of the models. The eclipsing problem is used to reflect the amount of deviation from the model results, and the accuracy is used to reflect the amount of correctness from the model results.

(iii) Calculation of standard uncertainty. The water quality level will be quantified, the standard uncertainty is calculated as follows:

\[
SU = s(x) = \left( \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \right)^{1/2}
\]

Where \(SU\) means standard uncertainty, \(s(x)\) is the standard deviation of the measurement result, \(x_i - \bar{x}\) is the residual of the data result, and \(N\) is the number of samples.

(iv) Hypothesis testing of the model. Student’s t-test (well known as t-test) is a concise method of hypothesis testing (Shaw, 2017), which can be used to demonstrate that the SU calculation is credible through objective results (Uddin et al., 2023a). The t-test is calculated as follows:

\[
t = \frac{\bar{x} - \mu_0}{s_x/\sqrt{N}}
\]

\[
\nu = N - 1
\]

Where, \(N\) is the sample size, and \(\bar{x}\) is the mean of the sample, \(\mu_0\) is the population mean, \(s_x\) is the standard deviation of the sample. Where \(\nu\) is the degree of freedom. The t-value is the statistical quantity. It can be used to get to the p-value based on the t-value, which in turn indicates whether there is a significant difference.

Table 1: Overview of different WQI model aggregation functions.

<table>
<thead>
<tr>
<th>Models</th>
<th>Aggregation functions</th>
<th>Sub-indices</th>
<th>Refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted Quadratic Mean (WQM)</td>
<td>WQI = \sqrt{\sum_{i=1}^{n} W_i S_i^2}</td>
<td>DLIR (Uddin et al., 2022a)</td>
<td></td>
</tr>
<tr>
<td>Log-weighted Quadratic Mean (QQM)</td>
<td>WQI = \ln(\sum_{i=1}^{n} W_i S_i)</td>
<td>EF Proposed by authors.</td>
<td></td>
</tr>
<tr>
<td>Sinusoidal Weighted Mean (SWM)</td>
<td>WQI = \frac{100 \sum_{i=1}^{n} W_i S_i}{\sum_{i=1}^{n} W_i}</td>
<td>ISF Proposed by authors.</td>
<td></td>
</tr>
</tbody>
</table>

(i) Distribution features of the model results. The generation of probability density function (PDF) is used to determine whether the distribution of WQI results satisfies a normal distribution (Uddin et al., 2023c). When the model is normal distribution, the model is considered to be fine. Calculation and testing of uncertainty can be performed.

(ii) The determination of the standard water quality grade. In the calculation of standard uncertainty before the priority to determine the standard water quality level. The standard water quality grade determination method is as follows: establish criteria for the guideline of water grades in the WQI model. Determine the number of indicators monitor values that exceed water quality standards. Based on criteria to determine the standard water quality grade of each water sample.

(iii) Calculation of standard uncertainty. The water quality level will be quantified, the standard uncertainty is calculated as follows:

\[
SU = s(x) = \left( \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \right)^{1/2}
\]

Where \(SU\) means standard uncertainty, \(s(x)\) is the standard deviation of the measurement result, \(x_i - \bar{x}\) is the residual of the data result, and \(N\) is the number of samples.

2.3.5. Water quality assessment of WQI

The proper classification of water quality is very important for the work of water quality assessment (Uddin et al., 2023d). Water quality is divided into five grades according to the China National Environmental Quality Standard for Surface Water (GB3838-2002). Liu et al. (2010) improved the classification criteria of inferior to Class V by adding the grade of “inferior to Class V”. The water quality standard of “inferior to Class V” was also added in this study. Sutdian et al. (2018) assigned the WQI values of the five water quality grades as 100–90, 90–75, 75–50, 50–25, 25–5. The classification criteria of WQI for inferior to Class V water quality is improved by adding 5–0 as the WQI value of inferior to Class V in this study.

The results of the WQI values calculated according to the model in Section 2.3.4 are matched with their water quality grades, and the corresponding water quality conditions are derived. The water quality
established above and calculated by the new water quality model. The results of the study were based on the water quality assessment system shown in Fig. 3 e and Fig. 3 g. Only a small number of samples in box plots. The statistical results of each indicator were shown in Fig. 3.

The statistical results of metals and metalloids. Fig. 3 i samples at sites 9, 10, 11, and 12 met the standard. Fig. 3 o showed that the sample concentration at each sampling site exceeded the III water standard, and only a small number of water samples exceeded the standard at site 14.

3. Results

Chaobai River Basin was taken as a case study, and the monitored data from 17 sampling sites in 2021 were used as the data source. The results of the study were based on the water quality assessment system established above and calculated by the new water quality model.

3.1. Results of water quality indicators statistics

This study collated the data collected from 17 sampling sites firstly, and used mathematical statistics to draw the values of indicators into box plots. The statistical results of each indicator were shown in Fig. 3 below.

The statistical results of nitrogen concentrations in the water samples were shown in Fig. 3e and Fig. 3g. Only a small number of samples in Fig. 3e had NH$_3$–N concentrations exceeding the Class III water standard. In Fig. 3g, the mean value of TN concentration at each sampling site exceeded the III water standard, and only a small number of water samples at sites 9, 10, 11, and 12 met the standard. Fig. 3i–3n showed the statistical results of metals and metalloids. Fig. 3i–3n showed the concentrations of metals and metalloids in the samples. Hg, Pb, and As were far from meeting the Class I water standard. Cd in site 2 and site 4 had higher concentrations and the average value exceeded the Class I standard, the other sites did not exceed. Zn in site 2 and site 4 had higher concentrations and the average value exceeded the Class I standard, the other sites did not exceed.

