

Research on river water quality evaluation based on the GA-PP and improved fuzzy model

Zhenggang Huo^{a,*}, Xiaoting Zha^b, Yuhong Chu^c, Mengyao Lu^a and Sensen Zhang^a

^a College of Civil Science and Engineering, Yangzhou University, Yangzhou 225127, China

^b Jiangsu Hydraulic Engineering Construction Co., Ltd, Yangzhou 225127, China

^c GenChem & GenPharm (Changzhou) Co., Ltd, Changzhou 213033, China

*Corresponding author. E-mail: zganghuo@126.com

ABSTRACT

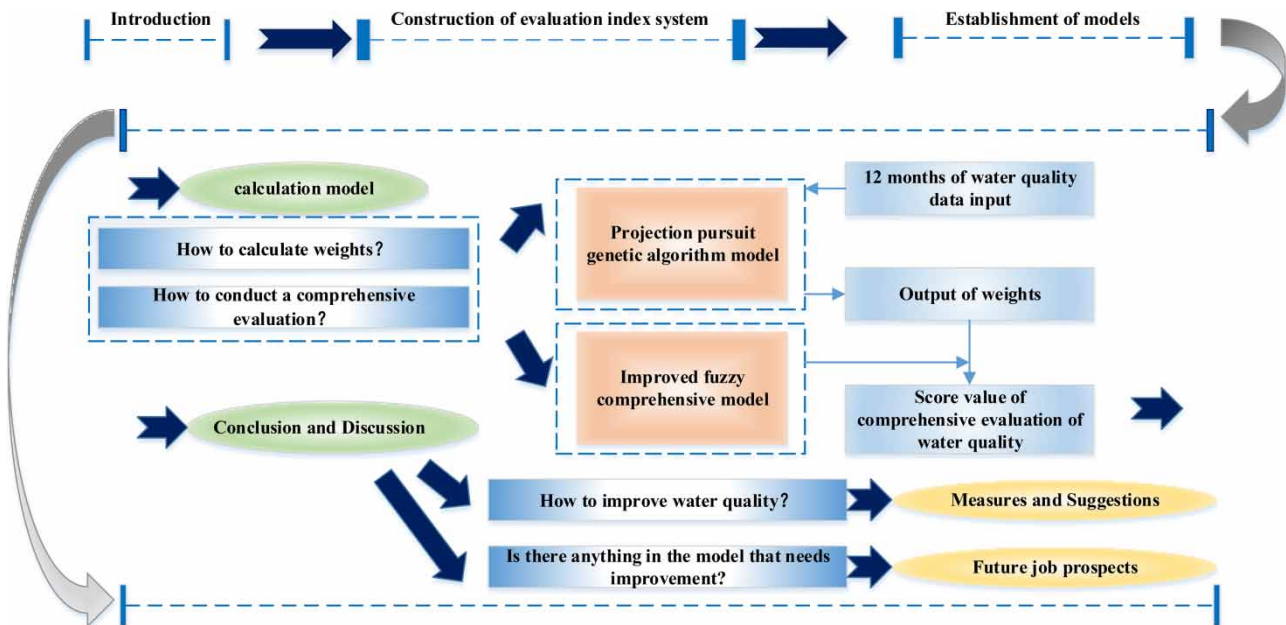
To ensure the water quality of rivers, it is crucial to scientifically evaluate their water quality status. This study takes a river in Jiangsu, China, as an example to establish six targeted main indicators for river water quality evaluation and uses a projection pursuit model optimized by the genetic algorithm to determine weights. Applying the improved fuzzy evaluation model to the final evaluation of water quality, the results indicate that this article adopts a weight calculation model that reduces dimensionality without losing data features, and the comprehensive evaluation model is also more complete, resulting in more accurate evaluation results. According to model analysis, the summer water quality is good and peaks from June to July. This article proposes corresponding measures and suggestions in response to the reasons behind this seasonal change. The evaluation model used in this article is superior to other models in terms of accuracy and portability, making it an excellent choice for river water quality evaluation. It can provide valuable technical guidance for similar river water quality evaluations.

Key words: improve fuzzy evaluation model, projection pursuit, river water quality

HIGHLIGHTS

- In the evaluation of river water quality, the high-dimensional data dimensionality reduction operation optimized by the genetic algorithm projection pursuit (GA-PP) is not easy to lose the original features of the data and is quite suitable for calculating weights.
- The improved fuzzy evaluation model is also superior to the traditional fuzzy comprehensive model and can more comprehensively evaluate water quality data.

GRAPHICAL ABSTRACT



Research on River Water Quality Evaluation Based on GA-PP and Improved Fuzzy Model

1. INTRODUCTION

In 1983, it was approximated that there were around 50,000 rivers in China. Nevertheless, when China conducted its inaugural national water census in 2013, approximately 28,000 of these rivers were found to be missing. This significant decline can be attributed to rapid population and economic growth, excessive exploitation of water resources, and pollution. Since 2003 when Changxing County in Zhejiang Province has taken the lead in implementing the *river chief system*, river management in China has been taken to an unprecedented level. Evaluating river water quality is an essential basic work in water resources and environment management through the rational evaluation of water quality monitoring data to formulate scientific remediation planning to take adequate measures.

Since the 1970s, numerous scholars have been actively researching the water quality of rivers, leading to the development of water quality evaluation indicators. In essence, the concentration of various physical and biochemical variables in a river should be maintained within specific limits. If these levels surpass certain established standards, it is deemed detrimental to the environment (Banda & Kumarasamy 2020). The indicators that measure the level of these physical and chemical substances and the biological parameters that affect them are the water quality indicators (WQIs). According to Poonam *et al.* (2013), Animashaun *et al.* (2023), Sutadian *et al.* (2016) and Paun *et al.* (2016), there is no standardized design of WQIs, and WQIs are supposed to be different in different areas and for different uses, and therefore, WQIs need to be based on the local water quality and the status of the relevant uses. In terms of special uses of water, Almeida *et al.* (2012) established swimming water grade evaluation indexes from pH, chemical oxygen demand (COD), nitrate, phosphate, detergent, enterococci, total coliforms, fecal coliforms, and *E. coli*. Kim *et al.* (2014) in their study of water quality improvement measures for agricultural drains used suspended solids, turbidity, biochemical oxygen demand (BOD), COD, nitrogen, phosphorus, and fecal coliform as evaluation indexes to evaluate the water quality of drains. A supplemental mixing system was developed to limit this pollution, and after the improvement, suspended solids in drains were reduced by 10.0–98.3%, turbidity was reduced by 25.2–98.4%, BOD was increased by 21.1–91.1%, COD was increased by 19.2–75.4%, T-N was reduced by 21.0–7.3%, T-P was reduced by 5.9–91.2%, and fecal coliform decreased by 35.7–97.6%. In different regions of the world, Rubio-Arias *et al.* (2013) selected acid-base (pH), electrical conductivity (EC), dissolved oxygen (DO), temperature (T), turbidity, total dissolved solids (TDS), total hardness (TH), and chloride (Cl⁻) as the physico-chemical variables in the study of water quality in the Mexican dams. Sharifinia *et al.* (2016) used biological indices such as macroinvertebrate

