



# OPEN Research on the surface water quality in the Huaihe River Basin and the gensis based on multivariate statistical analysis

Shuzhen Feng<sup>1</sup>, Chaokai Zhang<sup>2,3,4</sup>, Jiaheng Yan<sup>2,3,4</sup>✉, Ke Ren<sup>5</sup>, Ningbo Peng<sup>2,3,4</sup>, Wei Jiang<sup>1</sup> & Shouhua Liu<sup>2</sup>

The Huaihe River Basin (HRB) is an important water system in eastern China, and its water quality has received widespread attention. This study explored the latest spatial variation patterns of surface water quality in the HRB to cope with the increasingly severe challenges of water resource management. By integrating multidimensional water quality data from surface water monitoring stations, including dissolved oxygen (DO), chemical oxygen demand (COD<sub>Mn</sub> and COD), biochemical oxygen demand (BOD<sub>5</sub>), ammonia nitrogen (NH<sub>3</sub>-N), total phosphorus (TP), and total nitrogen (TN), this study utilized a cluster analysis technique to categorize the water quality data and reveal changes in the geographic variability of water quality. Among the 382 monitoring stations in the HRB, 258 stations had TN content lower than Class V, which was the highest among all monitoring indicators. The entropy weight method used to assess the comprehensive water quality showed that there were 157 and 163 monitoring stations belonging to Class III and IV, respectively, and stations with poor water quality were distributed downstream in the river network and estuary area. Correlation and cluster analyses indicated that agricultural and organic matter pollution were the two main factors affecting water quality in the HRB, particularly in the downstream area, and the high loading of nutrient salts such as TP and NH<sub>3</sub>-N reflected the significant influence of agricultural activities. In addition, the study examined the potential driving role of factors such as topography, geomorphology, and human activities on water quality changes and visualized the relationship between water quality class and cluster categories through spatial distribution maps and Sankey diagrams to clarify the regional patterns of water quality problems.

**Keywords** Cluster analysis, Huaihe River Basin, Surface water quality, Agricultural pollution, Spatial change rules

As the global population continues to expand and industrialization accelerates, the demand for water resources is increasing at an unprecedented rate. This unprecedented surge in demand, coupled with the challenges posed by climate change, has led to the gradual depletion of surface water resources globally<sup>1,2</sup>. Surface water, mainly rivers, lakes, and reservoirs, is a key component of the global water cycle and a valuable freshwater resource indispensable for human drinking water and agricultural irrigation. It also plays a crucial role in maintaining ecosystem balance and biodiversity<sup>3,4</sup>. However, surface water availability and quality are declining at an alarming rate owing to human activities such as deforestation, urbanization, and excessive water withdrawals<sup>5–9</sup>.

Over the past few decades, rivers in China have suffered from severe water quality degradation caused by rapid and energy-intensive economic development<sup>5</sup>. Addressing river water quality and freshwater security in China in the twenty-first century is widely recognized as an urgent task<sup>10</sup>. In response to the deterioration of surface water quality, the Chinese government has implemented five phases of the Key Basin Water Pollution Prevention and Control Plan, which have been deployed to prevent and control water pollution and promote significant

<sup>1</sup>Nanjing Vocational Institute of Railway Technology, Nanjing 210031, China. <sup>2</sup>Faculty of Architecture and Civil Engineering, Huaiyin Institute of Technology, Huaian 223003, China. <sup>3</sup>HYIT-CREC Institute of Science and Technology for Conservation of Cultural Heritage, Huaiyin Institute of Technology, Huaian 223003, China. <sup>4</sup>Key Scientific Research Base of the State Administration of Cultural Heritage for Integrated Technology and Application of Grotto Cultural Heritage Protection Engineering Department, Chengdu 610036, China. <sup>5</sup>Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, China. ✉email: jiahengyan@126.com

improvements in the quality of the water environment. The Chinese government has adopted a source control policy that has led to a substantial reduction in point-source pollution<sup>11</sup>. However, nonpoint source pollutants may still cause deterioration of surface water quality. As a key water system in eastern China, the Huaihe River Basin (HRB) covers the provinces of Henan, Anhui, Jiangsu, and Shandong, and its abundant water resources are not only crucial for the survival of hundreds of millions of people in the region, but also an important natural resource to support the prosperity of agriculture and industrial development in this region. However, the HRB faces complex climate change situations (e.g., drought and heavy rainfall)<sup>12–14</sup> and anthropogenic pollution caused by rapid economic and social development<sup>15,16</sup>. In particular, the downstream plain area faces a series of challenges, such as increasing water pollution, prominent contradictions between the supply and demand of water resources, and frequent floods<sup>17</sup>, all of which indicate the need for higher requirements in the management and protection of surface water.

To implement effective water resource management, scholars have primarily studied the water quality temporal change processes in the HRB, with few studies using statistical relationships between these indicators to explain the reasons for water quality changes. For example, Zhai et al.<sup>17</sup> used the Mann–Kendall trend and regression analyses to reveal the trends of chemical oxygen demand ( $\text{COD}_{\text{Mn}}$ ), ammonia nitrogen ( $\text{NH}_3\text{-N}$ ), and dissolved oxygen (DO) in the surface water of the HRB from 1994 to 2005. Dou et al.<sup>18</sup> focused on the pollutant contents of surface water and evaluated the changes in biochemical oxygen demand ( $\text{BOD}_5$ ),  $\text{COD}_{\text{Mn}}$ ,  $\text{NH}_3\text{-N}$ , and total phosphorus (TP) at 20 stations in the HRB from 2003 to 2012. Li et al.<sup>19</sup> quantified the factors influencing the spatiotemporal variations in surface water quality in the HRB from 2011 to 2018 using various statistical methods. Xu et al.<sup>20</sup> assessed temporal variations in  $\text{NH}_3\text{-N}$  concentrations in the HRB in relation to pollution source control policies. Although existing research has promoted people's understanding of the temporal variation characteristics of specific water quality indicators in the HRB, there is still a lack of comprehensive understanding of the overall characteristics of water quality. The intrinsic correlation between water quality indicators of rivers and lakes indicates the control mechanism of water quality changes, especially in large-scale spatial areas such as the HRB, where multiple water quality mechanisms often exhibit spatial variability. Therefore, it is necessary to strengthen research in this area.

