

Validating Statistical Zonation for Groundwater Natural Background Levels Using an Integrated SOM-KNN Machine Learning Approach in a Heterogeneous Mountainous Terrain, North China

Zhiqiang Gong

Hebei Geological Environment Monitoring Institute

Xun Zhang

Hebei University of Architecture

Zonglong Zheng

Hebei University of Architecture

Zhiwen Dang

13644076549@163.com

Hebei University of Architecture

Zhisen Ma

Hebei University of Architecture

Research Article

Keywords: Groundwater natural background levels, Self-Organizing Map, K-Nearest Neighbors, Hydrogeological heterogeneity, Statistical characterization

Posted Date: September 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10945655/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Validating Statistical Zonation for Groundwater Natural Background Levels Using an Integrated SOM-KNN Machine Learning Approach in a Heterogeneous Mountainous Terrain, North China

Zhiqiang Gong^{1,#}, Xun Zhang^{1,2,#}, Zonglong Zheng², Zhiwen Dang^{2,*}, Zhisen Ma²

¹ Hebei Key Laboratory of Geological Resources and Environment Monitoring and Protection, Hebei Geological Environment Monitoring Institute, Shijiazhuang 050021, China

² Hebei University of Architecture, Zhangjiakou, Hebei 075000, China

* Corresponding author: Zhiwen Dang

Email: 13644076549@163.com

[#]These authors contributed equally to this work.

Abstract: Determining reliable groundwater natural background levels (NBLs) in heterogeneous mountainous regions remains challenging because multiple flow domains (pore, fracture, and karst) generate complex hydrochemical patterns governed by recharge–flow–discharge processes. A critical yet unresolved question is whether predefined hydrogeological units are consistent with hydrologically controlled groundwater organization. In this study, we develop a process-informed framework that integrates unsupervised Self-Organizing Maps (SOM) and supervised K-Nearest Neighbors (KNN), and use this combined SOM–KNN approach as a diagnostic tool to evaluate whether statistical zonation reflects hydrologically meaningful groundwater organization and to quantify transitions between flow regimes. The framework is applied to a mountainous aquifer system in North China, where 480 groundwater samples were collected across six predefined units. The SOM analysis reveals a dominant mineralization gradient that reflects progressive groundwater evolution from recharge-dominated zones to flow and discharge zones, while trace elements exhibit localized signatures associated with lithological heterogeneity. Projection of hydrogeological unit labels onto the SOM topology demonstrates that most predefined

units are consistent with the underlying hydrologically organized structure, while transitional zones emerge at unit boundaries. The KNN classification further quantifies inter-unit separability and identifies overlapping regions that correspond to hydrologically mixed flow regimes. These overlaps provide a quantitative basis for refining hydrogeological zonation and improving the delineation of NBLs. Cross-validation with conventional statistical approaches confirms the robustness of the identified outliers and transitional samples. Overall, the proposed framework moves beyond purely data-driven classification by explicitly linking hydrochemical patterns to hydrological processes and by providing a transferable approach for validating conceptual hydrogeological models. This contributes to more reliable NBL estimation and supports groundwater quality assessment in complex aquifer systems.

Key words: Groundwater natural background levels; Self-Organizing Map; K-Nearest Neighbors; Hydrogeological heterogeneity; Statistical characterization

1 Introduction

Determining reliable groundwater NBLs is essential for distinguishing natural hydrochemical signatures from anthropogenic pollution and for supporting water quality assessment, source apportionment, and sustainable groundwater management (Bondu et al., 2022; Voutchkova et al., 2021). In mountainous terrains, this task is complicated by pronounced hydrogeological heterogeneity, where pore, fracture, and karst media coexist and produce complex, high-dimensional hydrochemical fingerprints that are inherently difficult to interpret using conventional statistical approaches (Ayed et al., 2017; Hartmann et al., 2014). The coexistence of multiple water-bearing media not only amplifies spatial variability but also generates nonlinear interactions among chemical components, which traditional univariate or weakly multivariate methods cannot adequately resolve (Bi et al., 2023; Khadra et al., 2022). Consequently, reliable NBL characterization in such settings demands analytical techniques capable of handling non-linearity, multi-media complexity, and high-dimensional data structures (Huang et al., 2023). Such capabilities lie beyond the reach of conventional parametric frameworks, which in turn motivates the adoption of data-driven machine learning strategies (Haggerty et al., 2023). This study addresses

this need by developing a hybrid unsupervised-supervised machine learning framework (Shang et al., 2024).

Early efforts to establish groundwater NBLs relied primarily on univariate statistical techniques. The iterative Grubbs test, the mean $\pm$ 2- or 3-standard-deviation criterion, and cumulative probability distribution methods were widely applied across various hydrogeological settings (Busico et al., 2024; Preziosi et al., 2021). For instance, in Xingtai City of the North China Plain, cumulative probability statistics combined with multivariate techniques were employed to determine NBLs for major ions, successfully identifying groundwater over-extraction and industrial activities as primary drivers of contamination (Q. Wang et al., 2025). While such approaches provided a quantitative basis for distinguishing natural from anthropogenic influences, they suffer from a fundamental limitation: they assume that the underlying data follow a specific distribution (e.g., normality) and that the predefined spatial units are chemically homogeneous, assumptions frequently violated in heterogeneous mountainous terrains where pore, fracture, and karst media coexist (Parrone et al., 2022; Zanotti et al., 2022). Consequently, these methods risk either retaining anthropogenic outliers as “natural” extremes or excluding genuine natural variability as “anomalies,” leading to biased background estimates (Vieira et al., 2026). To address these shortcomings, we gradually employed multivariate approaches such as hierarchical cluster analysis, principal component analysis, and pre-screening procedures (Sako & Ouangaré, 2023). In the Pinggu basin of Beijing, for example, a combined approach of cluster analysis, iterative Grubbs’ test, median $\pm$ 2MAD, and nitrate isotopes successfully distinguished industrial, agricultural, and domestic sources of groundwater contamination (Li et al., 2021). Similarly, in the Guadalhorce River basin of southern Spain, the iterative 2σ technique and distribution function method were compared, providing recommendations for their respective application under varying hydrogeological conditions (Urresti-Estala et al., 2013). However, these approaches are not formulated as integrated modeling frameworks, which limits their ability to capture complex data structures and to provide consistent classification across heterogeneous hydrogeological settings.

More recently, machine learning algorithms have begun to appear in NBL studies, primarily for outlier detection (e.g., Isolation Forest) or hydrochemical pattern recognition (e.g., KNN, random forest)(Haggerty et al., 2023; Q. Liu et al., 2026). In the Pearl River Delta, a coupled Cl/Br ratio, KNN algorithm, and empirical Bayesian kriging approach was applied to quantify nitrate NBLs across hydrogeological units, revealing substantial spatial heterogeneity linked to permeability and redox variations (R. Liu et al., 2026). While these studies demonstrate the applicability of machine learning, they are typically not designed as integrated modeling strategies, limiting their ability to jointly represent data structure and classification behavior in complex hydrochemical systems. As a result, the modeling consistency and interpretability of hydrochemical zonation remain insufficiently addressed.

This leaves three interconnected research gaps that must be addressed to advance NBL characterization in heterogeneous mountainous settings. First, existing methods predominantly estimate background values within predefined zones, but they lack an independent, data-driven means to verify whether those zones truly reflect underlying hydrochemical clusters rather than arbitrary administrative or geometric partitions (Y. Wang et al., 2025; Zhou et al., 2023). Second, the degree of chemical overlap between adjacent statistical units has never been systematically quantified, leaving uncertainty as to whether current unit boundaries should be refined, merged, or completely redefined (Li et al., 2025; Zhou et al., 2023). Third, these two deficiencies are particularly acute in terrains where pore, fracture, and karst media coexist, precisely the settings where conventional univariate and multivariate methods are most prone to failure due to violated distributional assumptions and unaccounted spatial heterogeneity (Costanzo et al., 2026). Together, these gaps highlight the urgent need for a framework that not only validates the hydrochemical coherence of statistical units but also quantifies inter-unit transitions, thereby providing a robust basis for background value determination in hydrogeologically complex regions.

To fill these gaps, this study presents a hybrid SOM–KNN modeling strategy that integrates unsupervised mapping and supervised classification to improve the representation and interpretation of complex hydrochemical data. The unsupervised

SOM extracts inherent hydrochemical topologies to verify whether predefined units are chemically meaningful, directly addressing the first gap. The supervised KNN quantifies inter-unit separability through classification accuracy and identifies compositional outliers, directly addressing the second gap. Together, this integrated approach offers three major advances over conventional methods: explicit validation of statistical unit rationality, data-driven outlier verification, and quantification of inter-unit overlaps. The framework provides a transferable, data-driven methodology for NBL characterization in hydrogeologically complex regions, supporting science-based groundwater quality assessment and pollution source identification.

2 Study Area

2.1 Geographical and geological setting

Situated between latitudes 39°40'-40°28'N and longitudes 117°31'- 118°55'E, the study area lies in the northern sector of Tangshan City, Hebei Province, China. The mountainous area of Tangshan City covers approximately 5131 km², with ground elevations generally ranging from 50 to 842 m above sea level; the highest point is the Bameian Peak in Qianxi County at 842 m. The landscape has been shaped by prolonged erosion and denudation, resulting in complex mountainous terrain characterized by rugged ridges and incised valleys (Dang et al., 2025; Ma et al., 2014). Several large intermountain basins are developed within the area, including the Zunhua Basin, the Qian'an Basin, and the Hazelzizhen Basin. These basins are topographically open and flat, with rivers meandering through their central parts and well-developed river terraces. Regional groundwater flow is intrinsically governed by topography and geomorphology. In mountainous areas, the dominant flow direction is from north to south, closely following the elevation gradient that declines from highlands toward basins and alluvial plains (C. Wang et al., 2025); in the basin areas, flow converges from the periphery toward the center, with the Qian'an and Zunhua basins exhibiting a dominant northwest-to-southeast flow direction. Historical water quality data indicate that the predominant hydrochemical type in the mountainous area is bicarbonate-rich water, with only localized occurrences of sulfate-type water; TDS are generally below 0.5 g/L in most of the area(Wen et al., 2025).