populations, the river bottom diatom index, and the Hilsenhoff bioindex to evaluate the ecological health of the Shahrud River. The results showed that the water quality of benthic diatoms and macroinvertebrate populations was 'fair' in the case of phosphate concentration and heavy metals, and the health status was 'quite severe' in the case of organic pollution.

Other scholars, on the other hand, focus on studying water quality issues from the perspective of evaluation modeling. In terms of subjective evaluation, *Munyika et al. (2014)* judged river water quality subjectively from expert scoring and used the South African River Scoring System to assess the current water quality and the overall health of the Orange River in Namibia. The evaluation was categorized into five grades, and the results showed that the Orange River in Namibia was in health class C, and the water quality condition of the river was within the acceptable range. This type of method relies too much on the experience level of the evaluator. For objective evaluation, *Zhang & Sihui (2009)* established a new water quality evaluation model based on the projection pursuit (PP) technique. In order to improve the accuracy of the model, a genetic algorithm is used to optimize its parameters, and examples show that the model can evaluate the water quality appropriately. It can reduce the dimensionality without losing the data characteristics and objectively determine the indicator weights based on the available data, which is difficult to do with other assessment methods. *Xu et al. (2021)* used principal component analysis (PCA) to determine the weights combined with fuzzy comprehensive evaluation (FCE) and evaluated the water quality of four lakes in China. They found that the water quality of these four lakes was better in spring and summer than in fall and winter. Class I water quality was the highest in spring at 28.9% and the lowest in fall at 25.1%. *You et al. (2021)* used improved FCE to evaluate the water quality of the Yangjiabao aquaculture base. The worst water quality in the Yangjiabao aquaculture base was in winter. Meanwhile, the evaluation results of the single-factor evaluation method and the traditional fuzzy evaluation method were compared and compared with the other two methods, the improved method can comprehensively reflect the changes in water quality problems.

In summary, water quality evaluation is one of the hotspots in current river ecology research. The selection of river WQIs needs to be based on the actual situation of the specific river, and the expert scoring method relies too much on the professionalism of the evaluator, so objective evaluation becomes very important. The novelty of this article lies in the weight calculation adopting genetic algorithm optimized projection pursuit (GA-PP), which means it is not easy to lose the original features of high-dimensional data in dimensionality reduction operations, thus ensuring the accuracy of the weight. The improved fuzzy evaluation model is also superior to the traditional fuzzy comprehensive model and can comprehensively evaluate water quality data, thus ensuring the comprehensiveness of the evaluation. The organic combination of the two algorithms makes river evaluation closer to reality.

2. CONSTRUCTION OF THE EVALUATION INDEX SYSTEM

Combing through the relevant literature (*Huang & Lu 2014; Liu et al. 2018; Liu et al. 2019b*), the indicators usually selected in the evaluation of river water quality are divided into three categories: ① common indicators: water temperature, PH, DO, 5-day biochemical oxygen demand (BOD₅), permanganate index (COD_{Mn}), chemical oxygen demand (COD_{Cr}), total nitrogen (TN), ammonia nitrogen (NH₃-N), total phosphorus (TP); ② toxicity indicators: copper (Cu), zinc (Zn), cadmium (Cd), chromium (Cr), lead (Pb), volatile phenols (Fb), lead (Pb), and other toxicity indicators, volatile phenols (FN), oils (OILS), selenium (Se), mercury (Hg), sulfide, cyanide, fluoride, anionic surfactants; ③ biological indicators: fecal coliform and other suspended solids.

The general principle for selecting river WQIs should be based on their potential threats to water environment pollution, human health, and socio-economic development. Through on-site monitoring of the Zhouxiang River in Yangzhou City, it was found that non-metallic ions, such as selenium, arsenic, and mercury, heavy metal ions such as copper and zinc, sulfides, petroleum, and volatile phenols all meet first-class standards, and do not pose a threat to the environment or human health. In addition, the river's water temperature and pH values are within an acceptable range. Therefore, these indicators can be ignored in the water quality evaluation of this river.

In summary, the water quality evaluation of the Zhouxiang River in this paper selected six evaluation indicators, respectively, DO, permanganate index (COD_{Mn}), BOD₅, TP, TN, and ammonia nitrogen (NH₃-N). See [Table 1](#) for details.

3. CALCULATION MODEL CONSTRUCTION

3.1. Weight calculation model

3.1.1. PP fundamentals

The essence of evaluation is the dimensionality reduction of multi-dimensional information (*Jia et al. 2022*) and the dimensionality reduction operation will inevitably lead to information distortion. Traditional evaluation methods, such as the

Table 1 | Water quality indicators, units, and analysis methods

Parameters	Abbreviations	Units	Analytical methods	Lowest detected limit
Dissolved oxygen	DO	mg/L	Iodimetry	0.20
Chemical oxygen demand	CODMn	mg/L	Potassium permanganate method	0.50
Biochemical oxygen demand	BOD5	mg/L	Dilution and inoculation test	0.40
Total phosphorus	TP	mg/L	Spectrophotometry	0.01
Total nitrogen	TN	mg/L	Ultraviolet spectrophotometry	0.05
Ammonia nitrogen	NH3-N	mg/L	N-reagent colorimetry	0.05

entropy weighting method, use simple dimensionality reduction operations to analyze the data, and it is difficult to make a complete and accurate evaluation. The PP method is a kind of exploratory emerging data processing model used to analyze and dimensionality reduction of multivariate data (Huber 1985). The basic idea of PP is to construct the corresponding projection index function, in order to observe the data from multiple perspectives, project the high-dimensional data to the low-dimensional subspace from different perspectives and find the optimal projection direction by maximizing or minimizing the projection function, from which the original high-dimensional data can be viewed from different perspectives. The best projection direction is found by maximizing or minimizing the projection function, from which the original high-dimensional data can be viewed, and the original high-dimensional data features can be obtained to the greatest extent or the information of the data can be fully explored.