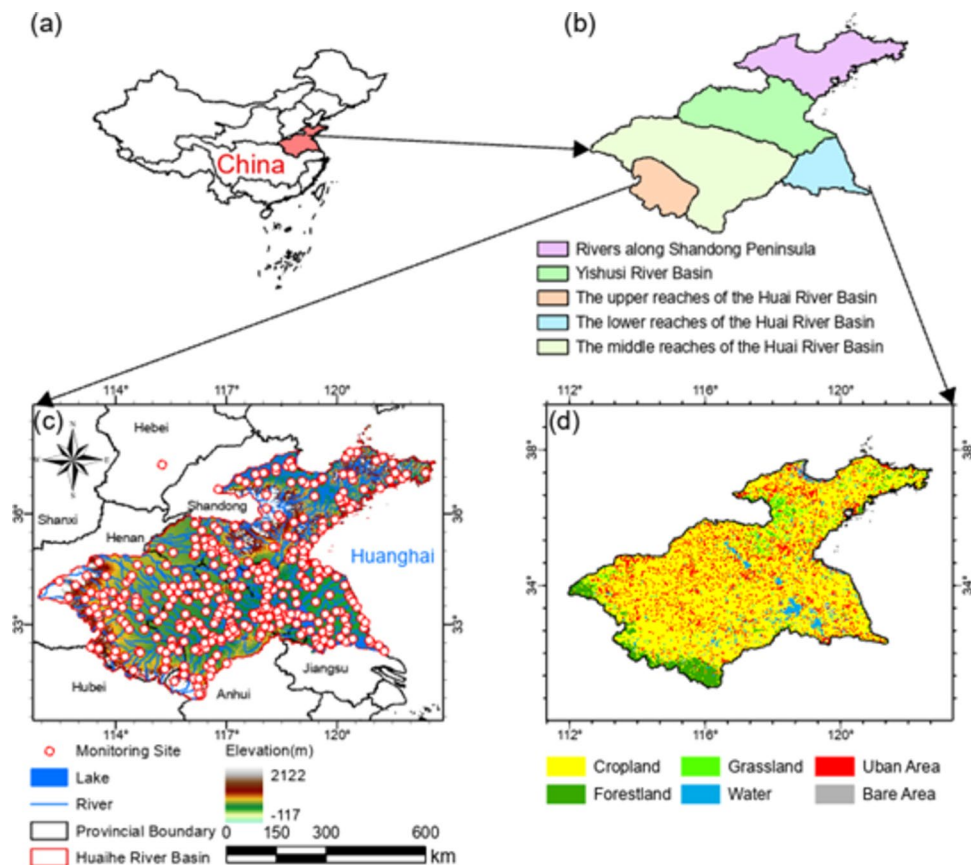
In 2023, China's Ministry of Ecology and Environment and five other departments jointly issued the "Key Basin Water Ecosystem Environmental Protection Plan," which is a top-level plan for water ecosystem environmental protection in the HRB that clearly defines the objectives and requirements for water ecosystem environmental protection and promotes the development of a new situation for the management of water resources, environment, and ecosystems based on the past focus on the prevention and control of water pollution ([https://www.mee.gov.cn/zcwj/zcjd/202306/t20230601\\_1032042.shtml](https://www.mee.gov.cn/zcwj/zcjd/202306/t20230601_1032042.shtml)). Considering the current situation and trends of surface water quality in the HRB and its influencing factors, it is critical to adopt scientific and effective analytical methods<sup>21–23</sup>. The HRB has a complete system of water quality monitoring indicators and a publicly available spatiotemporal dataset for water quality monitoring. However, the difficulty in analyzing such multivariate water quality indicator data lies in deciphering the quantitative relationships between variables. Cluster analysis, as a pattern recognition and data analysis technique, has shown significant advantages in processing a large amount of environmental monitoring data and revealing the changing patterns of the spatial distribution characteristics of water quality. This method can classify complex water quality datasets into several intrinsically related categories or clusters based on similarities and differences between water quality parameters, thus identifying regions with different water quality types and providing a scientific basis for further water quality management and pollution control strategies<sup>24</sup>.

The purpose of this study was to integrate multidimensional water quality data from surface water monitoring stations in the HRB and systematically explore the spatial distribution patterns of water quality parameters using cluster analysis techniques. Specifically, this study applied a clustering algorithm to classify the water quality data, explore the geographically differentiated patterns of change in water quality, analyze the correlation of water quality parameters within the clusters, identify the main pollutants and their sources, assess the environmental stress in different regions, and combine the topography and human activities in the watershed with the potential driving force for changes in water quality. This study is not only of theoretical significance, as it helps to deepen the understanding of the spatiotemporal evolution mechanisms of complex hydrological systems, but also provides strong technical support for the formulation of sustainable water resource management and protection strategies in the middle and lower reaches of the HRB, which is of significant socioeconomic value in practice.

## Study area

The HRB is located in east-central China (Fig. 1a), between longitude 111°55'–121°20' and latitude 30°55'–36°20', with an area of 270,000 km<sup>2</sup>. It begins from the Tongbai and Founou Mountains to the west and borders the Yangtze River Basin to the east. It is bordered by the Dabie Mountains, Jianghuai Hills, Tongyang Canal, and Rutai Canal to the south and the Yellow River Basin to the north by the southern embankment of the river and the Yimeng Mountain Range. The western, southern, and northeastern parts of the HRB are hilly areas, accounting for approximately 1/3 of the total area of the basin, whereas the rest are plains (including lakes and depressions), which are important parts of the Huanghai Plains (Fig. 1c).

The Huai River originates from Tongbai Mountain in Henan Province and flows from west to east through four provinces: Henan, Hubei, Anhui, and Jiangsu. The main stream enters the Yangtze River at Sanjiangying in Yangzhou, Jiangsu Province, with a total length of approximately 1000 km and a total elevation change of 200 m. The HRB consists of five subsystems: the upper, middle, and lower reaches of the Huai River, the Yishui River systems, and the coastal rivers of the Shandong Peninsula (Fig. 1b). The lower reaches of the Huai River are outlets for the waterway into the sea, the general irrigation canal in northern Jiangsu, and the Yellow River waste. The upper reaches of the Huaihe River have a large specific drop, the middle and lower reaches have a small specific drop, and there are many lakes and depressions on both sides of the main stream, with many tributaries.



**Fig. 1.** Study area.

The whole water system is asymmetrically distributed in the form of a fan-shaped plume. The Yishusi River system is located in the northeast (Fig. 1b).

The HRB is located in the climate transition zone between northern and southern China, with the northern part having a warm-temperate semi-humid monsoon climate and the southern part having a subtropical humid monsoon climate. Its weather system is complex and variable, and annual precipitation varies significantly. The average annual precipitation in the basin is 878 mm, with 600–700 mm in the northern part along the Yellow River, and 1400–1500 mm in the southern mountainous areas. Precipitation during the flood season (June–September) accounts for approximately 50–75% of the annual precipitation. The total water resources of the basin are 81.2 billion cubic meters on average, of which surface water resources, at 60.6 billion cubic meters, account for 75%.

The HRB is densely populated and dominated by arable land (Fig. 1d). Urban areas are distributed between rural areas, exhibiting local concentration characteristics. Forests and grasslands are concentrated in higher elevations such as the southwest and north of the watershed, respectively. There are scattered undeveloped wastelands in coastal areas. With fertile land, abundant resources, and convenient transportation, it is covered by the Yangtze River Economic Belt, Yangtze River Delta Integration, and Central Plains Economic Zone, as well as the main gathering area of the Grand Canal Cultural Belt. The basin accounts for approximately 11.8% of the total population of the country, with an urbanization rate of 54.2% and average population density of 607 people/km<sup>2</sup>, which is 4.2 times higher than the national average population density. The HRB is an important transportation hub in China, with railroad lines running north–south and east–west, highways in all directions, and waterways, such as the Beijing–Hangzhou Canal and the Huihe River, leading from the river to the sea.

## Materials and methods

### Sample collection

Surface water monitoring data were obtained in January, 2024 from China's National Automatic Surface Water Quality Monitoring System, which can realize real-time continuous and remote monitoring of water quality. During this period, the precipitation was relatively low, and the water quality indicators were relatively less affected by climate factors. There are totally 382 monitoring sites in the HRB and the monitoring items of the automatic water quality monitoring stations include water temperature, pH, DO, electrical conductivity (EC), turbidity, permanganate index, ammonia nitrogen, total nitrogen (TN), and TP for the automatic water quality monitoring stations in lakes. The monitoring data were transmitted to the host station of each automatic water quality station, provincial monitoring center station, and General Station of China Environmental Monitoring via virtual private network<sup>25,26</sup>.