The study area spans multiple stratigraphic units from the Proterozoic to the Cenozoic (Fig. 1). The Jixian and Qingbaikou systems (Mesoproterozoic) are predominantly composed of dolomite and sandy conglomerate, which serve as the primary carbonate aquifers in the recharge zones (Xiao et al., 2014). The Cambrian-Ordovician carbonate sequences occur extensively across the central and northern sectors of the study region. These sequences include the FuJUNSHAN Formation (thick-bedded leopard limestone, dolomitic limestone, and dolomite) and the Majiagou Formation (limestone and dolomite). These carbonate units are the principal sources of Ca^{2+} , Mg^{2+} , and HCO_3^- through calcite and dolomite dissolution, governing the dominant bicarbonate-rich hydrochemical type observed in most statistical units (Wu et al., 2025). The Carboniferous-Permian coal-bearing sequences, which contain significant pyrite (FeS_2) and are associated with the Kailuan coalfield, are distributed in the eastern and southern basin margins. Oxidation of pyrite within these strata releases SO_4^{2-} and H^+ , which likely accounts for the elevated sulfate levels observed in the transitional zones (Cai et al., 2025). The Quaternary system is composed of Pleistocene loess, gravel, coarse-medium sand, and sandy clay. In the basin areas, these lithologies contain unconsolidated pore water and constitute the principal groundwater storage units for the unconfined aquifers.

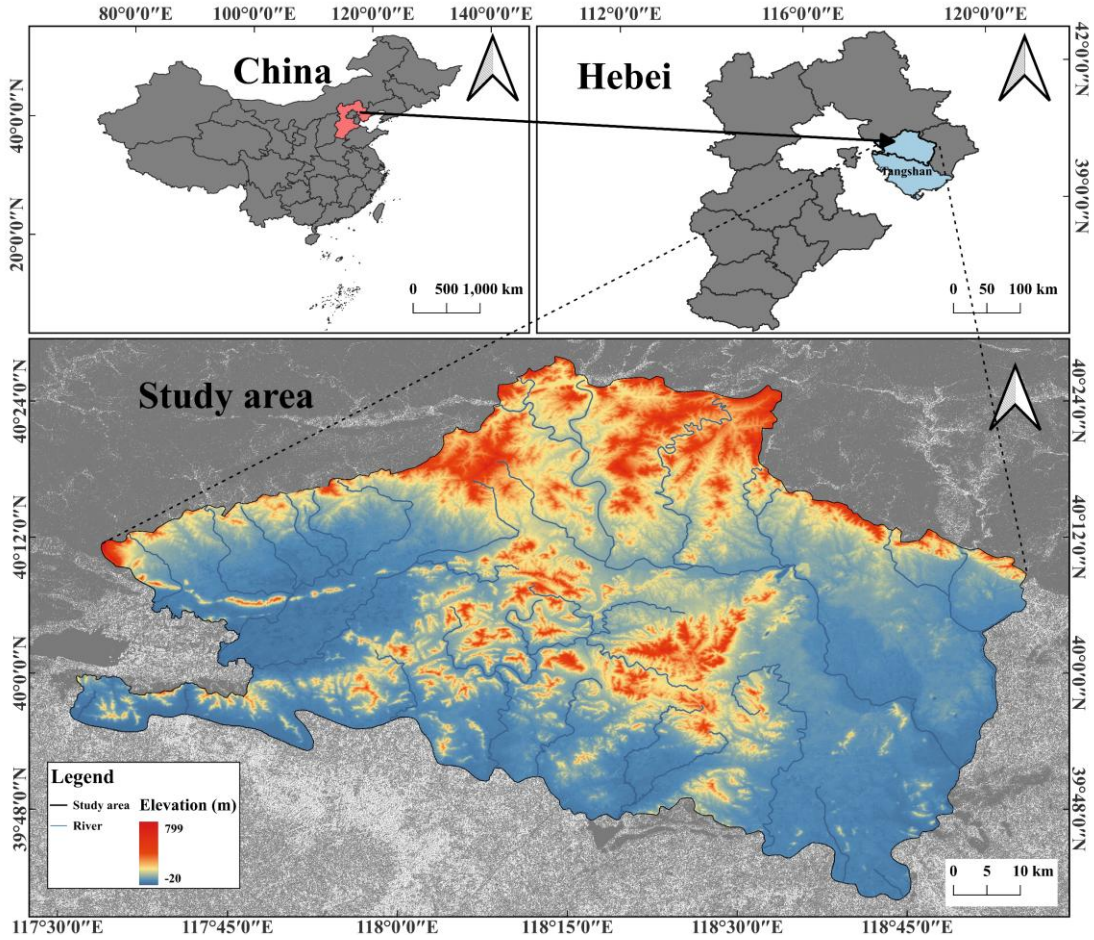


Fig. 1. Location and elevation of the study area in the Tangshan mountainous area

Based on geological structure, stratigraphic distribution, and groundwater drainage conditions, the mountainous area of Tangshan is divided into four level hydrogeological zones (Fig. 2)(Dang et al., 2025): (1) the Luanhe River Basin magmatic and metamorphic rock area, characterized by low-permeability crystalline rocks that serve as localized recharge zones with limited aquifer storage (Dang et al., 2026); (2) the East Hebei coastal Zhuhe metamorphic rock area, dominated by fractured metamorphic rocks with moderate permeability; (3) the Chaobaiji Canal magmatic and metamorphic rock area, where magmatic intrusions have created complex fracture networks that facilitate groundwater flow (Wen et al., 2025); and (4) the Zunhua Basin unconsolidated sediment area, composed of thick Quaternary alluvial deposits that form the primary groundwater reservoir and discharge zone in the northern basin.

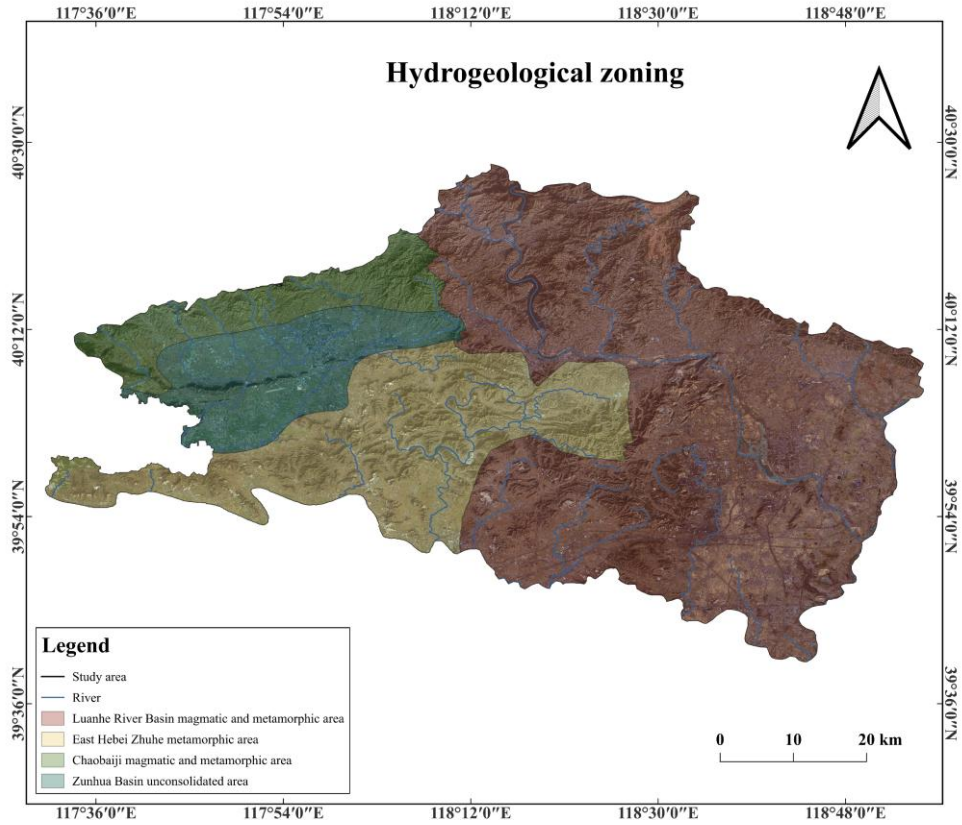


Fig 2. Geological and hydrogeological setting of the Tangshan mountainous area

2.2 Hydrogeological zoning and statistical units

The division of statistical units followed the technical requirements of the Statistical Characterization of Groundwater Environmental Background Values (Trial), combined with local hydrogeological conditions (Ministry of Ecology and Environment, 2023). The basis for unit division included: (1) distribution of hydrochemical types, (2) TDS concentration contours (generally grouped into ≤ 0.5 , $0.5 - 1$, $1 - 3$, $3 - 5$, > 5 g/L), and (3) abrupt zones of aquifer and groundwater runoff conditions (e.g., alluvial fan fronts) (He et al., 2022). Areas with similar hydrochemical types and TDS ranges were first grouped together, followed by fine-tuning based on the geographic patterns of each background component.