Let X denote the original high-dimensional data $p \times n$ matrix to be studied, among them, p and n , respectively, represent the number of samples and the dimension of high-dimensional data. The projection matrix or projection direction matrix A is represented by the $k \times p$ matrix with rank k , and in order to project from the high-order to the low-order, it is necessary to have $k < n$; the uppercase letter Z denotes the linear projection of the n -dimensional high-order data in the k -dimensional low-order space, and this linear projection is expressed mathematically as (Liu *et al.* 2019a):

$$Z = A * X, A \in R^k, X \in R^n \quad (1)$$

Let the original high-order data matrix X obey the feature distribution F ; The projection direction matrix Z follows the feature distribution F_A . Specifically, when the lower-order dimension $k = 1$, A becomes a single-column matrix a (the entire study focuses on single-column projection), F_a represents the one-dimensional feature distribution of Z . Mathematically, it can be proven that the feature function $\varphi(a)$ of the one-dimensional projection feature distribution F_a in the projection direction a is equivalent to the feature function $\varphi(F)$ of F , the projection in any direction a , the equivalence relation of the characteristic function of the linear projection in mathematical form can be expressed as (Barcaru 2019).

$$\varphi(F) = \varphi(F_a) \quad (2)$$

where $\varphi(F)$ is the function corresponding to the feature distribution of F , $\varphi(a)$ is the feature function of one-dimensional projection F_a . The above formula proves that the feature quantity is not lost after the high-dimensional data are downgraded and can still be represented in low dimensions, which is the mathematical basis for the realization of the basic idea of PP. Let $a = (a_1, a_2, \dots, a_n)$ be a single-column projection direction, the mathematical meaning of PP is to project $x(i, j)$ onto a , and get the one-dimensional projection value $z(i)$:

$$Z(i) = \sum_{j=1}^n a(j)x(i, j), i = 1, 2, \dots, n \quad (3)$$

3.1.2. Construct projection index function

Construct a projection metric function $Q(a)$ to project the high-dimensional data onto the lower dimensions. The requirement of the metric function is to distinguish certain features of the high-dimensional data better, firstly, the linearly projected points should be as many as possible in a certain local range, and secondly, the point clusters need to be distributed

as far as possible from each other. As D_z represents the projection value $Z(i)$ of the local density, S_z represents the projection value $Z(i)$ of the standard deviation, the projection indicator function $Q(a)$ mathematical significance is $S_z \times D_z$ to the maximum value, the construction formula (Huang & Bai 2018) is shown in the following equations:

$$S_z = \left[\left(\sum_{i=1}^n (Z(i) - E(z))^2 \right) / (n - 1) \right]^{1/2} \quad (4)$$

$$D_z = \sum_{i=1}^n \sum_{j=1}^n (R - r(i, j)) \times u(R - r(i, j)) \quad (5)$$

$$r(i, j) = |z(i) - z(j)| \quad (6)$$

where $E(z)$ is the arithmetic mean of the sequence of projection values $Z(i) | i = 1, 2, \dots, n$; R is the radius of the window of the local density of the projection distribution, the selection principle of R is that the sample points within the window cannot be too few, at the same time, when the number of sample points increases, the number of sample points within the window cannot be increased too much; there is no systematic theoretical basis to determine the width of the window. There is no systematic theoretical basis to determine the width of the window. The commonly used density window width recommended by Friedman is 10% of the variance of the projected eigenvalues of all samples, which is $0.1S_z$ (Friedman 1987), $r(i, j)$ is the distance between the clusters of dots and the clusters of dots; $u(R - r(i, j))$ is the Heaviside step function, the value of the function is 1 when $R - r(i, j) \geq 0$, and the value of the function is 0 when $R - r(i, j) < 0$.

3.1.3. Genetic algorithm model

Real-coded genetic algorithms have high efficiency, fast calculation speed, and high accuracy compared to traditional binary coding algorithms. This article uses real coded genetic algorithms for coding, and the specific implementation steps for its calculation are as follows:

- (1) Using a formula to perform a linear transformation on the data to be encoded, corresponding $x(j)$ on the interval $[a, b]$ to $y(j)$ on $[0, 1]$;

$$x(j) = a + y(j)(b - a), j = 1, 2, \dots, p \quad (7)$$

- (2) Generate a random population variable $y_i^0 (y_{i1}, y_{i2}, y_{i1} \dots y_{in})$ within the interval of $[0, 1]$, where i represents the sequence number of chromosomes, typically taking values of $[1, N]$, and N represents the population size. n represents the number of variables. 0 expresses primary chromosomes.
- (3) Calculate the objective function value, substitute the first generation chromosomes in step 2 into the objective function to be solved, calculate the function value, and sort the chromosomes according to the size of the function value to form y_i^1 .
- (4) Calculate the fitness evaluation function value based on order, where r represents the evaluation value of chromosome selection with sequence number 1 (less than 1).

$$P(y_i) = r(1 - r)^{i-1}, i = 1, 2, \dots, N \quad (8)$$

It can be verified that when $i = 1$, it indicates the best chromosome evaluation, and when $i = n$, it indicates the worst chromosome evaluation.

- (5) Calculate the cumulative probability of each chromosome being selected for y_i^1 (the cumulative probability is the sum of the probabilities of the sequence number i and all individuals preceding it).

$$P_i = \sum_{j=1}^i p(y_j) / \sum_{j=1}^N p(y_j), i = 1, 2, \dots, N \quad (9)$$

when selecting, a random number rnd is generated in $[0, 1]$. If the situation is like $p_{i-1} < rnd \leq p_i, p_0 = 0$, individuals with the sequence number i are selected to replicate into the next-generation population.

- (6) Repeat step 5 until the population size is reached, and chromosomes are replicated to form the next-generation chromosome population y_i^2 .

- (7) Cross-operation, generating population y_i^5 , given p_c as the crossover probability, for individual i , generating a random number rnd in $[0, 1]$, if $p_c < rnd$. Then select chromosome i for crossover operation. For two or more selected matrices, randomly select the same location for information exchange. After the exchange, a new gene combination was obtained and a portion of each parent was inhibited.
- (8) Mutation operation, generating population a , given p_t is the probability of variation, for an individual i , generate a random number rnd in $[0, 1]$, if $p_t < rnd$. Then select the chromosome i for mutation operation. Let the mutated individual be y'_i , and according to Michalewicz's step size formula, derive the mutated individual as

$$y'_i = \begin{cases} y_i + (1 - y_i) \times f(g) & \text{if random}(0, 1) = 0 \\ y_i + (y_i) \times f(g) & \text{if random}(0, 1) = 1 \end{cases} \quad (10)$$

$$f(g) = r_2 \times (1 - g/g_{\max})^\lambda \quad (11)$$

where r_2 is a uniform random variable between $[0, 1]$, g is the current evolutionary generation, and g_{\max} is the maximum evolutionary generation, λ . The shape factor of the curve representing the step size decreasing with algorithm generations is a fixed value.