### Comprehensive evaluation of water quality based on the entropy weight method

The latest surface water quality standards were used in this study<sup>26</sup>. Each water-quality monitoring indicator was categorized into six evaluation levels based on specified thresholds (Table S1). A comprehensive evaluation of water quality is a multi-objective decision-making problem<sup>27</sup>. The entropy weight method utilizes information entropy or Shannon's entropy, which is a multiobjective decision-making problem, as a measure of diversity in attribute importance<sup>28</sup>. In information theory, entropy is used to represent uncertainty or the degree of disorder in a system. The entropy weight method calculates the entropy value of each indicator to determine its degree of dispersion and then determines its weight in a comprehensive evaluation. Therefore, the entropy weight method is an objective assignment method for determining the weight of each evaluation index and its importance in a comprehensive evaluation. Specifically, the smaller the entropy value is, the greater the degree of dispersion of the indicator, and the greater the influence of the indicator on the comprehensive evaluation<sup>29</sup>. Calculation steps of the entropy weight method are as follows<sup>30</sup>:

1. Constructing a judgment matrix: First, it is necessary to construct a judgment matrix of the evaluation indicators for each evaluation object. For a total of  $m$  samples of evaluation objects, each sample has  $n$  evaluation indicators, and the judgment matrix is  $X = (x_{ij}) (i = 1, 2, \dots, m; j = 1, 2, \dots, n)$ .
2. Normalization: The judgment matrix is normalized to determine the normalized judgment matrix.

For larger and better type of metrics:

$$a_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (1)$$

For smaller and better type of indicators:

$$a_{ij} = \frac{\max(x_j) - x_{ij}}{\max(x_j) - \min(x_j)} \quad (2)$$

3. Calculation of the entropy value: The entropy value of each evaluation index is calculated based on the definition of entropy and the normalized judgment matrix,

$$e_i = -k \sum_{j=1}^n y_{ij} \ln y_{ij} \quad (3)$$

where  $k = -\ln \frac{1}{n} = \ln n$  is the standardization factor.

4. Definition of the entropy weight: Based on the entropy value of each indicator, its entropy weight is defined:

$$w_i = \frac{1 - e_i}{n - \sum_{i=1}^m e_i} \quad (4)$$

This leads to a weight vector for this comprehensive water quality evaluation  $\vec{W} = (w_1, w_2, \dots, w_n)$ .

5. Calculation of the entropy weight water quality index (EWQI): The vector of weights ( $\vec{W}$ ) and water quality indicators ( $\vec{x}_i$ ) for each monitoring station are multiplied, and the EWQI is obtained.

### Multivariate statistical analysis

Multivariate statistics focus on the study of interdependent statistical regularities among multiple variables, and cluster analysis is a multivariate statistical method used to classify samples into different class or clusters based on the similarities or differences between the samples. K-means cluster analysis is an unsupervised learning algorithm that classifies  $n$  observations in a dataset into  $K$  ( $K \leq n$ ) clusters, such that each observation belongs to the cluster corresponding to the closest mean (i.e., the center of the clusters)<sup>31</sup>. This algorithm is suitable for large-scale datasets owing to the relative simplicity of its computational processes. The algorithmic steps of K-means clustering analysis consist of three main parts:

1. Initial step: Randomly select  $K$  observations as the initial clustering centers.
2. Allocation: For each observation in the dataset, calculate its distance to the  $K$  clustering centers and assign it to the cluster corresponding to the closest clustering center.
3. Update: For each cluster, calculate the mean of all observations and set that mean as the new clustering center.

By repeating the above assignment and updating steps until the clustering centers no longer changed significantly, it was possible to divide the dataset into K well-defined regions with a high degree of similarity of data points and significant differences between regions.

## Results

### Characteristics of surface water quality in the HRB

The statistical characteristics of the water quality indicators in the HRB are shown in Table 1. The spatial variations of all water quality indicators are determined using Kriging interpolation (Fig. 2). Table 1 shows that the monitoring results for pH were 6.0–9.0, and the average value was 8.02, indicating that at most of the monitoring stations, the surface water was alkaline. For lakes and rivers, the pH can directly affect the survival and reproduction of organisms in water and the form of nutrients. Surface water in an alkaline environment not only favors the growth of alkaline-tolerant plants but also regulates nitrogen, phosphorus<sup>32</sup>, and heavy metal elements<sup>33,34</sup>. The sites with relatively low pH are concentrated in the southwest direction of the HRB (Fig. S1). According to the terrain and other water quality parameters of the region (Fig. 2), the lower pH value may be related to the higher terrain and abundant aquatic life in the area, as indicators such as COD, BOD<sub>5</sub> and COD<sub>Mn</sub> show high values. The respiration of aquatic organisms releases CO<sub>2</sub> into the water, which helps to lower the pH value of the surface water. The mean value of EC was 120.24 µs/cm, with a high standard deviation of 240.13 µs/cm, showing that EC varied greatly among different sampling sites and was related to the geographic differences in the dissolved mineral and pollutant contents in each site. As shown in Fig. 2a, monitoring stations with high EC were concentrated in the coastal rivers of Shandong Province. According to previous research, as Shandong is an important industrial and agricultural production base in China, the discharge of domestic sewage and industrial wastewater is a significant cause of some river water pollution in this area<sup>35</sup>. In addition, due to the proximity to the ocean, some areas in Shandong have been affected by seawater intrusion, resulting in salinization of surface water<sup>36</sup>. The mean value of turbidity was 17.97 NTU, indicating the presence of more suspended matter in certain sites. Although it was in the acceptable range, the turbidity fluctuated considerably (standard deviation of 32.11 NTU), indicating that the surface water was affected by suspended particulate matter to varying degrees, as was the transparency of water quality and photosynthesis. The mean value of dissolved oxygen (DO) was 12.15 mg/L, with a standard deviation of 2.08 mg/L, indicating that although the survival needs of most aquatic organisms were met overall, there was spatial heterogeneity, and hypoxia may have existed in localized areas in the southern part of the watershed (Fig. 2c). The mean values of COD<sub>Mn</sub> and COD were 4 and 14.86 mg/L, indicating that the load of organic pollutants was relatively high, particularly the high variability of COD (standard deviation of 4.32 mg/L), which showed that the water body was subjected to a significant difference in the degree of organic pollution in different regions (Fig. 2d). The mean value of BOD<sub>5</sub> was 2.47 mg/L, which indicated that the biochemical oxygen demand was moderate, but there was also high variability, reflecting the active and unstable degradation process of the organic matter and its relationship with the self-purification capacity of the water body and pollution discharge. This reflects the active and unstable degradation of organic matter, which is related to the self-purification ability of the water body and pollution discharge. NH<sub>3</sub>-N, TP, and TN represented the nutrient indicators, with mean values of 0.26, 0.067, and 4.01 mg/L, respectively, suggesting a risk of eutrophication, and in particular, a wide range of TN (Fig. 2g), suggesting that nitrogen is one of the main factors limiting the quality of the surface water in the HRB.