In the study area, groundwater chemical types are broadly similar and TDS is consistently below 5 g/L. Based on a four-level hydrogeological zoning framework and the distinction of water-bearing media (pore water, fracture water, karst water), the area was preliminarily divided into six statistical units (Fig. 3), hereafter referred to as Units I-VI.

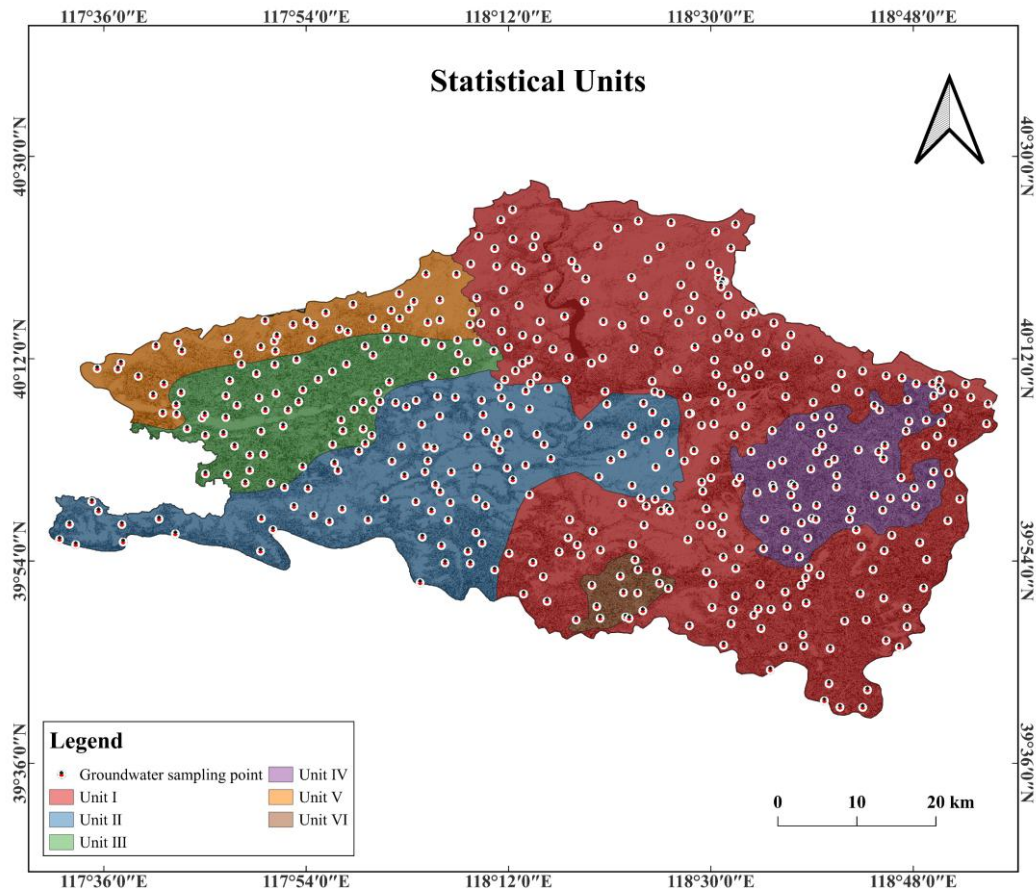


Fig.3. Statistical units and sampling sites in North Tangshan mountainous area

3 Materials and Methods

3.1 Groundwater sampling and analysis

Twelve hydrochemical parameters were selected for background value characterization: eight major ions (Na^+ , K^+ , Ca^{2+} , Mg^{2+} , Cl^- , SO_4^{2-} , HCO_3^- , CO_3^{2-}), TDS and three trace components (F^- , Mn, As). The major ions and TDS serve as the primary indicators for evaluating conventional groundwater quality and identifying anthropogenic contamination signals (Voutchkova et al., 2021). F^- and Mn were selected as trace components due to their prevalence in local carbonate aquifers and their sensitivity to pH and redox variations, which are critical for assessing the vulnerability of groundwater to geogenic contamination (Amiri et al., 2021). Arsenic, though present at ultratrace levels, was included because of its potential mobilization under reducing conditions, a phenomenon documented in neighboring basins of the North China Plain that warrants systematic evaluation in the Tangshan mountainous area (Li et al., 2024).

According to the requirements of the Statistical Characterization of Groundwater Environmental Background Values (Trial), the study area, characterized by complex mountainous terrain with an existing survey accuracy of 1:100,000, requires no fewer than 6 survey points per 100 km². Following these guidelines, sampling points were arranged to collect groundwater samples from the six statistical units. In the field, the collected samples were passed through 0.45 µm membrane filters, then placed into pre-rinsed polyethylene bottles, and finally delivered to the laboratory at 4 °C within 48 h. Major ions were analyzed by ion chromatography, TDS was determined gravimetrically, and trace components were measured using inductively coupled plasma mass spectrometry (ICP-MS). Quality control was ensured by analyzing reagent blanks, certified reference materials, and duplicate samples every 20 samples, with relative standard deviations consistently below 5 %.

Data screening followed the Statistical Characterization of Groundwater Environmental Background Values (Trial), excluding samples with unknown layer information, suspected contamination, anion-cation imbalance, abnormal macro-component content, incomplete test results, or unreliable analytical data. After this screening, a total of 480 groups of samples were retained for statistical characterization of groundwater background values, meeting the minimum sample size and spatial distribution density requirements for background value statistics (Fig. 3).

3.2 Data preprocessing and initial outlier identification

Normality of the background data in each statistical unit was examined using the Shapiro-Wilk test. For data that deviated significantly from normality ($p < 0.05$), transformations including Box-Cox, logarithmic, square root, reciprocal, or square root arcsine were applied to approximate a normal distribution (Joseph & Bhaumik, 1997). Only data that conformed to normality (after transformation, if applied) were subjected to subsequent parametric outlier identification methods.

Outliers for major ions and TDS were initially flagged by visual inspection of Piper trilinear diagrams. Quantitative outlier removal then followed a sample-size dependent strategy: for statistical units with ≤ 100 samples, the Grubbs test ($\alpha = 0.05$) was applied iteratively; for units with > 100 samples, the mean $\pm 3\sigma$ criterion was used

(He et al., 2022). When a sample was identified as an outlier for any major component, all its hydrochemical parameters (including TDS) were excluded. For trace components (F^- , Mn, As, NO_3^-), samples with a detection rate $>75\%$ had undetected values replaced by half the detection limit, and the same Grubbs/ 3σ procedure was applied. For components with a detection rate between 25 % and 75 %, only detected values were retained; if the detection rate was $< 25\%$, the detection limit itself was taken as the upper bound of the NBL. Given the non-point agricultural influence in the study area, NO_3^- concentrations exceeding 10 mg/L were directly removed as anthropogenic outliers (Parrone et al., 2019).

All removed outliers were manually checked against field records, local geology, and potential pollution sources to confirm that they did not represent natural extreme values (He et al., 2022). Samples from small areas with hydrochemical characteristics distinctly different from their assigned statistical unit were excluded from that unit and considered separately. After the above preprocessing and outlier removal, a total of 405 clean samples were retained for subsequent statistical characterization and machine learning analysis.

3.3 Statistical characterization of natural background levels

According to the distribution of each background value index, the rationality of the division of statistical units is analyzed, and the unreasonable statistical units are adjusted (Zhou et al., 2023). The background value data in each statistical unit after adjustment were counted. Analyze the causes of background values within the scope of investigation (Sako & Ouangaré, 2023). The maximum value, 5 % quantile and 95 % quantile of the fractional data of each group in each statistical unit after adjustment are used as the data basis to characterize the groundwater environmental background value (Masciale et al., 2021). The lower limit of the groundwater environmental background value is 5 % quantile and the upper limit is 95 % quantile (Parrone et al., 2019).

3.4 Self-organizing map (SOM)

The SOM is defined as an unsupervised ANN that transforms high-dimensional input vectors into a low-dimensional (commonly two-dimensional) topological grid while retaining the topological relations of the original data (Lazarev et al., 2025). In

this study, SOM was applied to reveal the intrinsic clustering structure of groundwater samples according to their hydrochemical signatures (Rahman et al., 2022). The network architecture of a SOM comprises an input layer and a mapping layer, the latter also termed the competitive layer. For each input vector in Eq. (1) and Eq. (2).

$$X=(x_1, x_2, \dots, x_m)^T \quad (1)$$

$$W_j=(x_{1j}, x_{2j}, \dots, x_{mj})^T \quad (2)$$

where m is the number of input variables (12 hydrochemical parameters), each neuron j in the mapping layer is associated with a weight vector.

First, the Euclidean distance between the input vector X and W_j each weight vector is calculated, as shown in Eq. (3), and the neuron with the smallest distance is declared the winner.

$$d_j = \|X - W_j\| = \sqrt{\sum_{i=1}^m (x_i(t) - w_{ij}(t))^2} \quad (3)$$

Second, the weight vectors of the BMU and its neighboring neurons are adjusted to move closer to the input vector according to the following update rule in Eq. (4) and Eq. (5).

$$\Delta w_{ij} = w_{ij}(t+1) - w_{ij}(t) = \rho(t)(x_i(t) - w_{ij}(t)) \quad (4)$$

$$\rho(t) = \frac{1}{t} \quad (5)$$

where $0 < \rho(t) < 1$ and $\rho(t)$ decreases to 0 with time change, and $x_i(t)$ and $w_{ij}(t)$ denote the input signal and the connection weight between input neuron i and map neuron j at iteration t , respectively.

Third, the final output for a given input is the BMU's position on the grid, as shown in Eq. (6).

$$O_k = f\left(\min \|X - W_j\|\right) \quad (6)$$

where O_k is a nonlinear function (here a simple assignment to the BMU coordinates). After training, each sample is assigned to a specific neuron, and neurons that are close on the grid represent samples with similar hydrochemical characteristics.