The DO concentrations at each sampling site in Fig. 3a met the Class III water standard of 5 mg/L. In the statistics of COD$_{Mn}$, COD$_{Cr}$, and BOD$_5$ indicators in Fig. 3b–3d, most of the water samples had reached the standard, but some of them still exceeded the Class III standard. Among them, the sampling site where the average concentration of COD$_{Mn}$ exceeded the standard was site 15. The sampling sites where the average concentration of COD$_{Cr}$ exceeded the standard were sites 13, 14, 15, 16, and 17. The sampling sites where the average concentration of BOD$_5$ exceeded the standard were sites 14 and 15. For site 15, the concentrations of COD$_{Mn}$, COD$_{Cr}$, and BOD$_5$ exceeded Class III. Only a small number of water samples in Fig. 3f had TP concentrations exceeding the Class III water standard. In particular, site 10 was located in Miyun Reservoir, according to the standard which should be distinguished from the standards of other sampling sites, the standard concentration of TP for Class III water should be 0.05 mg/L. The water samples from site 10 also met the standard in the more stringent standard. In Fig. 3h, the mean value of F$^-$ concentration at each sampling site all reached Class III, and only less than half of the water samples at site 6 exceeded the standard. Fig. 3i showed that the sample concentration of S is basically inferior to the Class I standard. Fig. 3p showed that the water quality standard for pH was limited to 6–9, and some samples exceeded the standard at site 14.

3.2. Results of weight calculation

In this study, subjective weights were obtained by AHP, objective weights were obtained by EWM and LightGBM. And two combination weights, CW$_{AE}$ and CW$_{AL}$, were obtained by combining subjective and objective weights through game theory. The rmse value calculated by LightGBM was equal to 0.0678. Some of the parameters of the LightGBM implementation were adjusted in Table 3.

The results of the calculation of the five weights were shown in Table 4. The calculation steps of AHP and EWM were shown in Tables S4 and Table S5. The maximum weight of the subjective weight was BOD$_5$ with $W_s$ value of 0.2305, and the minimum weight was As with $W_s$ value of 0.0041. The maximum weight of the objective weight calculated by EWM was DO with a $W_o$ value of 0.2436, and the minimum weight was Hg with a $W_o$ value of 0.0027. While the maximum weight of the objective weights calculated by LightGBM was TN with a $W_o$ value of 0.2114 and the minimum weight was pH with a $W_o$ value of 0.0039. The maximum value obtained by combining the weights CW$_{AE}$ was 0.2141 for DO and the minimum value was 0.0061 for Hg. The maximum value obtained by combining the weights CW$_{AL}$ was 0.1863 for BOD$_5$ and the minimum value was 0.0104 for Hg.

The weights of indicators in Table 4 were statistically analyzed and KMO Measure was done. The statistical and test results were shown in Table 5. The standard deviation (SD) of the objective weight values calculated by LightGBM was 0.0573, and the SD of EWM was 0.0632. The dispersion of the objective weights calculated by the two different methods was not much. SD of LightGBM is 0.0059 lower than the SD of EWM. When the objective weights were combined with the subjective weights, the SD of CW$_{AL}$ was 0.0021 lower than CW$_{AE}$, and the coefficient of variation (CV) of CW$_{AL}$ was 0.0336 lower than CW$_{AE}$, which indicated that the combined weights CW$_{AL}$ were more stable than the CW$_{AE}$. The KMO Measure of sampling adequacy was 0.614, which was greater than the untestable value of 0.5, indicating that the KMO Measure could be done and the data could be effectively extracted information (Mandal et al., 2022; Santiago et al., 2022). The largest KMO extraction value was 0.981 for CW$_{AL}$, the second KMO extraction value was 0.960 for CW$_{AE}$. The value of KMO extraction was: CW$_{AL}$ > CW$_{AE}$ > AHP > EWM > LightGBM. In summary, the combined weight CW$_{AL}$ was more stable and had the best explanatory capability for each variable, the result of weight meritocracy was CW$_{AL}$.

3.3. Results of aggregation functions

In this study, the combined weights CW$_{AL}$ obtained based on the game theory combination LightGBM and AHP were selected as the optimal weights. The optimal weights were applied to three aggregation functions WQM, LQM, and SWM to form new water quality assessment models WQM, LQM, and SWM respectively. The data from 17 sampling sites in Chaobai River Basin were calculated by the three models, and the WQI results obtained by the three models were shown in Fig. 4. The results of the calculation of sub-indices and aggregation functions are shown in Table S6–S8.

In Fig. 4a–4b, the trends of WQI values calculated by the site distribution of the three models were consistent, and the mean values of WQI, higher than the mean values of WQI, lower under the same conditions. All three models reflected the trend of decreasing water quality after site 10, and the water quality after site 11 was slightly worse than that before site 10. The mean values of WQI in the Chaobai River Basin calculated by the three models WQI, WQI, and WQI were 83.486, 91.160, 77.610, respectively. The order of the maximum, minimum, and mean values of the three models was WQI > WQI > WQI. In Fig. 4c–4d, the upper bounds of the three models did not differ much. Only after August, the upper bounds of the WQI model started to be lower than those of the other two models. The lower limit of WQI was higher than that of WQI and that of WQI in all three models. The mean folds of the three models had the same trend and all reached the minimum value in July.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>WQI value reflects the relationship of water quality conditions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water quality grade</td>
<td>Pollution description</td>
</tr>
<tr>
<td>Class I</td>
<td>Clean</td>
</tr>
<tr>
<td>Class II</td>
<td>Slightly polluted</td>
</tr>
<tr>
<td>Class III</td>
<td>Moderately polluted</td>
</tr>
<tr>
<td>Class IV</td>
<td>Heavily polluted</td>
</tr>
<tr>
<td>Class V</td>
<td>Seriously polluted, but neither black nor malodorous</td>
</tr>
<tr>
<td>Inferior to Class V</td>
<td>Seriously polluted, also black and malodorous</td>
</tr>
</tbody>
</table>