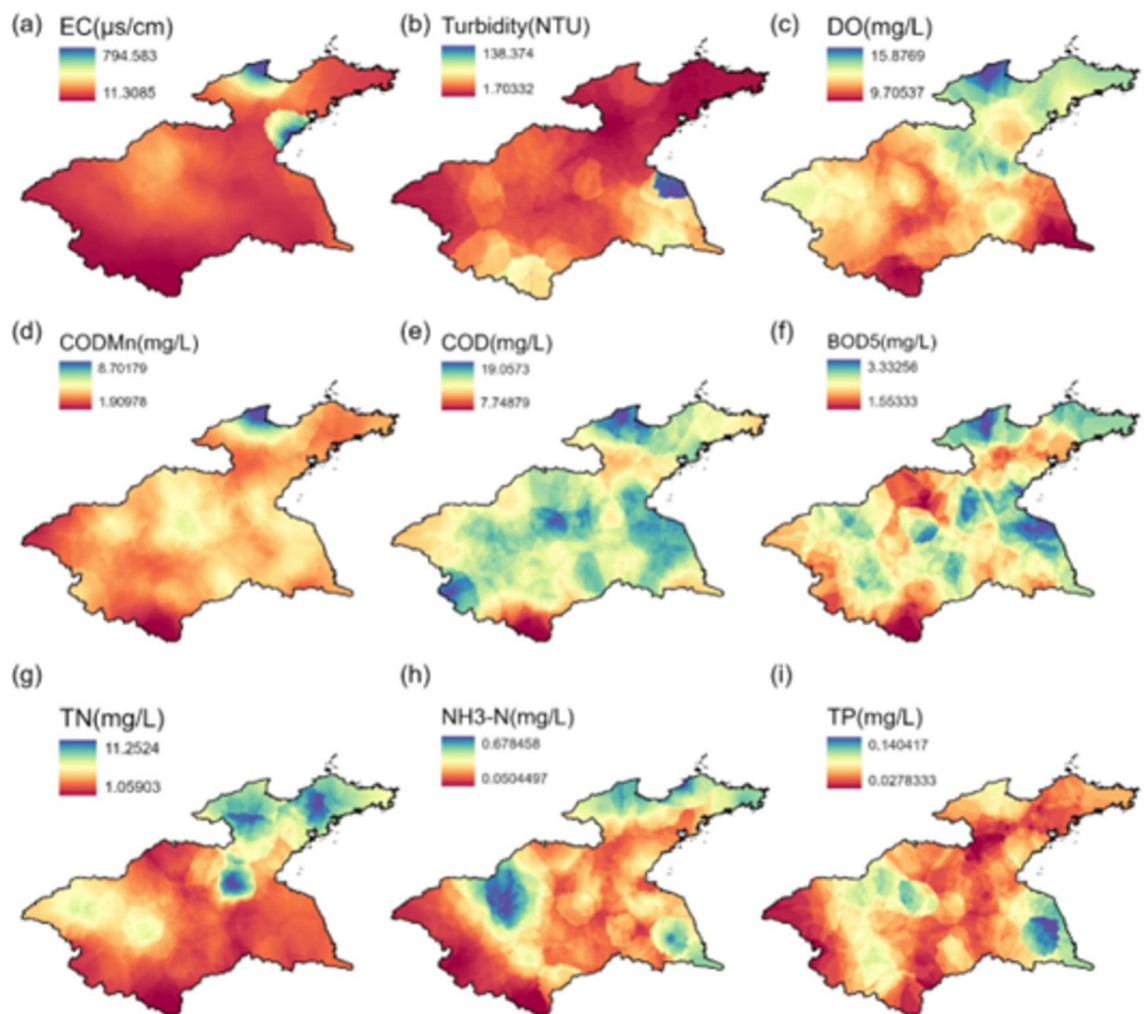
### Classification of surface water quality in the HRB

To illustrate the surface water quality in the HRB, quality indicators at each monitoring station were compared to surface water quality standards (Table S2). Because of relatively mild local temperatures in the HRB, it is easy for surface water to dissolve sufficient oxygen, and most surface water stations in the basin showed that the DO content reached the water quality status of Class I. COD<sub>Mn</sub>, COD, and BOD<sub>5</sub> reflected that most of the monitoring stations were in water quality Class II and III, indicating that organic pollution caused a certain degree of deterioration in the surface water quality. Especially, COD<sub>Mn</sub> indicated that only 23 stations were in Class I, which was significantly lower than that of COD and BOD<sub>5</sub> (217 and 287), indicating that the surface water was not only polluted by organic matter but also dissolved more oxidizable inorganic matter. Regarding

Chemical parameter	Unit	Number of samples	Maximum	Minium	Mean	SD
pH	-	382	9.0	6.0	8.02	0.43
DO	mg/L	382	22.9	5.3	12.15	2.08
EC	µs/cm	382	4100	6.9	120.2	240.1
Turbidity	NTU	381	511	0.7	17.97	32.11
COD <sub>Mn</sub>	mg/L	382	14.2	1.3	4.0	1.37
COD	mg/L	380	35.0	4.0	14.86	4.32
BOD <sub>5</sub>	mg/L	382	8.5	0.7	2.47	0.98
NH <sub>3</sub> -N	mg/L	382	2.33	0.02	0.26	0.29
TP	mg/L	382	0.42	0.005	0.067	0.046
TN	mg/L	382	31.3	0.33	4.01	3.87
F <sup>-</sup>	mg/L	382	1.2	0.13	0.56	0.21

**Table 1.** Characteristics of surface water quality in the HRB.

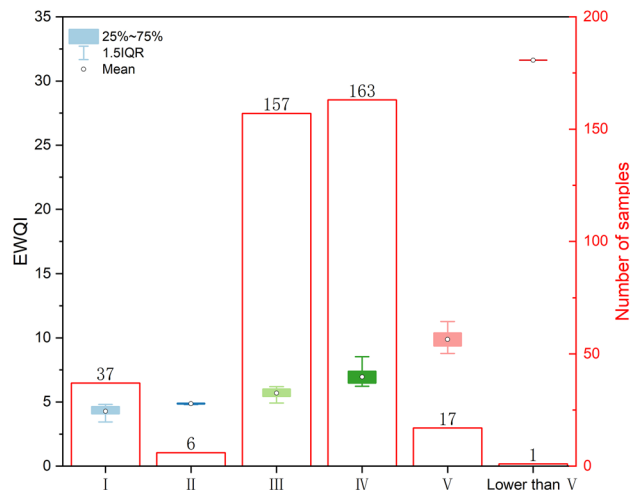




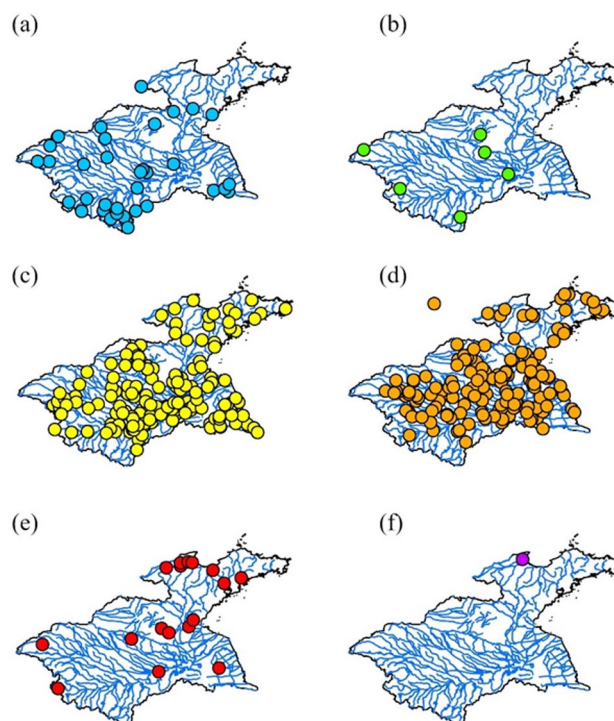
**Fig. 2.** Characteristics of spatial changes in surface water quality indicators in HRB (Generated by the Kriging method provided in ArcgisPro 3.01, url:<https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview>).

nutrients, the increase in nitrogen and phosphorus content had the same effect on surface water quality. The  $\text{NH}_3\text{-N}$  content showed that only 174 monitoring stations had Class I water quality, whereas TP and TN showed that 29 and 1 stations had Class I water quality, respectively, indicating that the nutrient pollution of water bodies is more serious than that of organic matter. In particular, high TN content was the main cause of water quality deterioration, with 258 water quality samples with levels lower than Class V, indicating that human activities leading to nitrogen input to surface water quality pollution are the most serious.