3.1.4. PP genetic algorithm model

When the sample set is given, the value of the projection metric function $Q(a)$ changes only with the projection direction a . Different projection directions can project different data features, and the optimal projection direction is the one that maximizes the possibility of exposing a certain type of feature structure of the high-dimensional data. The optimal projection direction is estimated by solving the maximization problem of the projection metric function. The formula (Lan & Huang 2018) is as follows:

$$\begin{cases} \max Q(a) = S_z D_z \\ \text{s.t. } \sum_{i=1}^n a^2(i) = 1 \end{cases} \quad (12)$$

The value of PP depends on the change of projection direction a , which is a complex nonlinear optimization maximum problem with $a = (a_1, a_2, \dots, a_n)$ as variables. In order to simplify the computational process, a genetic algorithm (Bhattacharjya 2012) is used to solve the global optimization problem with higher-order nonlinear data. Take $Q(a)$ as the objective function to construct the fitness function, and take $a(i)$ as the gene population, calculate the maximum value of $Q(a)$ by genetic algorithm, and its corresponding optimal gene. $a(i)$ is the optimal projection direction. According to the research of previous scholars (Yuan & Li 2017), the weight of the evaluation index j is $w_j = a_j^2$. The specific algorithm flow of the GA-PP (genetic algorithm for projective pursuit) model is as Figure 1.

In summary, the PP model can reduce dimensionality without losing data features, but its distinguishability is closely related to the selection of projection direction. Usually, it is necessary to combine optimization models to select the optimal projection direction to achieve the best distinguishability. In addition, the distinguishability of PP is closely related to the selection of window radius. This article draws on the experience of previous scholars.

3.2. FCE model

3.2.1. Traditional fuzzy comprehensive model

The factors affecting rivers are often numerous and complex, and it is more reasonable to analyze the reliability of river water quality from multiple dimensions than from a single perspective. FCE (Icaga 2007) plays a crucial role in assessing river water quality, particularly when dealing with factors that are ambiguous and challenging to quantify for precise analysis. Through the process of quantification, it becomes possible to delineate the evaluation indicators' association with water quality standards across all levels of assessment. Ultimately, employing comprehensive evaluation principles, such as the principle of maximum affiliation (Abadias & Miana 2015), allows for a thorough evaluation of the integrated water quality category. This method ensures a precise and unbiased reflection of the relationship between water quality and the established standards. However, the hierarchical calculation of FCE depends on the selection of membership functions, and the indicators also need to be unified as cost-based or consumption-based indicators, which will bring a large amount of computational workload. In the traditional FCE method, the steps are as follows:

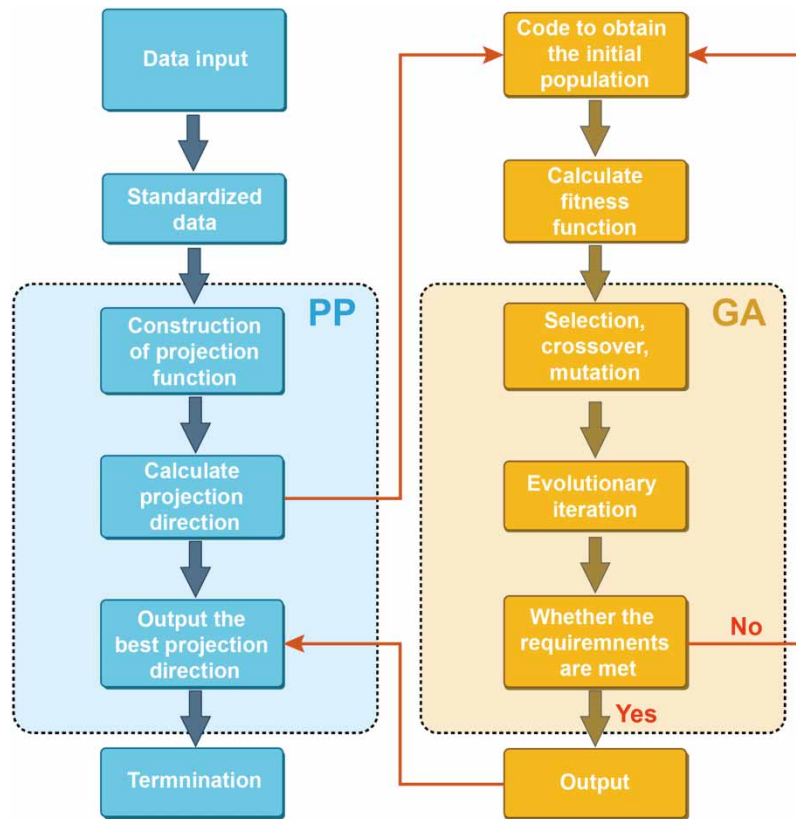


Figure 1 | Flow chart of the projection pursuit genetic algorithm.

- (1) Given the set of evaluation indicators, $U = \{u_1, u_2, u_3 \dots u_n\}$, given the set of evaluation levels, $V = \{v_1, v_2, v_3 \dots v_m\}$, calculated to determine the set of evaluation indicator weights, $W = \{w_1, w_2, w_3 \dots w_n\}$.
- (2) Single-factor affiliation judgment, to construct a set fuzzy matrix R for the membership degree of evaluation indicators and evaluation grade standards, a descending trapezoidal distribution of the affiliation function (Fu & Wang 2010) is often used to calculate the membership degree of evaluation indicators corresponding to various water quality standards. The larger the membership function value, the higher the membership degree of the evaluation indicator to a certain water quality grade. The fuzzy relationship matrix $R_{n \times m}$ for the degree of affiliation is shown in the following equation.

$$R = (r_{ij}) = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nm} \end{bmatrix} \tag{13}$$

where r_{ij} denotes the possibility of the i th factor to the j th evaluation, and the degree of affiliation of a water quality index i to the j th water quality level in this paper.