To determine the appropriate SOM grid size, we followed the heuristic guideline

of $5\sqrt{N}$, where N is the number of samples (D'Urso et al., 2020). For the 405 clean samples retained after outlier removal, this yields approximately 100 neurons. However, considering the relatively limited spatial extent and the need for a sufficiently detailed yet interpretable topological representation of the six predefined statistical units, a 5×6 hexagonal grid (30 neurons) was selected. This configuration ensures that each neuron represents approximately 13 samples on average, providing adequate resolution to capture the primary hydrochemical gradients while maintaining visual clarity for subsequent cluster interpretation. To validate this choice, we evaluated the mapping accuracy and topological preservation across a range of grid sizes (Bayat et al., 2025). The selected grid yielded a quantization error of 1.7511 and a topographic error of 0.0099, indicating that the 5×6 configuration effectively preserves the intrinsic nonlinear structure of the high-dimensional hydrochemical dataset (Rahman et al., 2022). The exceptionally low topographic error demonstrates that hydrochemical similarities among samples are consistently translated into spatial proximity on the SOM grid, providing a robust foundation for the subsequent validation of statistical unit rationality while avoiding over-fitting associated with excessive neurons.

The SOM was trained using the selforgmap function in MATLAB's Deep Learning Toolbox, with the grid configured as a 5×6 hexagonal topology. After training, the neurons were further clustered using hierarchical clustering with the Ward linkage method, and the optimal number of clusters was determined by the Davies–Bouldin index (S. Liu et al., 2023). The resulting SOM-derived clusters were then compared with the original six hydrogeological statistical units to evaluate the consistency between data-driven clustering and predefined zonation (Shang et al., 2024).

3.5 Outlier detection using K-nearest neighbor (KNN)

The KNN algorithm was employed as an independent, non-parametric verification of the outlier removal procedure described in Section 3.2. KNN is a distance-based method (Euclidean distance) that makes no assumptions about the underlying data distribution (Chen et al., 2021), making it particularly suitable for hydrochemical data that often deviate from normality (R. Liu et al., 2026). The fundamental principle involves sequentially calculating the average distance between each sample point and

its K nearest neighbors and then comparing the computed distance against a threshold to identify potential anomalies.

To avoid subjective judgment in parameter selection, the elbow method based on variance stability was applied to automatically determine the optimal K value for each statistical unit (Rhyu et al., 2024). The total variance change across all candidate K values is defined as in Eq. (7) and Eq. (8).

$$S_t = \sum_{i=1}^{n-1} |\text{Var}(K_{i+1}) - \text{Var}(K_i)| \quad (7)$$

where $\text{Var}(K_i)$ represents the local variance of average neighbor distances for a given K_i .

$$\sum_{i=1}^k |\text{Var}(K_{i+1}) - \text{Var}(K_i)| \geq 0.8S_t \quad (8)$$

where, $|\text{Var}(K_{i+1}) - \text{Var}(K_i)|$ is an absolute difference in variance at two neighboring values of K; S_t is the total variance change. This criterion identifies the “elbow point” where the incremental gain in variance stabilization diminishes, and 80% of the total variance change has been accumulated.

The 80% threshold in Eq. (8) is a commonly adopted empirical value in variance-based elbow methods for anomaly detection, as it balances sensitivity and robustness by capturing the point where the marginal gain in variance stabilization substantially diminishes (Elumalai et al., 2017; Ramos Marinho et al., 2020). To assess the sensitivity of this threshold, we performed a sensitivity analysis by varying the cumulative variance ratio from 70 % to 90 % at 5 % intervals. The number of detected outliers remained stable across this range (varying by less than 5 % for most units), confirming that the 80% criterion is not overly sensitive to the specific threshold value and that the elbow detection is robust against minor perturbations of this parameter (R. Liu et al., 2026). This automated approach eliminates subjective judgment in parameter selection and ensures consistent outlier detection across the six hydrogeological statistical units. Samples with average distances exceeding a pre-defined threshold (mean plus three standard deviations of the average distance distribution) were flagged as potential anomalies. Importantly, this KNN-based anomaly detection was used solely

for cross-validation purposes to assess the robustness of the parametric outlier removal applied earlier; no samples were excluded from the dataset based on this KNN detection. The classification experiments presented later in this study were conducted on the samples that passed the initial Grubbs or 3σ screening only, without further KNN-based exclusion, to provide an unbiased estimate of unit separability.

4 Results and discussion

4.1 Hydrochemical facies of the delineated statistical units

The Piper trilinear diagrams of the six statistical units reveal the overall hydrochemical characteristics of groundwater in the mountainous area of Tangshan (Fig. 4)(Sako & Ouangaré, 2023). On the ternary diagram for cations, most of the sample points are concentrated close to the Ca^{2+} apex, which suggests that calcium prevails as the dominant cation across nearly all statistical units. In the anion ternary plot, most samples plot in the HCO_3^- region, while a subordinate fraction shifts toward the SO_4^{2-} corner, suggesting bicarbonate-dominant water with sporadic sulfate enrichment. Overall, two major hydrochemical facies are identified: $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ type and $\text{Cl}\cdot\text{SO}_4\text{-Ca}\cdot\text{Mg}$ type. The $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ type is the most widespread, characterizing the background water chemistry of most units (Dang et al., 2025). The $\text{Cl}\cdot\text{SO}_4\text{-Ca}\cdot\text{Mg}$ type, however, is primarily observed in the northern parts of the Qian'an Basin and the Zunhua Basin, mainly corresponding to the upper halves of Units I, III, and IV in Fig. 3, where enhanced evaporation or water-rock interaction involving evaporites may occur (Liu et al., 2025). Units I, III, and IV exhibit a mixed hydrochemical signature with both $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ and $\text{Cl}\cdot\text{SO}_4\text{-Ca}\cdot\text{Mg}$ types. In the anion ternary plot, these samples drift from HCO_3^- toward SO_4^{2-} and Cl^- , indicating enhanced evaporation, evaporite dissolution, or prolonged water-rock interaction (Filipović et al., 2024). In contrast, Unit II is tightly clustered in the $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ field, reflecting carbonate dissolution and short flow paths without anthropogenic signals (Zhang et al., 2025). Units V-VI remain dominated by $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ type with a minor transitional shift. In the cation ternary plot, some samples from Units I, III, IV deviate from the bicarbonate corner toward the no-dominant region, suggesting local Mg^{2+} or Na^+ enrichment via ion exchange or mixing (Liu et al., 2025).

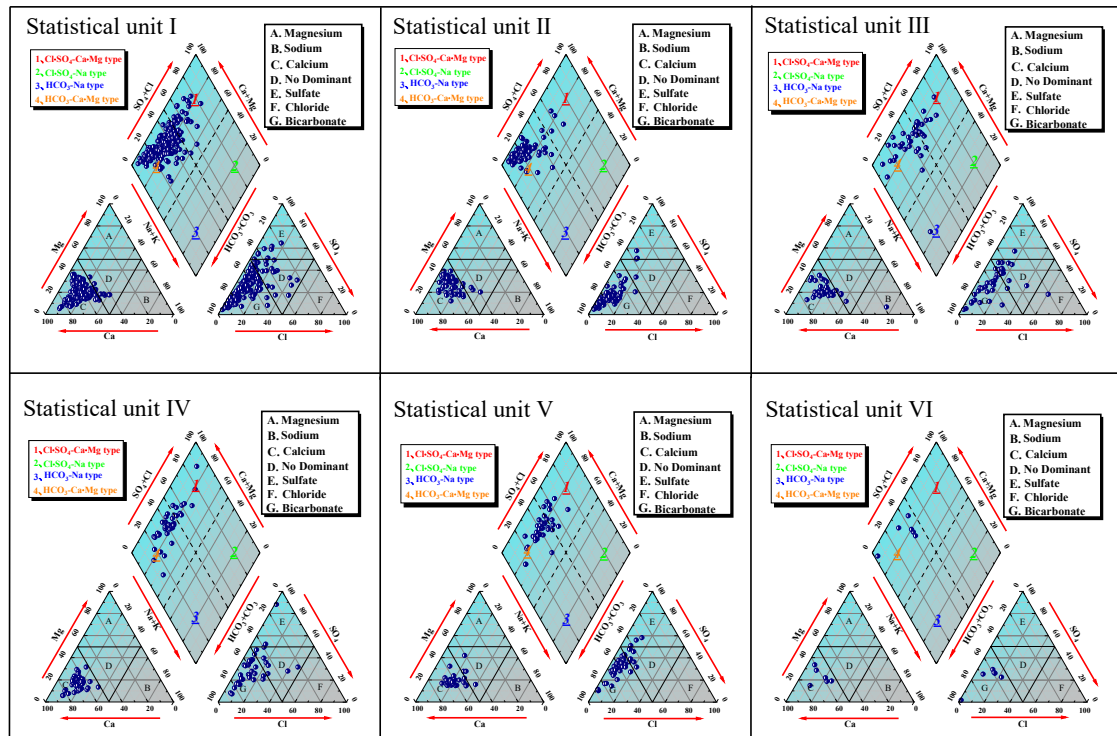


Fig. 4. Piper three-wire distribution for each statistical unit

4.2 Background value statistics

The background value statistics for the six statistical units are summarized in Table 1. According to the national groundwater quality standard (GB/T 14848-2017), each background value corresponds to a certain water quality class (General Administration of Quality Supervision & Standardization Administration of the P.R.C., 2017). Overall, TDS shows a clear increasing trend from the mountainous recharge zones (Units I-III) towards the basin-margin areas (Units IV-VI). A unit-by-unit comparison reveals distinct hydrochemical signatures. These differences validate the necessity of separate statistical treatment (Zhou et al., 2023).