The level represents the water quality condition and its corresponding WQI value is shown in Table 2.
Fig. 3. Data box plots of water quality indicators of 17 sampling sites in the Chaobai River Basin.
3.4. Results of water quality assessment

The values calculated based on WQI\(_W\), WQI\(_L\), and WQI\(_S\) models were used to evaluate the water quality of the Chaobai River Basin, and the assessment results were shown in Fig. 5 below.

Class III water could be evaluated in the WQI\(_W\) and WQI\(_S\) models in Fig. 5, but could not be evaluated in WQI\(_L\) model. In Fig. 5a, the WQI\(_W\) results had 11 Class I water samples, 178 Class II water samples, and 15 Class III water samples, accounting for 5.39%, 87.25%, and 7.35% of the total samples, respectively. In Fig. 5b, WQI\(_W\) results had 152 Class I water samples, 52 Class II water samples, and none of the Class III water samples, accounting for 74.51%, 25.49%, and 0.00% of the total samples, respectively. WQI\(_W\) and WQI\(_S\) had a higher percentage of Class II water quality, and WQI\(_L\) had a higher percentage of Class I water quality.

The final assessment results of the WQI\(_W\) and WQI\(_S\) were all Class II, and assessment results of WQI\(_L\) was Class I. The pollution level evaluated by WQI\(_W\) and WQI\(_S\) was "slightly polluted", the pollution level evaluated by WQI\(_L\) was "clean".

4. Discussions

Based on the above results, this study focused on the selection of weight values, the selection of aggregation functions, also the spatio-temporal analysis of typical pollutants in the basin.

4.1. Excellent selection of weights

In this study, ranked fitting was performed to explore the fitting of the objective weights and the combined weights for the four weights: EWM, LightGBM, CW\(_{AE}\), and CW\(_{AL}\). Based on the high KMO extraction values of the four weights, fitting the weights in order from low to high could explain the structure of the weights well, and the higher the fitting indicated that the structure of the weights was more stable. The linear fitting results of the four weights were shown in Fig. 6 below. In Fig. 6, the R\(^2\) result of the fitting of LightGBM weights was 0.8389 higher than that of EWM weights, and the R\(^2\) result of CW\(_{AL}\) was 0.8308 higher than that of CW\(_{AE}\). It indicated that the fitting of the objective weights based on LightGBM was better than EWM, and CW\(_{AL}\) was better than CW\(_{AE}\).

For the excellent selection of parameter weights, CW\(_{AL}\) was selected as the optimal weight for the subsequent steps in this study. The reasons for discussing the results were: (1) The KMO extraction values of the three single weights of AHP, EWM, and LightGBM are not as good as the combined weights. (2) The objective weights EWM and LightGBM did not have the low CV of the combined weights and were unstable. The KMO extraction value was not as high as the combined weights and the representativeness was low. (3) The combined weights CW\(_{AL}\) and CW\(_{AE}\) were stable and representative and were excellent weights. CW\(_{AL}\) had the lowest CV and KMO extraction value was the largest, and the ranking fitting was also better, so CW\(_{AL}\) was selected as the optimal weight.

4.2. Excellent selection of aggregation functions

The excellent selection for the aggregation function focused on the effect of the aggregated WQI model. The WQI\(_W\), WQI\(_L\), and WQI\(_S\) models were obtained based on the combined weight CW\(_{AL}\) combined with three aggregation functions, respectively.