- (3) Synthesize the weight matrix W with the affiliation matrix R to get the FCE result vector B . That is, the weight matrix W and the affiliation matrix R will be synthesized.

$$B = WR = (w_1, w_2, w_3, \dots w_n) \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nn} \end{bmatrix} = (b_1, b_2, b_3, \dots b_n) \tag{14}$$

- (4) According to the principle of maximum affiliation, you can use the comprehensive analysis and evaluation results vector B to determine the grade of water quality. That is, in a group of evaluation objects, the largest value of b_j value on the level corresponding to the evaluation factor will determine the comprehensive evaluation grade of water quality in the reorganized group.

The traditional FCE method has the following advantages: it can define the boundaries of water quality classification through the description of relative affiliation, thus well reflecting the fuzzy nature of the water quality status and level division, and can relatively objectively reflect the status of comprehensive water quality. However, the traditional FCE method also has some shortcomings. The selection of indicators is easily affected by subjective factors, the assignment of weights may not be considered comprehensively, and the comprehensive evaluation processing easily leads to data loss (Xiong & Xian 2003). These problems may affect the accuracy and reliability of the evaluation results. In summary, the traditional FCE method has some advantages in describing the comprehensive status of water quality, but it needs to pay attention to the selection of indicators, assignment of weights and comprehensive evaluation processing in order to improve the reliability and practicability of the evaluation results.

3.2.2. Affinity function construction

The form of the affiliation distribution function is various, such as half trapezoidal distribution, rectangular distribution, normal distribution, and other functional forms. In the use of the fuzzy integrated evaluation method for water quality evaluation, the standard value of its evaluation domain for the real number, assuming that the degree of affiliation of water quality is linear distribution, then its distribution function can be used in the form of a descending semicircular trapezoidal distribution. The larger the value of the affiliation function, the higher the affiliation of the evaluation factor for a certain water quality level. For positive indicators (negative indicators, first convert them into positive through some transformation), indicator i to the j th level of water quality class affiliation formula is as follows:

- (1) When characterizing the water quality of the first level (level 1), when $j = 1$, the corresponding affiliation function is expressed as follows:

$$r_{ij} = \begin{cases} 1, & x_i < v_{i,j} \\ (v_{i,j+1} - x_i) / (v_{i,j+1} - v_{i,j}), & v_{i,j} \leq x_i \leq v_{i,j+1} \\ 0, & x_i > v_{i,j+1} \end{cases} \quad (15)$$

- (2) When characterizing (intermediate) water quality, when $j = 2, 3, 4, \dots, (m - 1)$, its corresponding membership function is characterized as follows:

$$r_{ij} = \begin{cases} 0, & x_i < v_{i,j-1} \text{ or } x_i > v_{i,j+1} \\ (x_i - v_{i,j-1}) / (v_{i,j} - v_{i,j-1}), & v_{i,j-1} \leq x_i \leq v_{i,j} \\ (v_{i,j+1} - x_i) / (v_{i,j+1} - v_{i,j}), & v_{i,j} \leq x_i \leq v_{i,j+1} \end{cases} \quad (16)$$

- (3) When characterizing the tail level (last level) water quality, when $j = m$, the corresponding membership function formula is as follows:

$$r_{ij} = \begin{cases} 0, & x_i < v_{i,j-1} \\ (x_i - v_{i,j-1}) / (v_{i,j} - v_{i,j-1}), & v_{i,j-1} \leq x_i \leq v_{i,j} \\ 1, & x_i > v_{i,j} \end{cases} \quad (17)$$

Among them, x_i represents the actual monitoring concentration value of the evaluation indicator i , $v_{i,j-1}$, $v_{i,j}$, $v_{i,j+1}$, respectively, represent the boundary values of the $j - 1$, j , and $j + 1$ water quality levels corresponding to the i th indicator.

3.2.3. Improved fuzzy comprehensive model

Considering the drawbacks of traditional fuzzy averaging that can easily cause information loss, this article proposes to adopt the improved FCE principle. The principle is to combine the membership degree b_j and level coefficient value $\{1, 2, \dots, m\}$ of each evaluation level j calculated by traditional algorithms for weighted calculation to determine the level of

comprehensive water quality, which can fully retain the original and processed data of each class to avoid the loss of water quality information. Because b_j is normalized, its value is less than or equal to 1, so there are $1 \leq B_g \leq n$. In this way, according to the value of B_g and the evaluation of which class coefficients of the class j is closest to the value of the sample closest to this class, the quantitative treatment of the water quality level, so that the evaluation results are more intuitive. The calculation formula of the improved fuzzy model is as follows:

$$B_g = \left(\sum_{j=1}^m j \times b_j^k \right) / \left(\sum_{j=1}^m b_j^k \right) \quad (18)$$

where b_j represents the membership degree of the evaluation object to the j th level water quality standard under the traditional fuzzy evaluation model; k is the weighting coefficient, and k is taken as 2 in this paper. When k approaches infinity, the above equation is transformed into the formula for calculating the maximum membership principle in traditional fuzzy evaluation models.

4. CASE STUDY

4.1. Example data

The Zhouxiang River is located in the Hanjiang District of Yangzhou, starting from the Miaozhuang Reservoir in the east and ending at the Youyi River in the west, with a total length of 3.06 km. The waterlogged water in the river is ultimately discharged into the Youyi River, which is the main flood control, drainage, and irrigation channel for Yangmiao Town in Hanjiang District. The measured data information of river WQIs of the Zhouxiang River from January to December 2022 was analyzed.

4.2. Standardization of indicators

The standardization of indicators represents different meanings, often resulting in significant differences in unit dimensions and unit orders of magnitude. If these indicators with significant differences are compared together, often no meaningful results can be obtained, and they will seriously affect the accuracy of the evaluation of WQIs. In order to make the comparison between indicators meaningful, it is necessary to unify each indicator at a single level, both for dimensionless processing and standardization of indicator data. Generally speaking, there are two dimensionless processing methods based on different categories of indicators:

- (a) Positive indicator: Usually also known as a benefit indicator, the larger the benefit indicator, the better the meaning of the indicator. In this article, DO belongs to this category of indicators. The standardized processing method for this type of indicator is as follows:

$$y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (19)$$

- (b) Negative indicator: Usually also known as a cost indicator, the smaller the cost indicator, the better the meaning of the indicator. The standardized processing method for this type of indicator is as follows:

$$y = \frac{x_{\max} - x}{x_{\max} - x_{\min}} \quad (20)$$

4.3. Index weight calculation

The program of the PP evaluation model (GA-PP) based on genetic algorithm optimization was prepared by using MATLAB. By setting the empirical parameters of genetic algorithm optimization and debugging the program, in order to find the optimal value of the function faster, the parameters were set as follows: population size N400, the maximum number of iterations $G_{\max} = 50$. The 12-month water quality data were calculated in a MATLAB environment, and the GA-PP program was run to obtain the optimal projection direction $a_i = (0.463, 0.033, 0.061, 0.579, 0.624, \text{ and } 0.238)$. Thereby the weights of the indicators w_i were also obtained, which is $w_i = (0.214, 0.001, 0.004, 0.335, 0.390, \text{ and } 0.057)$.