The weights of each water quality indicator obtained using the entropy weight method were 0.180, 0.155, 0.177, 0.156, 0.092, 0.128, and 0.112. The weights calculated from the entropy were relatively small because the monitoring results of ammonia nitrogen contained relatively little random information. The weights of the remaining indicators are approximately the same size, indicating that the randomness of the distribution of the remaining water quality evaluation indicators at each monitoring site is approximately the same, ensuring that the results of the comprehensive evaluation of water quality have good robustness. The evaluation results of the EWQI calculated using entropy weights are shown in Fig. 3, where the number of monitoring stations corresponding to each class of water quality was 37, 6, 157, 163, 17, and 1, indicating that most of the monitoring stations had a water quality Class of III or IV. Comparison with the analysis results of single rating indicators shows that the comprehensive water quality assessment reflects that the number of stations in Class IV has increased significantly, whereas the number of stations in Class I, II, V, or lower than V is less, suggesting that the water quality status reflected by a single evaluation indicator can be either more aggressive or more conservative and does not comprehensively reflect the state of surface water quality. Figure 4 shows the spatial distribution characteristics of the entropy weight method used to assess the water quality at all monitoring stations. In the HRB, monitoring stations with water quality Class I and II were concentrated in the western part of the basin (Fig. 4a,b), indicating that water quality was better in most of the upper part of the river network. Monitoring stations with water quality Class III and IV covered most of the study area (Fig. 4c and d), indicating that the surface water quality in the HRB deteriorated. The monitoring stations with water quality Class V and below are



**Fig. 3.** Statistical characteristics of surface water quality indices for different class.



**Fig. 4.** Characteristics of the spatial distribution for surface water quality in the HRB (a–f represent the spatial distribution of monitoring stations in the basin with water quality Class I, Class II, Class III, Class IV, Class V, and below Class V, respectively).

mainly distributed in rivers in the central part of the basin as well as in the northern part of the river, such as the Daqing River (Fig. 4e,f), indicating that seriously polluted areas temporarily exist locally<sup>37,38</sup>.

### Discussions5.1 Mechanisms for controlling spatial changes in surface water quality

Surface water quality in the basin are primarily controlled by natural processes and anthropogenic activities. Multivariate statistical methods, such as correlation and cluster analyses, can effectively identify the causes of water quality. The results of the correlation analysis of all water quality indicators in the HRB are shown in Fig. 5. The Pearson correlation coefficient revealed the degree of correlation between the single indicators. In Fig. 5, the correlation coefficients between COD, COD<sub>Mn</sub>, and BOD<sub>5</sub> were relatively high because these indicators reflect the oxygen consumption of surface water that undergoes oxidation under different conditions<sup>34</sup>. Organic matter oxidation is an important factor in the variation of surface water quality because of the presence of high amounts of organic matter in surface water. DO reflects the content of dissolved oxygen in the surface water; therefore,