### Table 3

<table>
<thead>
<tr>
<th>Objective parameters</th>
<th>Setting values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>rmse</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.11</td>
</tr>
<tr>
<td>Max. depth</td>
<td>6</td>
</tr>
<tr>
<td>Num. leaves</td>
<td>7</td>
</tr>
<tr>
<td>KFold</td>
<td>3</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th>Water quality parameters</th>
<th>AHP (W(_W))</th>
<th>EWM (W(_L))</th>
<th>CW(_{AE}) (W)</th>
<th>LightGBM RANK</th>
<th>LightGBM (W(_L))</th>
<th>CW(_{AL}) (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DO</td>
<td>0.1883</td>
<td>0.2436</td>
<td>0.2141</td>
<td>2</td>
<td>0.1488</td>
<td>0.1729</td>
</tr>
<tr>
<td>COD(<em>{b</em>{a}})</td>
<td>0.0716</td>
<td>0.1171</td>
<td>0.0928</td>
<td>6</td>
<td>0.0686</td>
<td>0.0704</td>
</tr>
<tr>
<td>COD(_{c})</td>
<td>0.1012</td>
<td>0.0394</td>
<td>0.0724</td>
<td>7</td>
<td>0.0582</td>
<td>0.0844</td>
</tr>
<tr>
<td>BOD(_{s})</td>
<td>0.2305</td>
<td>0.0893</td>
<td>0.1646</td>
<td>3</td>
<td>0.1175</td>
<td>0.1863</td>
</tr>
<tr>
<td>NH(_{3-N})</td>
<td>0.0252</td>
<td>0.0357</td>
<td>0.0301</td>
<td>8</td>
<td>0.0492</td>
<td>0.0346</td>
</tr>
<tr>
<td>TP</td>
<td>0.0448</td>
<td>0.0672</td>
<td>0.0552</td>
<td>9</td>
<td>0.0414</td>
<td>0.0435</td>
</tr>
<tr>
<td>TN</td>
<td>0.0499</td>
<td>0.1580</td>
<td>0.0998</td>
<td>1</td>
<td>0.2114</td>
<td>0.1125</td>
</tr>
<tr>
<td>F(_{s})</td>
<td>0.0064</td>
<td>0.0442</td>
<td>0.0241</td>
<td>5</td>
<td>0.0811</td>
<td>0.0356</td>
</tr>
<tr>
<td>As</td>
<td>0.0041</td>
<td>0.0330</td>
<td>0.0176</td>
<td>10</td>
<td>0.0345</td>
<td>0.0160</td>
</tr>
<tr>
<td>S</td>
<td>0.0152</td>
<td>0.0268</td>
<td>0.0206</td>
<td>12</td>
<td>0.0226</td>
<td>0.0181</td>
</tr>
<tr>
<td>Cu</td>
<td>0.0938</td>
<td>0.0112</td>
<td>0.0553</td>
<td>4</td>
<td>0.0967</td>
<td>0.0949</td>
</tr>
<tr>
<td>Zn</td>
<td>0.0753</td>
<td>0.0388</td>
<td>0.0583</td>
<td>11</td>
<td>0.0282</td>
<td>0.0569</td>
</tr>
<tr>
<td>Pb</td>
<td>0.0167</td>
<td>0.0370</td>
<td>0.0262</td>
<td>13</td>
<td>0.0173</td>
<td>0.0169</td>
</tr>
<tr>
<td>Hg</td>
<td>0.0091</td>
<td>0.0027</td>
<td>0.0061</td>
<td>14</td>
<td>0.0125</td>
<td>0.0104</td>
</tr>
<tr>
<td>Cd</td>
<td>0.0215</td>
<td>0.0095</td>
<td>0.0159</td>
<td>15</td>
<td>0.0081</td>
<td>0.0163</td>
</tr>
<tr>
<td>pH</td>
<td>0.0473</td>
<td>0.0465</td>
<td>0.0469</td>
<td>16</td>
<td>0.0039</td>
<td>0.0303</td>
</tr>
</tbody>
</table>

### Table 5

<table>
<thead>
<tr>
<th>Weights</th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>SD</th>
<th>CV</th>
<th>Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHP</td>
<td>0.1</td>
<td>0.0041</td>
<td>0.2305</td>
<td>0.0655</td>
<td>1.0480</td>
<td>0.905</td>
</tr>
<tr>
<td>EWM</td>
<td>0.1</td>
<td>0.0027</td>
<td>0.2436</td>
<td>0.0632</td>
<td>1.0112</td>
<td>0.601</td>
</tr>
<tr>
<td>LightGBM</td>
<td>0.1</td>
<td>0.0039</td>
<td>0.2114</td>
<td>0.0573</td>
<td>0.9168</td>
<td>0.551</td>
</tr>
<tr>
<td>CW(_{AE})</td>
<td>0.1</td>
<td>0.0061</td>
<td>0.2141</td>
<td>0.0572</td>
<td>0.9152</td>
<td>0.960</td>
</tr>
<tr>
<td>CW(_{AL})</td>
<td>0.1</td>
<td>0.0104</td>
<td>0.1863</td>
<td>0.0551</td>
<td>0.8816</td>
<td>0.981</td>
</tr>
</tbody>
</table>

Kaiser-Meyer-Olkin Measure of Sampling Adequacy: 0.614. SD: Standard Deviation, CV: Coefficient of Variation.
4.2.1. Reliability analysis

In this study, the reliability of the model was determined by comparing and analyzing the water environment research results related to the Chaobai River basin. Yaqian et al. (2020) evaluated the water quality of the upper Chaobai River Basin in 2016, and the assessment result was "mild eutrophication", which was basically consistent with the results of this study. Yu et al. (2020) studied the water quality of the Chaobai River and gave the eutrophication level as a "medium harmful level", which was slightly worse than the results of this study. Wu et al. (2020) studied the North Canal, which was adjacent to the Chaobai River, and the team concluded that the water quality of the North Canal is "good", which was slightly better than the "slightly polluted" findings of WQI

Through multi-dimensional comparison, the water quality assessment results from the three models in this study had certain reliability and could be used as a reference for the study.

4.2.2. Eclipsing problem and accuracy analysis

The applicability of aggregation function and surface water assessment was judged by eclipsing problem and accuracy of the model in this study.
(1) Eclipsing problems of the models. The number of eclipsing problems occurring in the calculation of WQI by each aggregation function was shown in Table 6 below. The procedure for calculating the eclipsing problem was shown in Table S9. When the water quality was poor, the eclipsing problems of the WQI\(_S\) model appeared least, and the WQI\(_W\) model appeared least when the water quality was good. None of the underestimate eclipsing problems appeared in the model of WQI\(_L\) (Table 6). The least number of overestimate eclipsing problems occurred in the model of WQI\(_S\), with only 0 cases in Class I–II and 12 cases (36.36%) in inferior to Class II. In terms of the total number of eclipsing problems, WQI\(_S\) could be kept at a low level both when water quality was good and poor, while WQI\(_W\) had significant eclipsing problems when water quality was poor. Uddin et al. (2023a) considered the eclipsing problems of the IEWQI model and calculated a maximum eclipsing percentage of 21.4%. The results obtained by using WQI\(_W\) and WQI\(_S\) models in this study were similar to this. It indicated that both WQI\(_W\) and WQI\(_S\) models had low eclipsing problems when assessing surface water quality, while WQI\(_L\) had high eclipsing problems.