Unit I (n=199) has the largest sample size. Its TDS range is moderate (P5-P95: 219-915 mg/L), corresponding to Class I-III water. The cation composition is dominated by Ca^{2+} (43.5-179 mg/L) and Mg^{2+} (11.9-54.1 mg/L). HCO_3^- (88-343 mg/L) is the predominant anion. This represents a typical $\text{HCO}_3\text{-Ca-Mg}$ type background. F⁻ (P95=0.43 mg/L), Mn (P95=0.01 mg/L) and As (P95=0.0014 mg/L) all fall within Class I. With the widest background interval and the largest sample size, Unit I represents the most typical natural groundwater background of the study area, with minimal anthropogenic disturbance. Unit II (n=79) exhibits a slightly narrower TDS

interval (280-809 mg/L, Class I-III). Its Cl^- and SO_4^{2-} background levels are lower than in Unit I. Ca^{2+} (57.9 - 132 mg/L) and Mg^{2+} (24.7 - 66.3 mg/L) remain dominant. HCO_3^- (146 - 383 mg/L) is even more dominant, indicating stronger carbonate dissolution control (T. Zhang et al., 2023). F^- (P95 = 0.91 mg/L), Mn (P95=0.01 mg/L) and As (P95=0.0013 mg/L) are all Class I. Unit II represents a “cleaner”, low-mineralization groundwater endmember compared to Unit I. Unit III (n=42) has a TDS range (236 - 779 mg/L, Class I-III) similar to Unit II, but its upper TDS values approach those of Unit VI. Elevated P95 values of Na^+ (49.9 mg/L), Cl^- (176 mg/L), and SO_4^{2-} (170 mg/L) indicate a transition towards $\text{Cl}\cdot\text{SO}_4$ type water (Du et al., 2026). F^- P95 reaches 1.29 mg/L (Class IV). Mn P95 is 0.02 mg/L (Class I). This transitional signature reflects local evaporite dissolution or prolonged water-rock interaction. Unit IV (n = 40) displays a wide TDS range (162 - 968 mg/L, Class I-IV), indicating strong internal heterogeneity. Its P5 Na^+ (10.0 mg/L) is higher than in Units I-III. The HCO_3^- range (77 - 253 mg/L) is relatively narrow. This suggests a mixed $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ type. F^- (P95=0.43 mg/L, Class I), Mn (P95 = 0.13 mg/L, Class III), and As (P95 = 0.0089 mg/L, Class I-III) show moderate enrichment. The low-TDS end has elevated Na^+ , and the high-TDS end has elevated SO_4/Cl . Unit 5 (n = 39) records the lowest TDS P95 (697 mg/L, Class I-III) among all units. Its major ion concentrations are uniformly low. Ca^{2+} (30.9 - 127 mg/L) and Mg^{2+} (9.1 - 46.7 mg/L) are relatively low, while HCO_3^- (73.9 - 285 mg/L) dominates. F^- (P95 = 0.61 mg/L, Class I), Mn (P95 = 0.01 mg/L, Class I) and As (P95 = 0.00076 mg/L, Class I) remain very low. Unit V represents a weakly evolved, low-mineralization groundwater endmember. Unit VI (n = 6) has a small sample size (Zhou et al., 2023). It shows the highest TDS range (320 - 1357 mg/L, Class I-IV) and the highest P95 values for most ions (Ca^{2+} up to 319 mg/L, HCO_3^- up to 549 mg/L, Cl^- up to 197 mg/L). F^- P95 reaches 1.57 mg/L (Class IV). Mn P95 is 0.75 mg/L (Class III). As P95 is 0.0010 mg/L (Class I). These elevated background values are natural, reflecting prolonged water-rock interaction and evaporative concentration. The limited sample size (n = 6) warrants caution in interpreting these statistics.

Trace components (F^- , Mn, As): Across all units, F^- background upper limits (P95)

are generally below 1.0 mg/L (Class I-III), except for Unit III (1.29 mg/L, Class IV) and Unit VI (1.57 mg/L, Class IV). Mn remains low (Class I) in Units I-V but reaches Class III in Unit VI. Arsenic stays within Class I-III across all units. These patterns indicate that trace elements are controlled by local redox conditions or mineral sources rather than by the regional TDS gradient (Zhou et al., 2023). In summary, the six units show distinct background ranges and hydrochemical signatures. They range from low-TDS $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ type with Class I-III water (Units I-II), through transitional mixed types with occasional Class IV F^- (Units III-V), to the high-TDS $\text{HCO}_3\text{-dominated}$ endmember with Class IV F^- and Class III Mn (Unit VI). This provides a robust quantitative baseline for groundwater quality assessment. The distinct water quality classes derived from unit-specific background values highlight that using a single fixed threshold for the entire region would lead to erroneous conclusions. This underlines the necessity of a unit-specific approach for reliable background characterization in heterogeneous mountainous terrains (Zanotti et al., 2022).

Table 1. Descriptive statistics of the 12 hydrochemical parameters for each statistical unit (mg/L).

unit		TDS	Na^+	K^+	Ca^{2+}	Mg^{2+}	Cl^-	SO_4^{2-}	HCO_3^-	CO_3^{2-}	F^-	Mn	As
Unit I (n=199)	Min	135	2.93	0.31	25.2	7.83	3.73	13.1	38	0.002	0.01	0.01	9.00E-05
	Max	1237	118	46.7	237	98.2	193	298	472	1.8	0.73	0.74	3.36E-03
	P5	219.4	5.81	0.54	43.49	11.89	8.1	23.49	88	0.03	0.01	0.01	9.00E-05
	P95	915.4	49.1	10.84	179	54.05	128.9	215	343.4	0.85	0.43	0.01	1.44E-03
	Med	392	17.7	2.12	75.7	26.6	22.4	73.9	201	0.24	0.21	0.01	3.00E-04
Unit II (n=79)	Min	159	2.45	0.38	29.8	8.78	5.48	10.4	66	0.007	0.006	0.01	1.20E-04
	Max	1066	68.6	37.3	150	84.3	104	232	476	1.16	0.91	0.09	2.94E-03
	P5	279.8	3.04	0.5	57.94	24.68	7.45	17.25	145.6	0.09	0.11	0.01	1.50E-04
	P95	809.2	50.06	11.82	132.3	66.33	90.24	134.5	383	0.91	0.65	0.01	1.27E-03
	Med	422	12.8	1.38	78.3	40.9	23.4	57.6	289	0.48	0.31	0.01	4.10E-04
Unit III (n=42)	Min	221	4.41	0.44	35.7	8.89	4.95	12	74	0.02	0.006	0.01	9.00E-05
	Max	1338	189	16	279	78.1	569	217	514	2.06	1.85	0.03	1.67E-03
	P5	235.7	6.19	0.49	42	14.55	13.83	32.23	88.1	0.04	0.006	0.01	9.00E-05
	P95	778.6	49.93	10.67	185.1	54.57	175.5	169.8	416.5	0.84	1.29	0.02	1.28E-03
	Med	504	24.15	1.65	85.2	30	41.05	92.65	192	0.15	0.22	0.01	2.20E-04
Unit IV (n=40)	Min	127	8.03	0.37	29.3	4.28	6.35	9.21	59	0.01	0.006	0.01	9.00E-05
	Max	1339	78.4	5.09	239	61.7	284	645	268	1.29	0.74	0.65	1.55E-02
	P5	162	10.03	0.48	36.17	7.52	10.87	24.6	76.95	0.02	0.006	0.01	1.10E-04

	P95	968.1	55.79	3.76	177.2	54.97	121.7	201.4	253.4	0.75	0.43	0.13	8.94E-03
	Med	455	19.5	1.46	91.4	25.8	27.35	97.9	179.5	0.25	0.2	0.01	5.95E-04
Unit V (n=39)	Min	133	9.25	0.72	26	7.39	2.8	13.5	63	0.03	0.006	0.01	9.00E-05
	Max	1539	109	27.7	200	118	148	210	404	0.91	1.23	0.01	2.22E-03
	P5	159.8	9.56	0.93	30.9	9.13	6.19	31.8	73.9	0.03	0.006	0.01	9.00E-05
	P95	696.9	50.08	5.74	127	46.69	75.22	199.7	285.1	0.61	0.66	0.01	7.56E-04
	Med	356	16.5	2.97	72.2	22.1	21.3	77	149	0.18	0.18	0.01	1.60E-04
Unit VI (n=6)	Min	247	4.36	0.29	54.6	30.2	3.59	4.9	205	0.31	0.27	0.01	3.00E-04
	Max	1414	89.4	2.28	349	72.9	213	247	556	1.66	0.8	0.02	1.19E-03
	P5	320.25	6.65	0.33	65.03	30.95	9.89	29.43	255.2	0.39	0.28	0.01	3.25E-04
	P95	1356.5	81.78	2.18	319.2	72.28	196.7	238.7	549	1.57	0.75	0.02	1.04E-03
	Med	748	48.85	1.09	137	55.65	75.15	154.5	368.5	1.05	0.41	0.01	4.20E-04

4.3 Unsupervised SOM clustering for hydrochemical pattern recognition

The SOM was applied to the entire cleaned dataset (n=405 samples, 12 hydrochemical parameters) to explore the intrinsic clustering structure of groundwater chemistry without using any prior unit labels (Zhang et al., 2026; Zhong et al., 2022). The SOM was configured as a 5×6 hexagonal grid (30 neurons), trained with PCA-initialized weights and a decaying learning rate over 200 epochs (Bayat et al., 2025). The resulting component planes (Fig. 5) visualize the distribution of each parameter across the SOM grid, where warm colors (yellow to brown) indicate high normalized values and cool colors (blue to green) indicate low values (Amiri & Nakagawa, 2021).