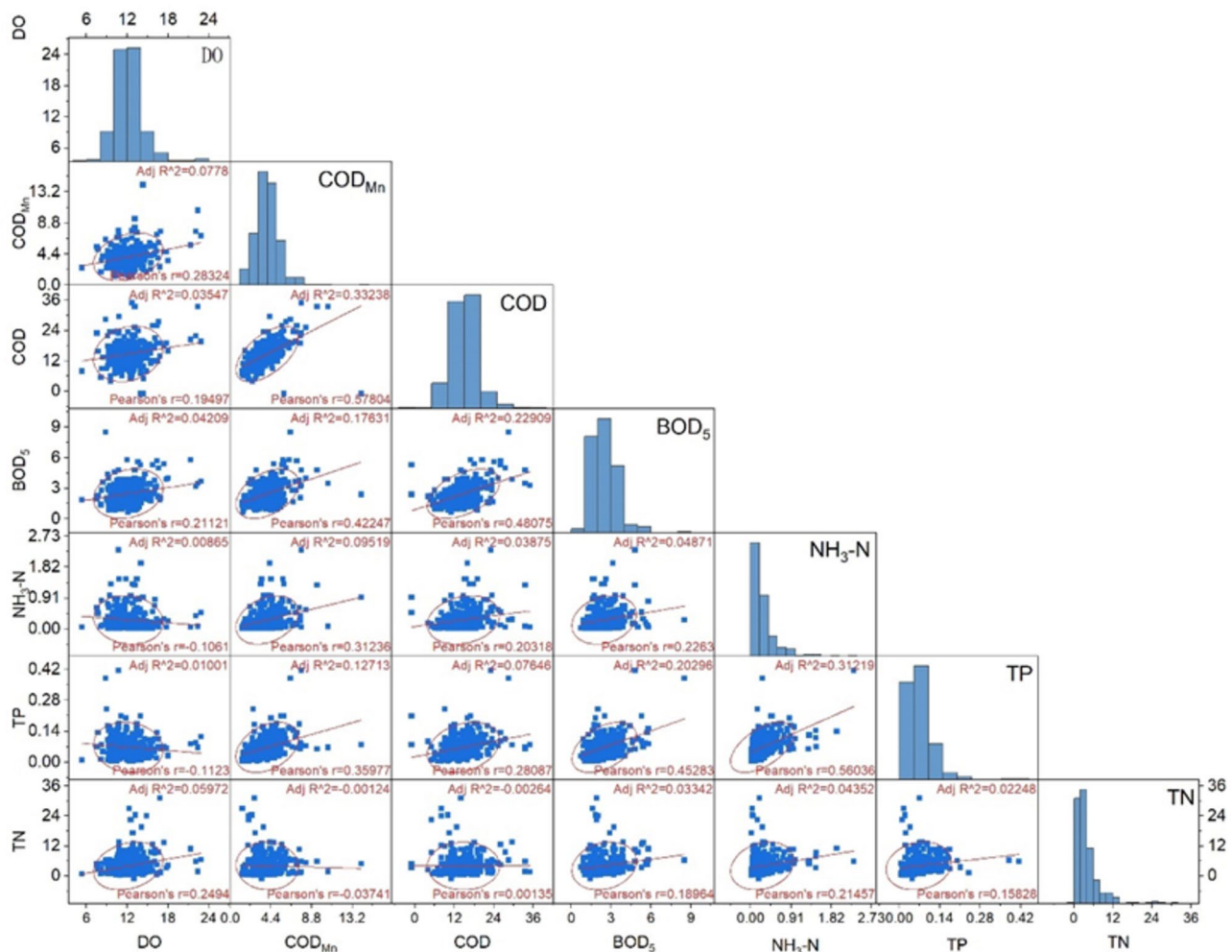


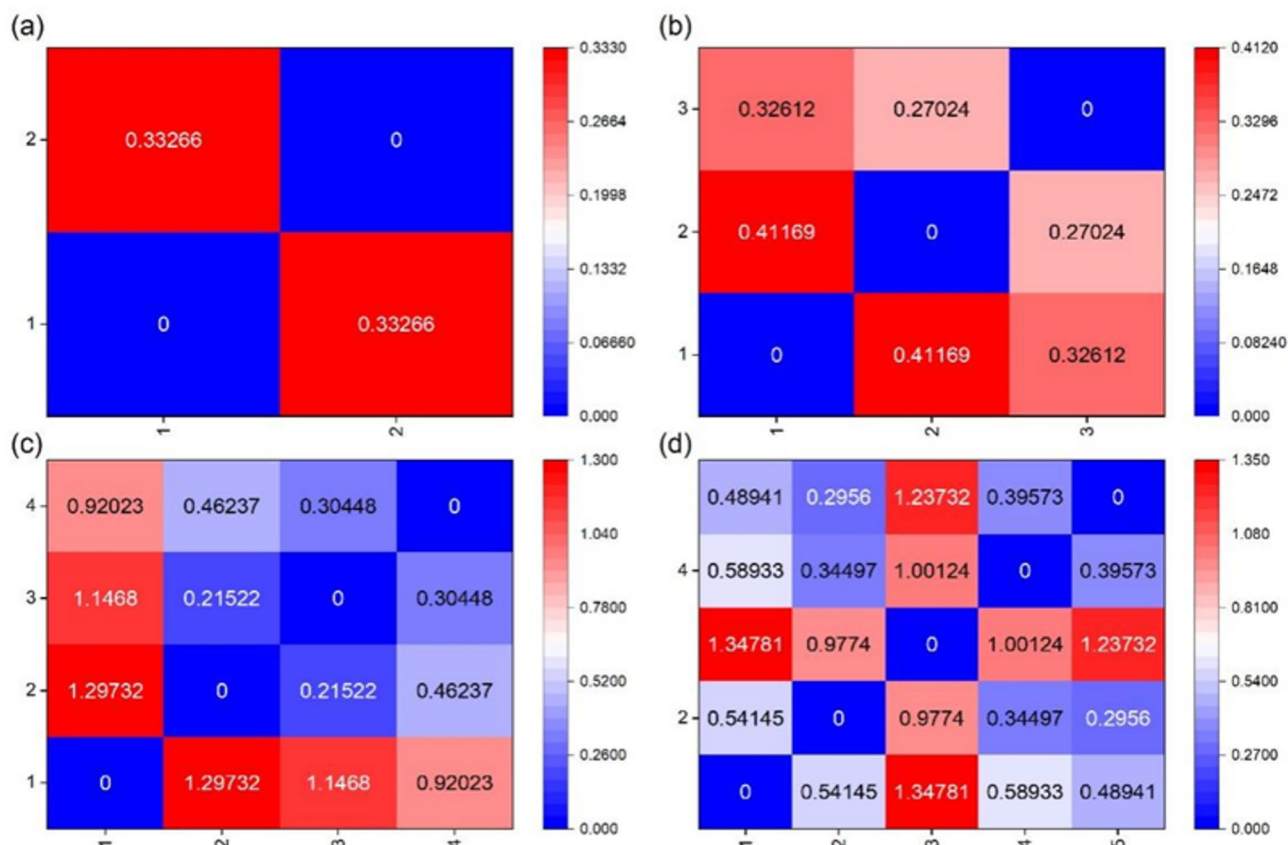
Fig. 5. Correlation of surface water quality indicators in HRB.

the correlation between DO and COD,  $\text{COD}_{\text{Mn}}$  and  $\text{BOD}_5$  was high. In addition, the correlation coefficient between TN and DO was 0.2494, which was close to the above three indicators and higher than the correlation coefficient between  $\text{NH}_3\text{-N}$  and DO. This can be attributed to the fact that the presence of DO could promote the nitrification reaction and inhibit the  $\text{NH}_3\text{-N}$  content. Among the indicators characterizing the nutrient chemistry of the water bodies, TP was highly correlated with  $\text{NH}_3\text{-N}$  ( $r=0.56$ ). Agricultural pollution is one of the main sources of phosphorus and nitrogen in surface water. For large amounts of phosphorus fertilizer and nitrogen fertilizer application, irrigation water easily inputs elements such as N and P into surface aquatic environments such as rivers and lakes<sup>3,39</sup>; thus, the correlation between TP and  $\text{NH}_3\text{-N}$  indicates the impact of agricultural pollution on water quality in the HRB.

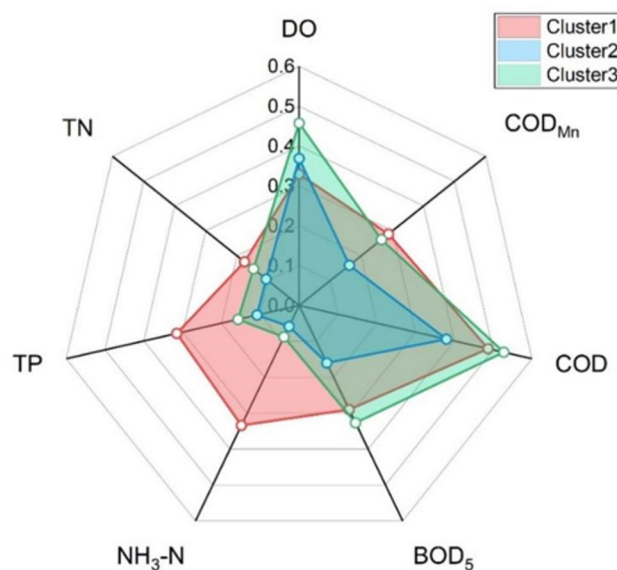
Cluster analysis of indicators from all monitoring stations as multivariate variables can provide further insights into the factors controlling water quality changes. The K-Means clustering method, based on Euclidean spatial relationships, requires presetting the number of clusters to obtain, which determines the reliability of the cluster analysis results. Therefore, the final statistical clustering characteristics corresponding to cluster numbers 2, 3, 4, and 5 were compared. As shown in Table S3 and Fig. 6, the number of cluster observations, variance, mean distance from sample points within a cluster to the center of clustering, and maximum distance reflect the structural characteristics of the classification for different cluster sizes. Ideal clustering results should ensure that the number of clusters is balanced, the within-cluster variance is relatively small, and the spatial distance between clusters is relatively uniform. When the number of clusters is four or five, clusters with only two sample points appear, and the intracluster sample variance is significantly smaller than that of the other clusters. In addition, a comparison of the cluster distances shows that clusters with sizes of four and five correspond to significantly different distances (red and blue parts in Fig. 6), indicating that too many clusters are not justified. When the number of clusters is two or three, the number of samples is relatively uniform, but the intra-cluster variance corresponding to cluster number 2 is significantly larger, indicating that the two clusters cannot evenly divide the original data structure. Therefore, three clusters were used for cluster analysis.