(2) Comparison of accuracy is an important part of assessing the strengths and weaknesses of a model. There are various methods for accuracy, Ding et al. (2022a, 2022b) determined the accuracy of the model by studying and analyzing the research results of others in the same region. In this study, the results of the model calculations were compared with the official water quality grades to verify the accuracy of the model.

(i) Analysis of water quality grade proportion. The percentage of each grade of water quality in the Chaobai River Basin in 2021 was shown in Fig. 7.

In Fig. 7a, 72.37% of water was Class I–II for the standard, 27.63% for inferior to Class II. And, 68.63% for Class I–II, and 31.37% for inferior to Class II in the water quality structure calculated by WQI\(_S\). The difference in the number of percentages between WQI\(_S\) and the Standard was only 3.74%. While, the difference in the number of percentages between WQI\(_W\) and the Standard was 20.28%, and the water quality structure calculated by WQI\(_L\) was all Class I–II. In Fig. 7, the accuracy of model in assessing water quality was ranked as: WQI\(_S\) > WQI\(_W\) > WQI\(_L\). The result of the comparison of the study was that the WQI\(_S\) model had the highest accuracy.

(ii) Analysis of differences. The reason for the difference between the water quality obtained in this study and officially published water quality in the Chaobai River Basin were: ① Assessment of water quality models was different. The official model for

Table 6
Numbers of sampling sites at which eclipsing problems occurred for WQI\(_W\), WQI\(_L\) and WQI\(_S\) models.

<table>
<thead>
<tr>
<th>Models</th>
<th>Underestimate eclipsing</th>
<th>Overestimate eclipsing</th>
<th>Total eclipsing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class I–II</td>
<td>Inferior to Class II</td>
<td>All</td>
</tr>
<tr>
<td>WQI(_W)</td>
<td>6 (3.51%)</td>
<td>0 (0.00%)</td>
<td>6 (6.67%)</td>
</tr>
<tr>
<td>WQI(_L)</td>
<td>0 (0.00%)</td>
<td>0 (0.00%)</td>
<td>0 (0.00%)</td>
</tr>
<tr>
<td>WQI(_S)</td>
<td>40 (23.39%)</td>
<td>0 (0.00%)</td>
<td>40 (23.39%)</td>
</tr>
</tbody>
</table>

The number of all samples was 204, the number of Class I–II was 171 and the number of inferior to Class II was 33. The number in brackets indicated the percentage of eclipsing problems that occurred.
assessment was not published, and the study took the new combination of assessment models. A few sampling sites were different. Some sampling sites announced by Beijing were not sampled in this study, and this study also counted Tianjin and Hebei (Fig. 8).

Combined with the above analysis, the WQI_{S} model not only had low eclipsing problems but also had high accuracy. The WQI_{W} had very low eclipsing problems in the condition of good water quality. The WQI_{L} model had high eclipsing problems and had low accuracy.

4.2.3. Uncertainty analysis

In this study, the standard uncertainty (SU) was calculated for the results of three models to evaluate water quality grades. The smaller the standard uncertainty obtained from the calculation, the better the model performed. The analysis process of uncertainty included four parts: (1) features of the distribution of model results; (2) the determination of the standard water quality grade; (3) calculation of model uncertainty; (4) hypothesis testing of the model uncertainty. The specific analysis process are as follows:

1. Features of the distribution of model results. PDF plots and cumulative distribution function (CDF) for the WQI results of the three models and the standard were shown in Fig. 9, and the details of the PDFs were shown in Table S10.

The WQI results of all three models were normal distribution (Fig. 9a). It showed that all the PDF plots of WQI showed a "bell-shaped" curve (high in the middle and low at both ends) in Fig. 9a. The mathematical expectation (the horizontal coordinate of the highest point of the bell curve) of WQI_{S} was close to the standard, but the standard deviation was larger (the bell curve showed short and fat). The mathematical expectation of WQI_{L} deviated from the standard the most, but the standard deviation was the smallest. Fig. 9b showed that WQI_{S} was closer to the standard when the water quality was poor, and WQI_{W} was closer to the standard when the water quality was good. WQI_{L} deviated from the standard more.

2. The determination of the standard water quality grade. Standard water quality grades are determined as follows: (i) Research reference the recommended guideline values of Uddin et al., (2022), and combined with our water quality standards division method, increased the value of Class IV, Class V, and inferior to Class V grades. The determined criteria for the guideline of water grades in the WQI model was shown in Table 7 below. (ii) Determine the number of monitoring values of water quality indicators that...
(iii) Based on Table 7 determined the standard water quality grade for each water sample. According to the Eq. (18) for the model standard uncertainty solution, three models calculated standard uncertainty results were shown in Table 8. The details of the grade results were shown in Table S10.