4.4. Improving FCE

4.4.1. Evaluation grade classification

River water is a kind of Earth's surface water, according to the *Environmental Quality Standards for Surface Water (GB3838-2013 Edition)*, Su *et al.* (2017) defined clear grade and concentration limits, and the water quality of surface water is divided into five water quality evaluation level. The domain interval values are divided into five levels: excellent I, good II, medium III, qualified IV, and unqualified V. Construct a set of evaluation levels: $V = \{I, II, III, IV, V\}$.

4.4.2. Affinity matrix calculation

According to Tables 2 and 3, formulas 15, 16, and 17 are used to calculate the membership values of different indicators for different level intervals under month k , forming a membership matrix R_k . For visual display, it is plotted in matrix form, with the horizontal axis representing WQIs and the vertical axis representing membership intervals. The number represents the degree of membership value as Figure 2.

4.4.3. Calculation of comprehensive evaluation value

The projective pursuit model optimized by the genetic algorithm calculates the weights and the improved fuzzy model improves the fuzzy evaluation of the weight matrix and the affiliation matrix, which results in the evaluation results. The Bg value of the water quality class in each month is referred to as method C. Method A is a traditional fuzzy comprehensive evaluation method, with its weights determined through AHP (Hierarchical Analysis) method (Vaidya & Kumar 2006), and method B is the traditional FCE method that calculates the weights by using the PP-GA method. Method D is the use of artificial neural networks (Najah *et al.* 2009), the monitoring data of historical time as the input value, the historical evaluation

Table 2 | Water quality monitoring data of Zhouxiang River

Monitoring indicators	DO	COD _{Mn}	BOD ₅	TP	TN	NH ₃ -N
January	9.53	1.33	1.41	0.09	1.62	0.14
February	9.81	1.65	1.21	0.12	1.65	0.18
March	9.12	1.43	1.12	0.11	1.58	0.21
April	8.10	1.58	1.21	0.08	1.46	0.21
May	6.91	1.62	0.91	0.04	1.34	0.23
June	6.61	2.21	0.51	0.08	2.36	0.13
July	5.70	2.16	0.82	0.03	2.05	0.19
August	5.30	2.87	0.73	0.05	1.91	0.12
September	4.41	2.72	0.51	0.06	1.82	0.32
October	7.62	1.94	0.61	0.05	1.76	0.15
November	8.11	1.67	0.70	0.09	1.68	0.09
December	8.31	1.53	0.52	0.07	1.52	0.07

Table 3 | Environmental quality standards for surface water (GB3838-2013)

Level interval	I	II	III	IV	V
DO	≥7.50	≥6.00	≥5.00	≥3.00	≥2.00
COD _{Mn}	≤2.00	≤4.00	≤6.00	≤10.00	≤15.00
BOD ₅	≤3.00	≤3.00	≤4.00	≤6.00	≤10.00
TP	≤0.02	≤0.10	≤0.20	≤0.30	≤0.40
TN	≤0.20	≤0.50	≤1.00	≤1.50	≤2.00
NH ₃ -N	≤0.15	≤0.50	≤1.00	≤1.50	≤2.00