4.3.1 Dominant hydrochemical gradient revealed by SOM

The component plane of TDS exhibits a clear and systematic gradient: low values consistently occupy the lower left corner, while high values progressively increase toward the upper right. This gradient represents the primary hydrochemical evolution axis, reflecting increasing mineralization from recharge zones (southwestern mountains) to discharge zones (northern basin margins)(Z. Liu et al., 2023). The major ions Na^+ , Ca^{2+} , Mg^{2+} , Cl^- , SO_4^{2-} , and HCO_3^- all show strong spatial concordance with this TDS gradient, confirming that TDS is co-determined by these dominant ions.

This pattern is consistent with progressive water-rock interaction along the regional flow path. In the recharge zones (Units I-III), groundwater has low TDS and is

of $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ type, reflecting active carbonate dissolution. The Cambrian-Ordovician carbonate sequences serve as the primary sources of Ca^{2+} , Mg^{2+} , and HCO_3^- . These sequences include the FuJUNSHAN Formation (thick-bedded leopard limestone, dolomitic limestone, and dolomite) and the Majiagou Formation (limestone and dolomite). As groundwater flows from the southwestern recharge areas toward the northern basin margins, residence time increases, allowing more extensive carbonate dissolution. Simultaneously, dissolution of evaporite minerals (gypsum and halite) progressively enriches groundwater with Na^+ , Cl^- , and SO_4^{2-} (S. Liu et al., 2023). This drives the evolution from $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ type toward more mineralized $\text{Cl}\cdot\text{SO}_4\text{-Ca}\cdot\text{Mg}$ type waters in downgradient areas (R. Wang et al., 2022). This evolution sequence is consistent with regional patterns observed in other North China Plain studies. In contrast, K^+ exhibited a weaker gradient and contributed less to the regional TDS variation. Its relatively stable distribution suggests that K^+ was mainly influenced by localized mineral sources and adsorption processes rather than the dominant groundwater evolution pathway.

In contrast, trace components (F^- , Mn, As) exhibited weak or localized gradients, indicating stronger controls from local geochemical conditions rather than regional mineralization. F^- enrichment in localized areas may be associated with fluorite dissolution under Na-HCO_3 type groundwater conditions, where elevated pH and carbonate alkalinity enhance fluoride mobilization (S. Wang et al., 2025). Mn showed scattered hotspots, reflecting heterogeneous redox conditions; reducing microenvironments can promote Mn oxide reduction and Mn release (X. Wang et al., 2022). As enrichment occurred in restricted zones, likely related to reducing and high-pH conditions that enhance As mobilization through the reduction of As(V) to more mobile As(III) (Li et al., 2024). These localized distributions highlight the importance of considering element-specific geochemical controls when establishing groundwater NBLs.

The decoupling between major ions and trace components justifies separate statistical treatment for background values of different parameter groups. Notably, the elevated SO_4^{2-} levels observed in Units III and VI (reflected in the upper-right region

of the SO_4^{2-} component plane) may be linked to oxidation of sulfide minerals in local coal-bearing sequences. The Kailuan coalfield contains Permian coal seams with significant pyrite content and lies within the study area. Oxidation of pyrite (FeS_2) releases SO_4^{2-} and H^+ (Shang et al., 2024). The acidic byproducts of pyrite oxidation would enhance carbonate dissolution, further contributing to the elevated TDS and major ion concentrations in these units. Additionally, infiltration of groundwater from coal-bearing strata may also contribute to the elevated SO_4^{2-} . This hypothesis is supported by the transitional mixed $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ and $\text{Cl}\cdot\text{SO}_4\text{-Ca}\cdot\text{Mg}$ water types observed on Piper plots for the upper sections of Units I, III, and IV. It also aligns with regional observations that coal-bearing strata significantly influence groundwater chemistry in the North China Plain.

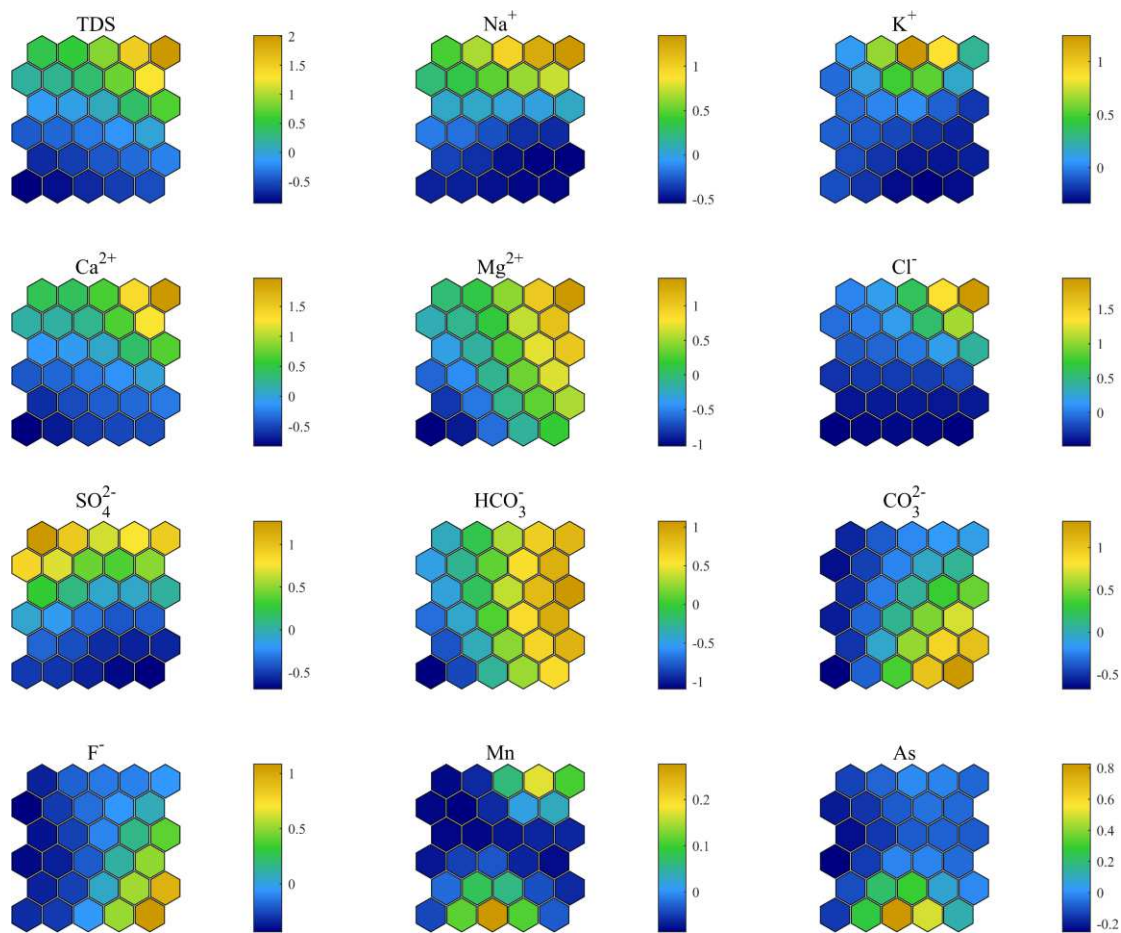


Fig. 5. Self-organizing maps of variables in shallow groundwater

4.3.2 Relationship between SOM clustering and original hydrogeological units

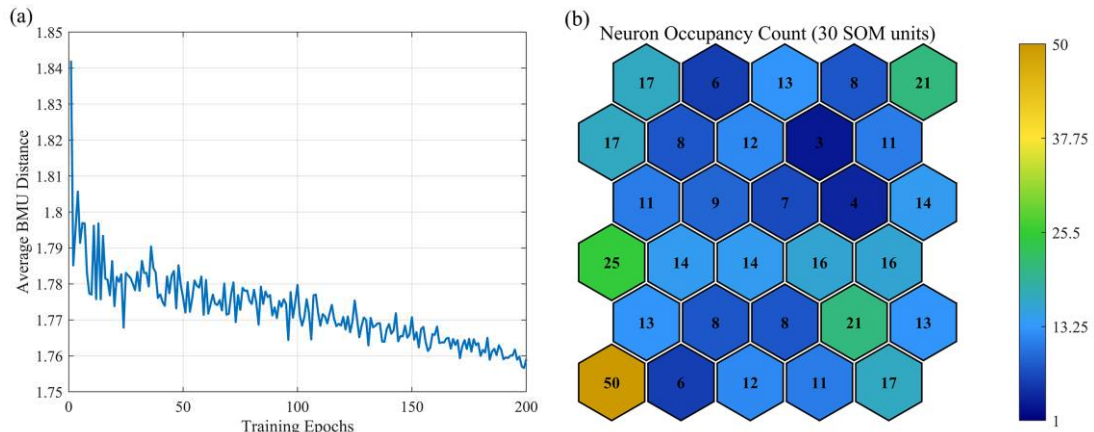


Fig. 6. SOM training and grid visualization. (a) Training convergence curve. (b) Sample counts per neuron in the 5×6 hexagonal SOM grid.