The indicator loads corresponding to the center point within each cluster reflect the overall water quality characteristics of the monitoring stations within the corresponding cluster (Fig. 7). For Cluster 1, all water quality indicators had high loadings, except for DO. The loads of TP and  $\text{NH}_3\text{-N}$  were significantly higher than those



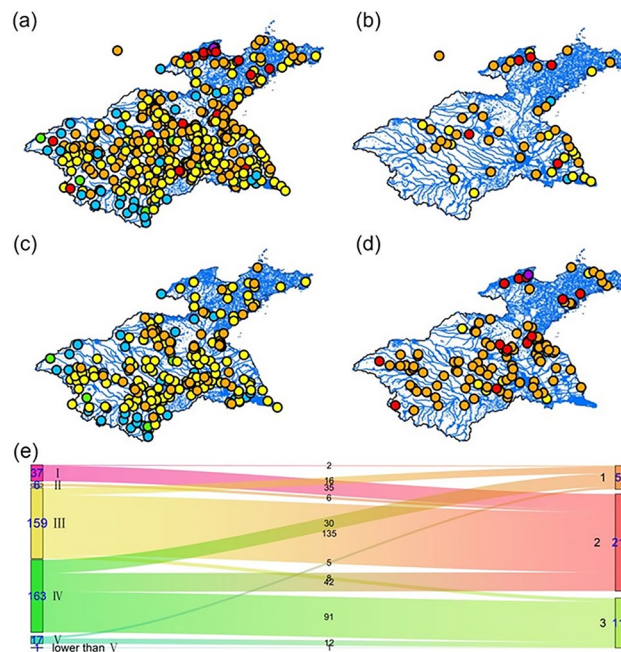


**Fig. 6.** Center distance of categories with different numbers of clusters.



**Fig. 7.** Characteristics of changes in standardized water quality indicators in different clusters.

of the other water quality indicators. Combined with the results of the correlation analysis, this cluster reflects monitoring stations affected by agricultural pollution. For Cluster 2, DO and COD had higher loadings, and the other indicators were lower than those of the other clusters, indicating that this cluster was only slightly polluted by organic matter and that the overall water quality was better. For Cluster 3, the distinguishing features were



**Fig. 8.** Relationship between surface water quality class and cluster result of monitoring stations (a) surface water quality class of all monitoring stations; (b) surface water quality class of monitoring stations in Cluster 1; (c) surface water quality class of monitoring stations in Cluster 2; (d) surface water quality class of monitoring stations in Cluster 3.

higher DO, COD, and BOD<sub>5</sub>. Combined with the correlation analysis findings, the water quality characteristics of this cluster indicate that the monitoring stations were affected by organic matter pollution.

### Spatial change pattern of water quality influencing factors in HRB

The spatial distribution of the different clusters demonstrates a regional pattern of water quality changes at the monitoring sites in the HRB (Fig. 8a). To illustrate the pattern of influence of each cluster on the water quality of the HRB, the cluster category to which the monitoring sites belonged was compared with the water quality class (Fig. 8b–d). The Sankey diagram demonstrates the quantitative correspondence between the cluster category and water quality class (Fig. 8e).

From the spatial distribution of the clusters, Cluster 1 contained a total of 53 monitoring stations, of which 2, 16, 30, and 5 monitoring stations belonged to water quality Class I, II, III, and IV, respectively. These stations were distributed in the eastern coastal area of the HRB as well as in the central and western hinterlands, indicating that nutrients accumulated in the downstream area. In addition, there were relatively few monitoring stations with good water quality grades in Cluster 1, suggesting that agricultural pollution is an important factor leading to river water quality pollution in downstream areas. In Cluster 2, 35, 6, 135, and 42 monitoring stations had water quality class belonging to Class I, II, III, and IV, respectively. The number of monitoring stations with a better water quality status was higher than that of the other clusters because the water quality indicator loads were lower in this cluster. In terms of spatial distribution, monitoring stations with water quality Class I and II were distributed in the upper reaches of the river, suggesting that less pollution in the upstream areas was the main reason for better surface water quality. Compared to Clusters 1 and 2, the water quality status of the monitoring stations in Cluster 3 was significantly worse, indicating that organic pollution was the most important cause of water quality deterioration in the HRB. There were 8, 91, 12, and 1 monitoring stations belonging to Class III, IV, V, and lower than V, respectively. Although most of these stations were located in the middle and downstream areas of the river network, localized river organic enrichment in the western part of the basin and other areas led to a sharp deterioration in water quality, indicating that the localized concentration of pollution is a potential cause of the pollution occurring in the upstream areas of the river network. Overall, the water quality status of the monitoring stations in Cluster 3 was significantly poorer than that of other clusters, indicating that combating organic matter pollution is a top priority for surface water quality control in the HRB.

### Conclusions

As a large-scale surface water system, the spatial distribution of water quality in river basins varies. Compared to previous studies that focused on the temporal variation of water quality indicators, this study successfully revealed the spatial distribution characteristics and temporal change patterns and controlling factors of the surface water quality in the middle and lower plains of the HRB through cluster analysis. Overall, most areas in the HRB exhibited good surface water quality, while low-quality surface water appears in some areas and is affected by local pollution. Specifically, water quality changes in the HRB were primarily reflected in organic

matter pollution indicators (DO, COD, COD<sub>Mn</sub> and BOD<sub>5</sub>) and nutrient salt pollution indicators (TP, TN, and NH<sub>3</sub>-N). In particular, TN has a wide range and is a main factor limiting the quality of water bodies. The entropy weighting method revealed that the overall water quality of the basin was Class III, and heavily polluted areas were distributed downstream. Correlation and cluster analyses revealed three causes of water pollution. The areas affected by agricultural activities (Cluster 1) were mainly concentrated in the lower reaches of the river network and sea outlets and showed higher nutrient salt concentrations. Cluster 2 represented the upstream areas polluted by slight organic matter and had relatively good water quality. Cluster 3 highlighted the severity of organic matter pollution in the middle and lower reaches, which is a key concern for surface water management. These findings emphasize the importance of developing different management strategies for different pollution sources, particularly for controlling agricultural surface pollution and enhancing organic pollutant reduction. In conclusion, this study provides a scientific basis and technical support for the sustainable management of water resources in the HRB and is of great practical value in guiding future water quality protection and improvement measures.

## Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to the confidentiality but are available from the corresponding author on reasonable request.