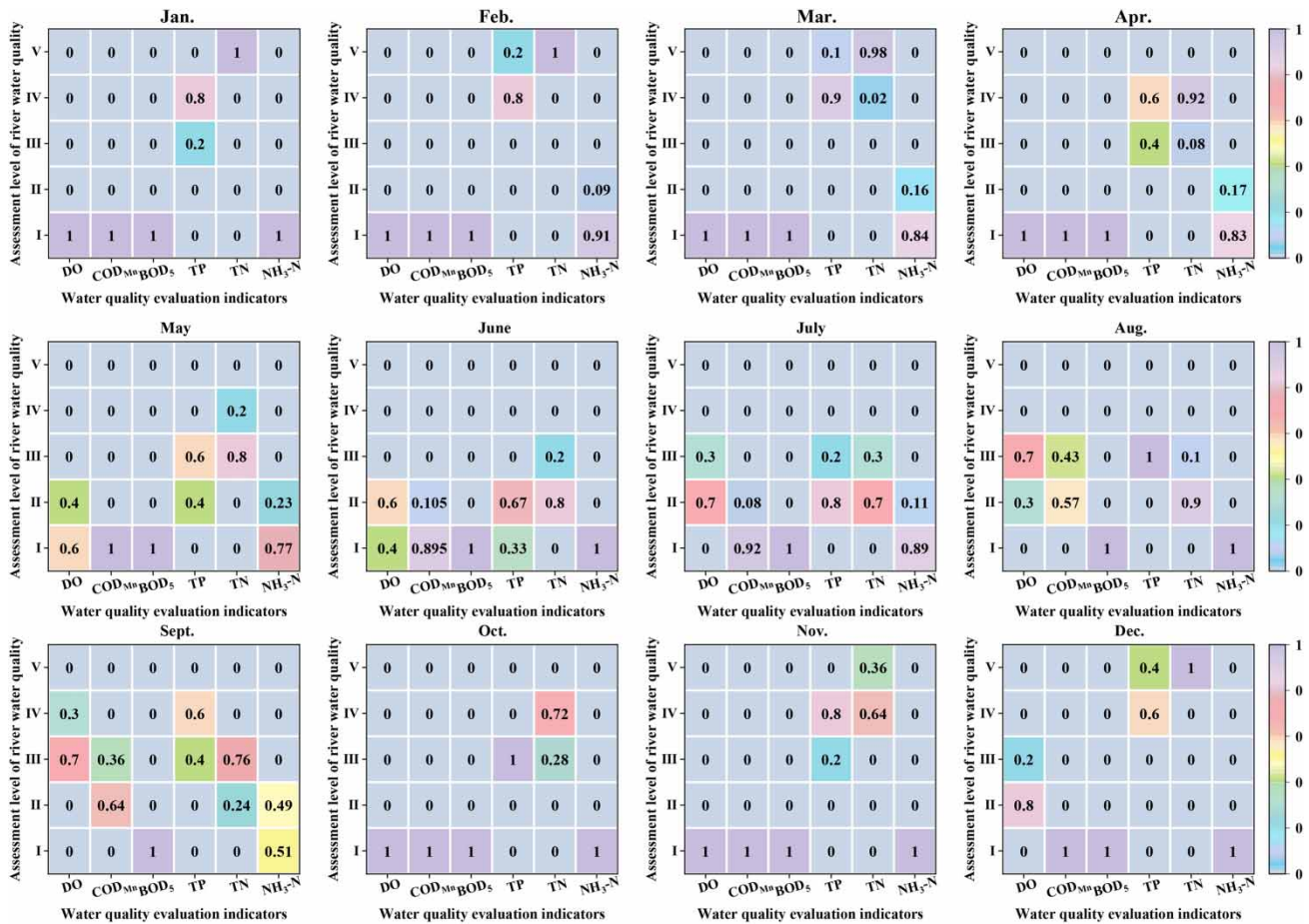


Figure 2 | The 12-month affiliation matrix pattern.

results as the output value, through the continuous forward and reverse feedback on the back propagation (BP) neural network training, and then substituting the data to be monitored to conclude, and the four kinds of results are compared (see Table 4).

In comparing the four evaluation methods, Method A’s weight calculation is too subjective. The evaluation adopts traditional FCE, which is incomplete in information synthesis, and the conclusions have a large gap with the other methods. Method B’s weight calculation is consistent with this paper and only changes the comprehensive evaluation method. It can be seen that the conclusions donot differ too much from the latter two, which shows that the effect of the weight calculation of PP-GA is better than the effect of AHP. Method C is the method adopted in this paper, which is improved in both weight calculation and comprehensive evaluation, and the results are consistent with the evaluation results of Method D, which indicates that these two improvements are of great help to the improvement of evaluation accuracy. Although the calculation results of Method D are consistent with this paper, the method requires tremendous database support. It is complicated to calculate, and it is not as good as the method of this paper for small samples or the portability of the method.

4.5. Summary and analysis

- (1) Analysis of the results: The weight analysis showed that TN, TP, and DO accounted for the most significant proportion of the weight of the water quality, so TN, TP, and DO are the main factors affecting water quality. The evaluation results of the four methods show that the river has the best water quality in June and July, and the laboratory results show that the TP of the water quality is low, and DO is high. Nitrogen is high in these 2 months. This is because these 2 months are the rainy season in the middle and lower reaches of the Yangtze River, with sufficient rainfall replenishment; it is also

Table 4 | Comparison of four evaluation results

Degree of affiliation	I	II	III	IV	V	Bg	A	B	C	D
January	0.276	0.000	0.067	0.268	0.39	3.735	V	V	IV	IV
February	0.270	0.005	0.000	0.268	0.456	3.968	V	V	IV	IV
March	0.267	0.009	0.000	0.309	0.415	3.880	V	V	IV	IV
April	0.266	0.010	0.165	0.559	0.000	3.417	IV	IV	III	III
May	0.177	0.233	0.513	0.078	0.000	2.688	III	III	III	III
June	0.259	0.663	0.078	0.000	0.000	1.881	II	II	II	II
July	0.055	0.697	0.248	0.000	0.000	2.106	II	II	II	II
August	0.060	0.415	0.524	0.000	0.000	2.601	II	III	III	III
September	0.033	0.122	0.580	0.265	0.000	3.126	II	III	III	III
October	0.276	0.000	0.444	0.28	0.000	2.792	II	III	III	III
November	0.276	0.000	0.067	0.517	0.140	3.422	III	IV	III	III
December	0.061	0.171	0.043	0.201	0.523	4.579	V	V	V	V

the planting season in the south, and the use of nitrogen fertilizers impacts the surrounding rivers. The water quality is the worst in winter. During that time, the TP concentration is high, COD is also high, with the dry season in winter, the water quality mobility is small, the concentration of TP and other pollutants is easy to accumulate; ammonia nitrogen concentration has a significant reduction; agricultural production is dormant in the winter.

- (2) Suggestions: First, to ensure the river's mobility and enhance the river's overflow capacity. Second, to increase the shoreline management, increase the vegetation cover, and effectively prevent soil erosion. In addition, strengthen the management of the surrounding farmland to prevent the irrigation water from the surrounding fields from flooding or pouring into the river, which wastes the fertilizer and causes environmental pollution. Finally, managing the river surface needs to be strengthened; debris, aquatic plants, and aquatic animals are densely covered with the river surface or corruption and deterioration will cause water quality deterioration.

6. CONCLUSION AND DISCUSSION

- (1) By monitoring the data of six indicators of river water quality for 1 year, the PP-GA method, which reduces dimensionality without losing data features, is used to more accurately determine the weights of indicators and make them closer to reality. Based on the improved FCE theory, a river water quality evaluation system was established and compared with four models: models with different weights, different evaluation methods, models with the same weight calculation, models with different evaluation methods, and artificial neural network models. Practice has shown that in terms of improving weight calculation and evaluation methods, our method is more accurate than neural network models, our method has higher portability and more minor sample requirements.
- (2) Using this evaluation model, the water quality comparison of the target river in each month can be accurately obtained, which is very consistent with the actual situation. Therefore, specific targeted measures to improve water quality are proposed.
- (3) The practice has proven that the river water quality evaluation model based on PP-GA and the improved fuzzy evaluation model has good practicality and applicability in river water quality evaluation and grading. This model can play a role in the water quality evaluation of other similar rivers, lakes, and seas.
- (4) However, in using such models, still the following aspects need further research, such as the selection of window radius for projection tracking and the use of software to calculate membership functions to simplify workload.

ACKNOWLEDGEMENTS

Z. H. conceptualized the whole article; X. Z. wrote the original draft; Y. C. rendered support in data curation.

The authors thank Scientific Research and Innovation Plans for Postgraduate of Jiangsu Province (No. KYCX22_3437).