The training convergence curve (Fig. 6(a)) shows that the average BMU distance decreased rapidly and then stabilized after 150 epochs, confirming that the SOM effectively captured the structure of the 12-dimensional hydrochemical data without overfitting (Rahman et al., 2022). The trained SOM grid (Fig. 6(b)) exhibits highly uneven neuron occupancy: a dense high-count region in the lower-left corner (up to 50 samples) and sparse low-count neurons in the upper-right and central areas (3-6 samples). This non-uniform distribution indicates that the SOM has compressed the chemical space into a meaningful topology, concentrating similar samples into specific neurons (Y. Zhang et al., 2023).

Projecting the original six statistical unit labels onto the SOM grid reveals a clear spatial segregation. Units I-III (recharge zones, $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ type) map predominantly to the low-TDS, high-density lower-left region, whereas Units IV-VI (transitional and basin-margin areas) concentrate in the high-TDS upper-right and right-central regions. The upper halves of Units I, III, and IV, which contain $\text{Cl}\cdot\text{SO}_4\text{-Ca}\cdot\text{Mg}$ water in the Piper diagrams, partially overlap the high-TDS zone, confirming that SOM captures internal heterogeneity within these units (Shang et al., 2024). The overall consistency between the SOM-driven clustering and the predefined statistical units demonstrates that the hydrogeological zoning is largely supported by the natural chemical clustering of groundwater (Fan et al., 2024), with only minor overlaps that suggest boundaries could be refined with higher-resolution data (Mia et al., 2023). Notably, the SOM-derived clusters show that these overlaps are not randomly distributed but systematically occur

541 along the TDS gradient axis (from lower-left to upper-right of the SOM grid),
542 suggesting that the transitions between adjacent hydrogeological units follow a
543 progressive chemical evolution pathway rather than sharp discontinuities.

544 4.4 KNN-based outlier identification and classification validation

545 The KNN algorithm was applied for two complementary purposes: (1) to detect
546 potential outliers within each statistical unit using the optimal neighbor number
547 determined by elbow curves, and (2) to evaluate the overall separability of the six
548 hydrogeological units via a supervised classification framework (R. Liu et al., 2026).
549 The analysis used the same cleaned dataset ($n = 405$) and standardized 12
550 hydrochemical parameters as in the SOM clustering. All computations were performed
551 in MATLAB.

552 4.4.1 Unit-specific KNN anomaly detection and optimal K selection

553 For each statistical unit, the optimal number of neighbors K for KNN-based
554 anomaly detection was determined from the elbow curve of distance variance (Fig. 7).
555 The elbow point, where the variance growth changes from steep to gentle, balances
556 sensitivity and robustness (Jones et al., 2020). The optimal K values varied considerably
557 across units: Unit I required $K = 14$, Unit II: $K = 9$, Units III-V: $K = 6$, and Unit VI: K
558 $= 2$. This variation is not arbitrary but reflects fundamental differences in the internal
559 hydrochemical structure of each statistical unit.

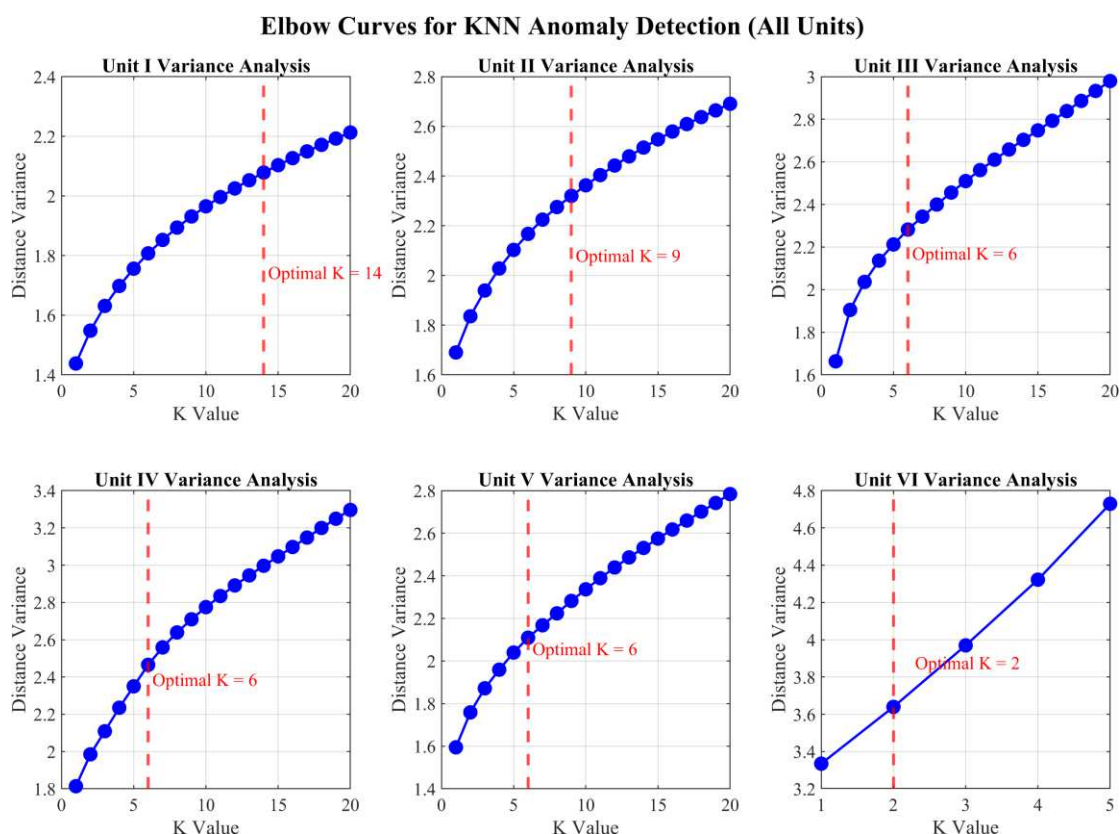


Fig. 7. Optimal k selection for KNN anomaly detection using distance variance elbow curves across Units I–VI

The red dashed line marks the 80% cumulative variance threshold used to identify the elbow point for each unit. The selected K values are annotated for each unit.

Unit I, with the largest sample size ($n = 199$) and the widest background interval (P5-P95 TDS: 219 - 915 mg/L), required the largest K value ($K = 14$). This indicates that Unit 1 contains the most diverse hydrochemical compositions, encompassing the full range of natural variability from low-TDS recharge waters to more evolved transitional waters. A larger K is necessary to capture this internal heterogeneity and to ensure that the neighborhood-based distance calculation is not overly influenced by local fluctuations. In contrast, Unit VI ($n = 6$) required the smallest K ($K = 2$), consistent with its highly compact chemical cluster, as evidenced by its uniform HCO_3 -dominated high-TDS composition in the Piper diagram and its exclusive projection onto the high-TDS region of the SOM grid (Tian et al., 2025). Units III-V, with intermediate sample sizes and transitional hydrochemical signatures, required moderate K values ($K = 6$), reflecting their intermediate positions in the chemical evolution sequence. The systematic decrease in optimal K from Unit I ($K = 14$) through Units III-V ($K = 6$) to Unit VI ($K = 2$) thus mirrors the progressive chemical homogenization from the most

heterogeneous recharge unit to the most uniform discharge endmember (Basharat et al., 2025). This pattern validates the conceptual model of hydrochemical evolution along the regional flow path.

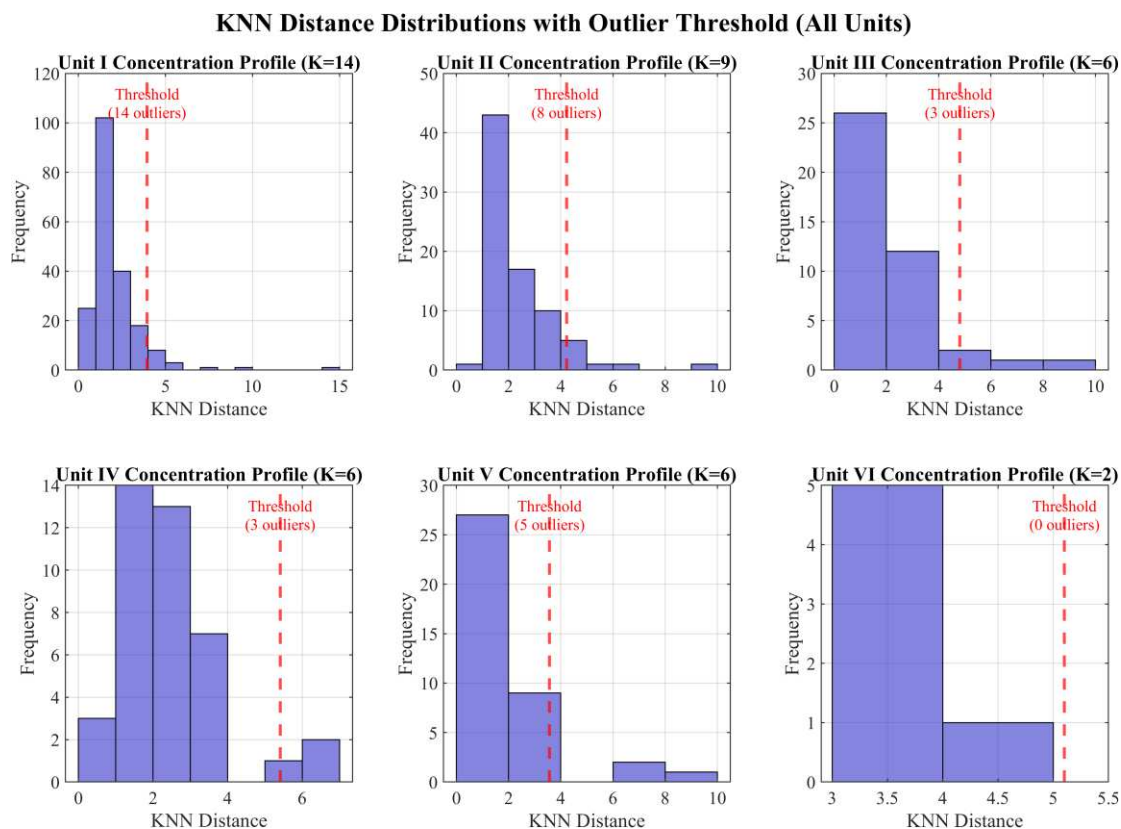


Fig. 8. Frequency histograms of KNN distances and outlier counts for each hydrochemical unit. The red dashed line indicates the anomaly threshold (mean +3 σ of the distance distribution); samples to the right of this line are flagged as potential outliers. The number of detected anomalies is shown in each panel.