Received: 15 August 2024; Accepted: 16 May 2025

Published online: 05 June 2025

## References

- Bocaniov, S. A. & Scavia, D. Temporal and spatial dynamics of large lake hypoxia: Integrating statistical and three-dimensional dynamic models to enhance lake management criteria. *Water Resour. Res.* **52**(6), 4247–4263 (2016).
- Park, Y., Pachepsky, Y. A., Cho, K. H., Jeon, D. J. & Kim, J. H. Stressor–response modeling using the 2D water quality model and regression trees to predict chlorophyll-a in a reservoir system. *J. Hydrol.* **529**, 805–815 (2015).
- Wang, Y. B. et al. Effects of land-use patterns on seasonal water quality at multiple spatial scales in the Jialing River, Chongqing, China. *Catena* **234**, 107646 (2024).
- Zhang, L. et al. Linkage and driving mechanisms of antibiotic resistome in surface and ground water: Their responses to land use and seasonal variation. *Water Res.* **215**, 118279 (2022).
- Chen, X. et al. Multi-scale modeling of nutrient pollution in the rivers of China. *Environ. Sci. Technol.* **53**(16), 9614–9625 (2019).
- Lang, M., Li, P. & Yan, X. Runoff concentration and load of nitrogen and phosphorus from a residential area in an intensive agricultural watershed. *Sci. Total Environ.* **458**, 238–245 (2013).
- Nguyen, T. G. & Huynh, N. T. H. Evaluating surface water quality using indexes of water quality and plankton diversity. *Civ. Eng. J.* **9**(5), 1187–1202 (2023).
- Oswald, C. J. et al. Integrating urban water fluxes and moving beyond impervious surface cover: A review. *J. Hydrol.* **618**, 129188 (2023).
- Sérodès, J. B. et al. Tracking domestic wastewater and road de-icing salt in a municipal drinking water reservoir: Acesulfame and chloride as co-tracers. *Water Res.* **203**, 117493 (2021).
- Huang, J. et al. Characterizing the river water quality in China: Recent progress and on-going challenges. *Water Res.* **201**, 117309 (2021).
- Shen, Z., Zhong, Y., Huang, Q. & Chen, L. Identifying non-point source priority management areas in watersheds with multiple functional zones. *Water Res.* **68**, 563–571 (2015).
- Cheng, H. et al. Impacts of different human activities on hydrological drought in the Huaihe River Basin based on scenario comparison. *J. Hydrol. Region. Stud.* **37**, 100909 (2021).
- Gao, C. & Ruan, T. The influence of climate change and human activities on runoff in the middle reaches of the Huaihe River Basin, China. *J. Geog. Sci.* **28**, 79–92 (2018).
- Wang, M. et al. Nonstationary flood and low flow frequency analysis in the upper reaches of Huaihe River Basin, China, using climatic variables and reservoir index as covariates. *J. Hydrol.* **612**, 128266 (2022).
- Wang, H. et al. A novel method for quantifying human disturbances: A case study of Huaihe River Basin, China. *Front. Public Health* **10**, 1120576 (2023).
- Zhang, Y., Xia, J., Liang, T. & Shao, Q. Impact of water projects on river flow regimes and water quality in Huai River Basin. *Water Resour. Manag.* **24**, 889–908 (2010).
- Zhai, X., Xia, J. & Zhang, Y. Water quality variation in the highly disturbed Huai River Basin, China from 1994 to 2005 by multi-statistical analyses. *Sci. Total Environ.* **496**, 594–606 (2014).
- Dou, M., Zhang, Y. & Li, G. Temporal and spatial characteristics of the water pollutant concentration in Huaihe River Basin from 2003 to 2012, China. *Environ. Monit. Assess.* **188**, 1–18 (2016).
- Li, H., Chen, S., Ma, T. & Ruan, X. The quantification of the influencing factors for spatial and temporal variations in surface water quality in recent ten years of the Huaihe River Basin, China. *Environ. Sci. Pollut. Res.* **29**(29), 44490–44503 (2022).
- Xu, J. et al. Assessing temporal variations of Ammonia Nitrogen concentrations and loads in the Huaihe River Basin in relation to policies on pollution source control. *Sci. Total Environ.* **642**, 1386–1395 (2018).
- Beckers, L. M. et al. Unraveling longitudinal pollution patterns of organic micropollutants in a river by non-target screening and cluster analysis. *Sci. Total Environ.* **727**, 138388 (2020).
- Yang, J. et al. Using cluster analysis for understanding spatial and temporal patterns and controlling factors of groundwater geochemistry in a regional aquifer. *J. Hydrol.* **583**, 124594 (2020).
- Yin, Z., Luo, Q., Wu, J., Xu, S. & Wu, J. Identification of the long-term variations of groundwater and their governing factors based on hydrochemical and isotopic data in a river basin. *J. Hydrol.* **592**, 125604 (2021).
- Yu, J. W. et al. Water quality forecasting based on data decomposition, fuzzy clustering and deep learning neural network. *Environ. Pollut.* **303**, 119136 (2022).
- Aytaç, E. Unsupervised learning approach in defining the similarity of catchments: Hydrological response unit based k-means clustering, a demonstration on Western Black Sea Region of Turkey. *Int. Soil Water Conserv. Res.* **8**(3), 321–331 (2020).
- The Ministry of Ecology and Environment, P.R.C. 2020. *Annual Bulletin of China's Ecology and Environment*. <http://www.cnemc.cn/jcbg/zghjzkgb/>.
- Yan, J., Ren, K. & Wang, T. Improving multidimensional normal cloud model to evaluate groundwater quality with grey wolf optimization algorithm and projection pursuit method. *J. Environ. Manag.* **354**, 120279 (2024).
- Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948).
- Chen, P. Effects of normalization on the entropy-based TOPSIS method. *Expert Syst. Appl.* **136**, 33–41 (2019).

30. Wang, D. et al. A multidimension cloud model-based approach for water quality assessment. *Environ. Res.* **149**, 113–121 (2016).
31. Hashemy, S. M. & Monem, M. J. Facilitation of operation and maintenance activities of irrigation networks using ak-means clustering method: case study of the ghazvin irrigation network. *Irrig. Drain.* **61**(1), 31–38 (2012).
32. Liu, Y., Wu, X. & Qi, W. Assessing the water quality in urban river considering the influence of rainstorm flood: A case study of Handan city, China. *Ecol. Indic.* **160**, 111941 (2024).
33. Gundersen, P. & Steinnes, E. Influence of pH and TOC concentration on Cu, Zn, Cd, and Al speciation in rivers. *Water Res.* **37**(2), 307–318 (2003).
34. Zhang, C. et al. Effects of sediment geochemical properties on heavy metal bioavailability. *Environ. Int.* **73**, 270–281 (2014).
35. Cui, H. M. & Yan, J. H. Investigation and analysis of water quality of Xiaohe River in Linyi, Shandong Province. *Environ. Sustain. Dev.* **42**(6), 177–178 (2017).
36. Feng, C. X. et al. Groundwater hydrochemical evolution in the muddy coastal zone of northern Shandong Peninsula. *Marine Geol. Front.* **38**(12), 16–25 (2022).
37. Yuan, R. Y., Zhao, D. M., Zhai, X. Z., Li, C. & Tian, Z. F. Characteristics and source analysis of fluoride ion pollution in the lower reaches of Daqing River Basin. *J. Green Sci. Technol.* **24**(8), 89–93 (2022).
38. Yuan, R. Y. et al. Analysis of priority control sources of heavy metal pollution in sediments of the lower of Daqing River Basin, China. *Environ. Sci.* **44**(7), 4042–4052 (2024).
39. Cui, J. et al. Evaluating the influence of hydrological condition on the phosphorus loads in an agricultural river basin using the SWAT model. *Hydrol. Res.* **52**(5), 1143–1158 (2021).

## Acknowledgements

This work was supported by Project of Intelligent Detection Research Center for Rail Transit Infrastructure (KYPT2023003), Natural Science Foundation of the Higher Education Institutions of Jiangsu Province, China (22KJB520014), Research Project of Jiangsu Province Higher Education Institutions (23KJB560003), Huaian Science and Technology Plan Project Natural Science Research (HAB202365).

## Author contributions

Shuzhen Feng—Data Curation, Formal Analysis, Writing—Original Draft; Chaokai Zhang—Methodology, Software, Visualization; Jiaheng Yan—Conceptualization, Investigation, Methodology, Writing—Review and Editing; Ke Ren—Methodology, Software, Visualization; Ningbo Peng—Resources; Wei Jiang—Validation; Shouhua Liu—Investigation.

## Declarations

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-02964-2>.

**Correspondence** and requests for materials should be addressed to J.Y.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025