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

- Abadias, L. & Miana, P. J. 2015 A subordination principle on Wright functions and regularized resolvent families. *Journal of Function Spaces* **2015**, 158145.
- Almeida, C., González, S. O., Mallea, M. & González, P. 2012 A recreational water quality index using chemical, physical and microbiological parameters. *Environmental Science and Pollution Research* **19** (8), 3400–3411.
- Animashaun, I. M., Abubakar, S., Mohammed, A. S., Adeoye, P. A. & Kuti, I. A. 2023 Determination of Water Quality Indices for Kofa Dam, Suleja, Nigeria. *PriMera Scientific Engineering* **2** (2), 48–58.
- Banda, T. D. & Kumarasamy, M. V. 2020 Development of Water Quality Indices (WQIs): a review. *Polish Journal of Environmental Studies* **29** (3), 2011–2021.
- Barcaru, A. 2019 Supervised projection pursuit – a dimensionality reduction technique optimized for probabilistic classification. *Chemometrics and Intelligent Laboratory Systems* **194**, 103867.
- Bhattacharjya, R. K. 2012 Introduction to genetic algorithms. *IIT Guwahati*, 12.
- Friedman, J. H. 1987 Exploratory projection pursuit. *Journal of the American Statistical Association* **82** (397), 249–266.
- Fu, G. & Wang, X. 2010 Chinese sentence-level sentiment classification based on fuzzy sets. In *Coling 2010: Posters*, 312–319.
- Huang, X. & Bai, H. 2018 Risk prediction of rural public security environmental carrying capacity based on the risk entropy. *Natural Hazards* **90** (1), 157–171.
- Huang, H. & Lu, J. 2014 Identification of river water pollution characteristics based on projection pursuit and factor analysis. *Environmental Earth Sciences* **72** (9), 3409–3417.
- Huber, P. J. 1985 Projection Pursuit. *The Annals of Statistics* **13** (2), 435–475.
- Icaga, Y. 2007 Fuzzy evaluation of water quality classification. *Ecological Indicators* **7** (3), 710–718.
- Jia, W., Sun, M., Lian, J. & Hou, S. 2022 Feature dimensionality reduction: a review. *Complex & Intelligent Systems* **8** (3), 2663–2693.
- Kim, M., Kim, S., Lee, S. & Cho, Y. 2014 Development of a hybrid best management practice system for control of agricultural nonpoint water pollution. *Journal of Agricultural Chemistry and Environment* **3** (4), 161–168.
- Lan, Z. & Huang, M. 2018 Safety assessment for seawall based on constrained maximum entropy projection pursuit model. *Natural Hazards* **91** (3), 1165–1178.
- Liu, D., Liu, C., Fu, Q., Li, T., Khan, M. I., Cui, S. & Faiz, M. A. 2018 Projection pursuit evaluation model of regional surface water environment based on improved chicken swarm optimization algorithm. *Water Resources Management* **32** (4), 1325–1342.
- Liu, D., Feng, J., Li, H., Fu, Q., Li, M., Faiz, M. A. & Khan, M. I. 2019a Spatiotemporal variation analysis of regional flood disaster resilience capability using an improved projection pursuit model based on the wind-driven optimization algorithm. *Journal of Cleaner Production* **241**, 118406.
- Liu, D., Zhang, G., Li, H., Fu, Q., Li, M., Faiz, M. A. & Khan, M. I. 2019b Projection pursuit evaluation model of a regional surface water environment based on an Ameliorative Moth-Flame Optimization algorithm. *Ecological Indicators* **107**, 105674.
- Munyika, S., Kongo, V. & Kimwaga, R. 2014 River health assessment using macroinvertebrates and water quality parameters: a case of the Orange River in Namibia. *Physics and Chemistry of the Earth, Parts A/B/C* **76–78**, 140–148.
- Najah, A., Elshafie, A., Karim, O. A. & Jaffar, O. 2009 Prediction of Johor River water quality parameters using artificial neural networks. *European Journal of Scientific Research* **28** (3), 422–435.
- Paun, I., Cruceru, L., Chiriac, F. L., Niculescu, M., Vasile, G. & Marin, N. M. 2016 *Water Quality Indices-Methods for Evaluating the Quality of Drinking Water. INCD ECOIND – International Symposium-SIMI 2016. The Environment and the Industry. Proceedings Book*.
- Poonam, T., Tanushree, B. & Sukalyan, C. 2015 Water quality indices-important tools for water quality assessment: a review. *International Journal of Advances in Chemistry* **1** (1), 15–28.
- Rubio-Arias, H., Ochoa-Rivero, J. M., Quintana, R. M., Saucedo-Teran, R., Ortiz-Delgado, R. C., Rey-Burciaga, N. I. & Espinoza-Prieto, J. R. 2013 Development of a water quality index (WQI) of an artificial aquatic ecosystem in Mexico. *Journal of Environmental Protection* **4** (11), 1296–1306.
- Sharifinia, M., Mahmoudifard, A., Gholami, K., Namin, J. I. & Ramezanzpour, Z. 2016 Benthic diatom and macroinvertebrate assemblages, a key for evaluation of river health and pollution in the Shahrood River, Iran. *Limnology* **17** (1), 95–109.
- Su, J., Ji, D., Lin, M., Chen, Y., Sun, Y., Huo, S. & Xi, B. 2017 Developing surface water quality standards in China. *Resources Conservation and Recycling* **117**, 294–303.
- Sutadian, A. D., Muttill, N., Yilmaz, A. G. & Perera, B. J. C. 2016 Development of river water quality indices – a review. *Environmental Monitoring and Assessment* **188** (1), 58.

- Vaidya, O. S. & Kumar, S. 2006 Analytic hierarchy process: an overview of applications. *European Journal of Operational Research* **169** (1), 1–29.
- Xiong, D. G. & Xian, X. F. 2003 Improvement of fuzzy comprehensive evaluation method. *Journal of Chongqing University* **26** (6), 93–95.
- Xu, S., Cui, Y., Yang, C., Wei, S., Dong, W., Huang, L. & Wang, W. 2021 The fuzzy comprehensive evaluation (FCE) and the principal component analysis (PCA) model simulation and its applications in water quality assessment of Nansi Lake Basin, China. *Environmental Engineering Research* **26** (2), 200022.
- You, G., Xu, B., Su, H., Zhang, S., Pan, J., Hou, X. & Ding, R. 2021 Evaluation of aquaculture water quality based on improved fuzzy comprehensive evaluation method. *Water* **13** (8), 1019.
- Yuan, J. & Li, C. 2017 Intuitionistic trapezoidal fuzzy group decision-making based on prospect Choquet integral operator and grey projection pursuit dynamic cluster. *Mathematical Problems in Engineering* **2017**, 1–13.
- Zhang, C. & Sihui, D. 2009 A new water quality assessment model based on projection pursuit technique. *Journal of Environmental Sciences* **21** (S1), S154–S157.

First received 1 August 2023; accepted in revised form 7 September 2023. Available online 5 October 2023