All three models had low uncertainty (Table 8). The magnitude of uncertainty in numerical terms for the three models was: \( WQI_W \) < \( WQI_L \) < \( WQI_S \). The maximum difference between the total SU results of \( WQI_L \), \( WQI_W \) and \( WQI_S \) was 0.117, indicated that the uncertainties of the three models were close. In the condition of poor water quality (inferior to Class II), \( WQI_S \) had lower uncertainty. In the condition of good water quality (Class I–II), \( WQI_W \) had lower uncertainty. Overall, the \( WQI_S \) model was recommended for assessing poor water quality and the \( WQI_W \) model was more recommended when assessing good water quality.

### 4.3. Spatial and temporal analysis of water quality

#### 4.3.1. Features of water quality

In this study, \( WQI_L \) was selected for the spatial and temporal analysis of water quality in the Chaobai River Basin, and the map of the basin was

**Table 7**

<table>
<thead>
<tr>
<th>Water grades</th>
<th>Quantification of water grades</th>
<th>Expected number of breaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Class II</td>
<td>2</td>
<td>1–2</td>
</tr>
<tr>
<td>Class III</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Class IV</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Class V</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Inferior to Class V</td>
<td>6</td>
<td>&gt;5</td>
</tr>
</tbody>
</table>

The term “criteria” referred to the threshold value of standard guidelines. Details are given in Table 2 above.

**Table 8**

<table>
<thead>
<tr>
<th>Models</th>
<th>Aggregation function</th>
<th>Sub-index</th>
<th>Sample size</th>
<th>Grade mean</th>
<th>SU Inferior to Class II Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( WQI_w )</td>
<td>WQM</td>
<td>DLIR</td>
<td>2.020</td>
<td>0.259</td>
<td>0.502</td>
</tr>
<tr>
<td>( WQI_L )</td>
<td>LQM</td>
<td>EF</td>
<td>1.255</td>
<td>0.322</td>
<td>0.174</td>
</tr>
<tr>
<td>( WQI_S )</td>
<td>SWM</td>
<td>ISF</td>
<td>2.309</td>
<td>0.397</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Table 9**

<table>
<thead>
<tr>
<th>Models</th>
<th>( t )-value</th>
<th>P</th>
<th>Mean</th>
<th>SE</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>( WQI_w )</td>
<td>80.723</td>
<td>0.000***</td>
<td>2.020</td>
<td>0.025</td>
<td>0.357</td>
</tr>
<tr>
<td>( WQI_L )</td>
<td>41.027</td>
<td>0.000***</td>
<td>1.255</td>
<td>0.031</td>
<td>0.437</td>
</tr>
<tr>
<td>( WQI_S )</td>
<td>69.621</td>
<td>0.000***</td>
<td>2.309</td>
<td>0.033</td>
<td>0.474</td>
</tr>
</tbody>
</table>

\*, **, *** represent 1%, 5%, 10% significance level respectively.

4. Hypothesis testing of the model uncertainty. Hypothesis testing which used \( t \)-test determined whether the model was credible. The statistical results of the \( t \)-test of the model were shown in Table 9.

In Table 9, the \( t \)-values of the three models ranged from 41.027 to 80.723, and the model significance p-values were all 0.000***. Uddin et al. (2023a) calculated the statistical range of \( t \)-values for IEWQI in various domains from 40.64 to 194.27, confirmed that the IEWQI model could be a good assessment of seawater quality. The \( t \)-value results of this study were within the range of the results of Uddin et al. (2023a), indicated that the uncertainty results of all three models were credible.

According to the reliability, uncertainty, eclipsing problems and accuracy analysis: All three models had good reliability. In reducing the uncertainty of the models, \( WQI_S \) was applicable for assessing poor water quality, \( WQI_W \) was applicable for assessing good water quality. Both \( WQI_S \) and \( WQI_W \) models had low eclipsing problems, while \( WQI_L \) had high eclipsing problems. \( WQI_S \) model also had the highest accuracy. Overall, the \( WQI_L \) model was recommended for assessing poor water quality, and the \( WQI_W \) model was more recommended when assessing good water quality.

### 4.3. Spatial and temporal analysis of water quality

#### 4.3.1. Features of water quality

In this study, \( WQI_L \) was selected for the spatial and temporal analysis of water quality in the Chaobai River Basin, and the map of the basin was

*Fig. 9. PDF and CDF plots of WQI models calculated by different aggregation functions and standard WQI (including 95% confidence level of the probability functions, the solid line indicated the calculated WQI value and the dashed line indicated the standard WQI value. The standard was calculated by the mean of the interval).*
drawn with the water quality level of each sampling site, which was shown in Fig. 10.

1. Spatial distribution features. It can be seen the upstream water quality was generally better than the downstream in Fig. 10. Since site 10 was located at Miyun Reservoir, the water quality around site 10 was relatively better. The spatial distribution of water quality in the basin shows that the water quality downstream at site 10 is worse than that upstream at site 10, which was related to the topography of the Chaobai River Basin. According to the DEM map shown in Fig. 10, it showed that the overall topography of the Chaobai River Basin was high in the northwest and low in the southeast. Moreover, sites 13 and 14 were located in an intensive agricultural area, and the agricultural wastewater might cause the water quality to be worse than the upstream (Zhang et al., 2021). While the upstream area was mostly mountainous with less anthropogenic agricultural wastewater, so the water quality was better than the downstream all year round (Zheng et al., 2013).

2. Temporal distribution features. Temporally, the water quality at site 10 was Class I in January, while the water in the remaining months
was Class II. The overall water quality of the basin was better in October, which was only in site 17 and site 13 for Class III, the rest water quality of the sampling points was Class II. Water quality was similar in April, September, November and December. The worst water quality in the basin was in July, with a higher number of Class III water compared to other months. Zhao et al. (2021) found that the recharge of reclaimed water was greatest in July and August. And Li et al. (2017) showed that the recharge of reclaimed water to the river brought nitrogen and phosphorus, and other ions, which might contribute to poor water quality in the basin. It was also possible that the pollution in July was related to the rainy season (Xu et al., 2009), which caused surface source pollution in the watershed and made water quality worse. Therefore, it was speculated that either reclaimed water recharge or rainfall might cause the worst water quality in July, but further research was needed to determine which cause was responsible.

3. Analysis of correlation. We plotted the Pearson correlation coefficient heat map based on the collected water quality data combined with rainfall and temperature. The rainfall information was obtained from Beijing Water Authority and the temperature information was obtained from China Meteorological Administration. The Pearson correlation coefficient results were detailed in Fig. 11. From the Pearson correlation graph, we could also see that DO had a strong negative correlation with rainfall and temperature, TN had a strong negative correlation with temperature, and COD$_{Mn}$, COD$_{Cr}$, BOD$_5$ had a strong correlation with each other. Nitrogen had a relatively strong positive correlation with phosphorus, and nitrogen also showed a negative correlation with temperature, which was consistent with the results of Zhao et al. (2021). Weak correlations were observed between metals, and metalloid (As) had strong positive correlations with organics and nitrogen.

4.3.2. Spatio-temporal analysis of nitrogen

1. Spatial distribution features. According to the box plot of water quality parameters in Fig. 3 above, it can be seen that the key pollutant in the basin is TN with slight pollution of NH$_3$-N. This indicated a certain risk of eutrophication in the water body, which is consistent with the results of Zhao et al. (2021). The team hypothesized that the source of total nitrogen could be poor water quality due to high background values of nutrients and minerals in the wastewater treatment plant return water. Fig. 12g–12h showed that ammonia nitrogen was more polluted downstream than upstream; the pollution of total nitrogen far exceeded the Class III water standard in terms of spatial distribution. The high ammonia nitrogen downstream might be due to the land structure downstream (Su et al., 2021). Where a large amount of nitrogenous pollutants were imported into the water body might be because of surface pollution caused by agricultural fertilizers, resulting in higher ammonia nitrogen concentrations downstream. The lower concentration of total nitrogen at sites 11 and 12 might be because of the occurrence of denitrification microreactions (Xia et al., 2022). Due to the confluence of the Jian River and the Chaobai River (the channel was widened, the flow slowed, and phytoplankton increased), a favorable anaerobic microenvironment was created to denitrification, reducing nitrate nitrogen and thus total nitrogen levels: $2\text{NO}_3^- + 10\text{H}^+ (\text{Hydrogen donor-Organic}) \rightarrow \text{N}_2 + 2\text{OH}^- + 4\text{H}_2\text{O}$. The nitrogen community mainly includes nitrate nitrogen, nitrite nitrogen, ammonia nitrogen, and organic nitrogen. It could be found in the data of Zhang et al. (2021) that

Fig. 11. Heat map of Pearson correlation coefficient for water quality parameters.
the Chaobai River had higher ammonia nitrogen downstream and almost no nitrate nitrogen and nitrite nitrogen pollution, which was consistent with the results in Fig. 12g. It indicated that there might be organic nitrogen pollution at the upstream, while there was organic nitrogen pollution and some ammonia nitrogen pollution at the downstream. Overall, recycled water and phytoplankton might be the causes of nitrogen pollution spatially.

2. Temporal distribution features. In Fig. 12g–12h, it could be seen that ammonia nitrogen pollution was mainly in July, and total nitrogen pollution was not differentiated in terms of the month. The Chaobai River had a larger recharge of reclaimed water in July and August, and Zhao et al. (2021) thought that the reclaimed water from the treatment plant might contain nutrients and minerals. Li et al. (2017) also found that the recharge of reclaimed water made the nitrogen level increase, which confirmed that the recharge measures were also a source of nitrogen. Chen et al. (2022) found that atmospheric deposition was a non-negligible source of total nitrogen in aquatic ecosystems, which contributed to the increased concentration of nitrogen in summer. Overall, the main causes of temporal pollution of nitrogen might be surface source pollution and atmospheric deposition.

4.3.3. Spatio-temporal analysis of metal and metalloid

Fig. 12 could be seen the gap between the various types of metals and the Class I water quality standard, in which the concentrations of Pb and As were lower than the Class I standard.

1. Spatial distribution features. Both Zn in Fig. 12b and Cd in Fig. 12d had exceeded Class I contamination at site 2 and site 4. Yu et al. (2013) studied the soils along the Chaobai River channel and found that Hg exceeded Class I standards and hypothesized that the exceedance of Hg mainly came from the development of surrounding mineral resources. No exceedance of Hg was found in the water quality of this study, but it was also possible that the concentration of Zn and Cd exceeded the Class I standard with the exploitation of mineral resources. Ji et al. (2016a, 2016b) pointed out the presence of metal deposits at the left tributary of the upper Chaobai River, which further confirmed that metal pollution might come from underground metal deposits.

2. Temporal distribution features. Fig. 12a showed that the concentration of Cu exceeded the Class I standard only at site 14, and the months with relatively large concentrations were May and August, which might be related to the period of abundant water (May to September). Yu et al. (2013) speculated that Cu may come from coal combustion. It might also be related to the type of crops and the way they were grown. When there was in the rainy season, the rain brought heavy metals from the soil into the river, causing heavy metal pollution of the river. The pollution of Zn occurred in October, and the pollution of Cd occurred in July and October. When the groundwater recharged to the river during the dry period, it brought heavy metal pollution. In Fig. 12c, the monthly pollution of Hg was inferior to the Class I standard and the annual cumulative pollution concentration exceeded Class I. The pollution of Hg probably originated from the soil in the basin (Yu et al., 2013).
4.3.4. Spatio-temporal analysis of organic

1. Spatial distribution features. Fig. 3 showed that the average value of COD$_{10t}$ concentration at site 15 exceeded the Class III water standard; the average value of COD$_C$ concentration at sites 13, 14, 15, 16, and 17 exceeded the Class III water standard; and the average value of BOD$_5$ concentration at sites 14 and 15 exceeded the Class III water standard. Organic pollution is more severe downstream and less severe upstream. Zhang et al. (2021) found that anthropogenic inputs, active microorganisms, and phytoplankton influence the level of organic pollution in Chaobai River. Su et al. (2021) found that the land structure in the downstream contributed to water quality changes. The downstream land in the Chaobai River Basin was mainly farm and farmland, while the upstream was mainly villages and towns, which caused more organic pollution downstream than the upstream. Overall, land structure, human activities, microorganisms, and phytoplankton all might contribute to organic pollution.

2. Temporal distribution features. Fig. 11 showed that there was a strong positive correlation between COD$_{10t}$, COD$_C$, BOD$_5$, temperature, and rainfall, and they had a negative correlation with DO. The water quality of the Chaobai River was stable except during the rainy season (July) in Fig. 10. Among them, local organic carbon was mainly derived from organic matter in the soil, and related to the local climatic characteristics. When precipitation was high, the organic matter in the soil sank into the Chaobai River system through runoff, resulting in poor water quality during the rainy season (Lu et al., 2013). This might also be related to the full ecological recharge of the Chaobai River in April 2021 from reservoirs such as Miyun Reservoir, Huairou Reservoir, and the South-North Water Transfer Project (Gao et al., 2022a, 2022b). Overall, surface source pollution and ecological recharge caused higher concentrations of organic pollution during the rainy season (July-August).

Overall, we concluded that the water quality of Chaobai River Basin was good, and the water quality varied upstream and downstream as well as climate change, with the overall performance of the upstream water quality being better than the downstream water quality, and the worst water quality during the July and August. The overall water quality in the basin as a whole was “slightly polluted” Class II water quality, and the key pollutant was TN. Pollution of nitrogen might come from the atmosphere, soil, and reclaimed water. Metals were less polluted, exceeded the Class I standard (less than Class III standard) in only a few months, and were less than the Class I standard in the rest. Soil, coal combustion, and nearby metal deposits could be sources of metals and metalloids.

5. Conclusion

The weights and aggregation functions are important structures of the WQI model and determine the model accurately. Machine learning was used to calculate the weights in this study. The combined weights of LightGBM and AHP were established by game theory and two aggregation functions were proposed to optimize the WQI model. The Chaobai River Basin was used as a study case for the new water quality assessment system. The conclusions of this study are as follows:

(1) To determine the optimal weights, the KMO extraction values and CV values of weights were compared in this study. The combined weights CW and CW$_A$ had larger KMO extraction values of 0.981 and 0.960, respectively. The CW$_A$ had well interpretability. The CV value of the combined weight CW$_A$ was 0.8816, which had the highest relative stability and better ranking fitting. Overall, the combined weight CW$_A$ was the optimal combination weight.

(2) The analysis results of reliability, uncertainty, eclipsing problems and accuracy showed that: All three models had good reliability. The WQIs, WQI$_S$, and WQI$_W$ models had low uncertainty of 0.474, 0.437 and 0.357 respectively. The WQIs had low uncertainty (0.000) in assessing poor water quality, and WQIs had low uncertainty (0.259) in assessing good water quality. Both WQIs and WQIs$_W$ models had low eclipsing problems (25.49% and 16.83%). The accuracy of the models was ranked as WQIs > WQIs$_W$ > WQIs.

(3) The improved WQI models were better in reducing the uncertainty and improving the accuracy of the model, which used machine learning, combined weights and aggregation function. The results of the analysis showed that the WQIs model was recommended for assessing poor water quality, and the WQIs$_W$ model was recommended for assessing good water quality.

(4) The assessment results showed that the Chaobai River Basin was assessed as Class II, “slightly polluted”. In the water quality of the basin, the content of TN exceeded the standard, nitrogen pollution might come from the atmosphere, soil, and reclaimed water. Metal pollution was slight, only a few months exceeded the Class I standard (less than the Class III standard), and the rest were less than the Class I standard, for metals and metalloids pollution could come from the soil, nearby metal deposits, or coal combustion. The overall water quality, upstream was better than that downstream, and the water quality was the worst in July and August.

This study proposed an optimized WQI model, and the new assessment system established can provide a reference for other decision-makers and researchers. Due to the limitation of monitoring data, the transformation of families of identifiable chemical substances was not studied enough. For example, the nitrogen groups in this assessment indicators were only ammonia nitrogen and total nitrogen, and the assessment of other forms of nitrogen was not conducted. If monitoring data was complete, an in-depth analysis of the interconversion of different forms of nitrogen could be assessed. In addition, extensive experiments are needed to assess the sensitivity of the model to these variables.

Declaration of Competing 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.

Data availability

The data that has been used is confidential.

Acknowledgments

This study is supported by the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (32004019201501).

Supplementary materials


References