Using the respective optimal K values, the average KNN distance was calculated for each sample. The concentration profiles (Fig. 8) show unimodal right-skewed distributions in all units, where the majority of samples cluster at low distances (representing the normal background population) and a small number of tail points exceed the threshold (Ma et al., 2021). The numbers of detected anomalies were: Unit I: 14, Unit II: 8, Units III-V: 3 - 5, and Unit VI: 0. Notably, Unit 6 showed no anomalies, reinforcing its chemical homogeneity and consistent with its very small K value. The higher anomaly counts in Unit I (14 samples) are consistent with its larger sample size and greater internal diversity. This inevitably increases the probability of encountering statistical outliers (Dhapre et al., 2025). Some of these outliers may represent genuine natural extremes rather than measurement errors or anthropogenic contamination.

4.4.2 Supervised KNN classification for unit separability

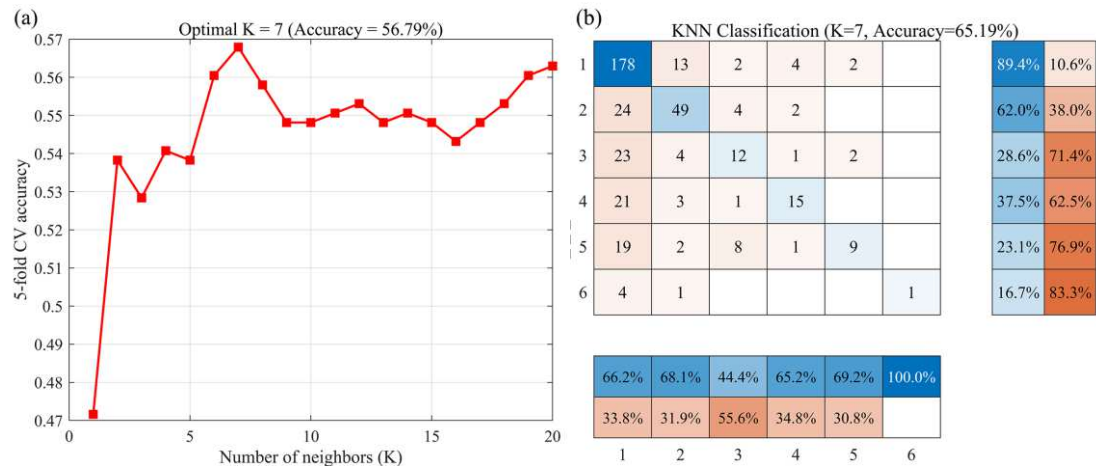


Fig. 9. KNN classification results. (a) 5-fold cross-validation accuracy as a function of K, with the optimal K=7 marked by a red circle. (b) Confusion matrix for the KNN classifier at K=7; rows represent true unit labels; columns represent predicted labels. Diagonal cells show correctly classified samples; off-diagonal cells indicate misclassifications. Color intensity reflects the number of samples.

The KNN classification was performed on the 405 samples that passed the initial Grubbs/ 3σ screening, without further exclusion based on the KNN anomaly detection, to provide an unbiased estimate of the natural separability of the six statistical units. A 5-fold cross-validation was performed to select the optimal K for classification. The cross-validation accuracy curve (Fig.9(a)) shows that the accuracy increased rapidly from K = 1 to K = 7, reaching a peak of 54.8%, and then fluctuated downward. Thus K = 7 was chosen as the optimal value for classification, balancing under- and over-fitting. The confusion matrix for K = 7 (Fig. 9(b)) yields an overall classification accuracy of 65.2 %. While this accuracy is moderate and suggests that the six units are partially distinguishable, the complementary 34.8 % misclassification rate is equally informative, it directly quantifies the degree of hydrochemical overlap between adjacent units (Adem et al., 2026). This overlap is not a methodological failure but rather a reflection of the intrinsic geological reality that groundwater chemistry evolves gradually along flow paths, and the boundaries between hydrogeological units are rarely sharp (Jalali et al., 2024). The misclassification patterns reveal important insights into the nature of these transitions.

Unit I shows the highest classification accuracy (89.4%, 178 correct samples),

indicating that its $\text{HCO}_3\text{-Ca}\cdot\text{Mg}$ type water is chemically distinct from the other units. This is consistent with its role as the most typical recharge-zone endmember with minimal anthropogenic disturbance. In contrast, Unit II is frequently misclassified as Unit I (24 samples), suggesting that the hydrochemical signatures of these two units substantially overlap. This overlap is geochemically plausible, as both units are dominated by carbonate dissolution processes. However, their distinguishing characteristics (specifically Unit II's slightly lower Cl^- and SO_4^{2-} background levels) may not be sufficiently pronounced for a distance-based classifier to separate them reliably. Unit III also exhibits moderate misclassification, especially with Units IV and VI, confirming the transitional nature of its hydrochemical signature as previously identified in the SOM component planes and Piper diagrams. Units IV and V show moderate accuracies (around 60 - 70 %), reflecting their intermediate positions in the chemical evolution sequence. Unit VI (only 6 samples) shows 100% correct classification, but this result is statistically unreliable due to the extremely small sample size and should be interpreted with caution.

4.5 Complementarity of the SOM and KNN analyses

The unsupervised SOM and supervised KNN analyses were designed to address distinct but complementary questions regarding the six-unit hydrogeological zoning. Their integration within a single analytical framework is not merely additive but synergistic, as each method compensates for the inherent limitations of the other while reinforcing their respective strengths (Gholami & Sahour, 2024).

The complementarity is clearly illustrated by the specific findings. The SOM component planes revealed that the overlaps between adjacent units are not random but systematically occur along the TDS gradient axis, from the lower-left to the upper-right of the grid. This suggests that the chemical evolution from recharge zones to basin margins is a continuous process, and the boundaries between units are gradational rather than abrupt. The KNN classification then quantified this gradation: the 34.8% misclassification rate directly measures the extent of overlap, with Units I-II and Units III-IV showing the highest confusion (Amiri & Nakagawa, 2021). Thus, SOM provides the qualitative spatial context (where and along which gradient overlaps

occur), while KNN provides the quantitative measure (how much overlap exists). Together, they transform a qualitative observation into a statistically grounded assessment.

Another layer of synergy lies in the outlier detection. The unit-specific KNN anomaly detection, with its non-parametric distance-based logic, served as an independent verification of the parametric Grubbs/ 3σ removal. The $> 90\%$ agreement between the two methods confirms that the data cleaning was not arbitrary, while the $< 10\%$ discrepancies (samples flagged by only one method) highlight borderline cases that deserve individual scrutiny. This cross-validation is rarely achievable with a single method and significantly enhances the reliability of the cleaned dataset (Shang et al., 2024).

The two methods also connect directly to the research gaps identified in the introduction. The SOM addresses the first gap, which concerns whether predefined statistical units are hydrochemically meaningful, by demonstrating that unit labels projected onto the SOM grid form coherent, spatially distinct clusters. The KNN addresses the second gap, which concerns whether inter-unit overlaps can be quantified, by providing the confusion matrix and the 65.2% accuracy figure. The third gap, which concerns the particular severity of these issues in heterogeneous mountainous terrains, is implicitly addressed by the fact that both methods consistently reveal gradational boundaries rather than sharp discontinuities, a finding that would be less pronounced in hydrogeologically simpler settings.

5 Conclusion

This study developed a hybrid SOM-KNN framework to validate groundwater NBL zonation in a hydrogeologically heterogeneous mountainous area of Tangshan, North China. By integrating unsupervised hydrochemical pattern recognition with supervised quantitative evaluation, this framework provides a process-informed approach for assessing whether predefined statistical units reflect meaningful groundwater organization. Based on six predefined hydrogeological units, 480 groundwater samples were collected, and 405 samples with 12 hydrochemical

parameters were retained after rigorous outlier removal.

The SOM analysis revealed a dominant hydrochemical evolution gradient associated with groundwater recharge-flow-discharge processes. TDS increased from 219-915 mg/L in recharge-dominated units to 320-1357 mg/L in basin-margin units, indicating progressive groundwater mineralization along the flow path. Major ions exhibited consistent spatial patterns with this evolution trend, whereas trace elements showed localized variations, suggesting stronger controls from lithological heterogeneity and local geochemical processes. Projection of predefined units onto the SOM topology demonstrated that most statistical units corresponded well with intrinsic hydrochemical structures, while overlaps mainly occurred at unit boundaries, reflecting transitional groundwater regimes rather than clearly separated hydrochemical domains.

The KNN analysis provided complementary quantitative validation of the proposed zonation framework. Unit-specific anomaly detection showed >90% agreement with conventional Grubbs/ 3σ methods, supporting the reliability of the identified anomalous samples. Supervised KNN classification achieved an overall accuracy of 65.2%. The moderate classification performance does not indicate a failure of the zonation scheme but instead reveals the continuous nature of groundwater chemical evolution in heterogeneous mountainous aquifers, where rigid boundaries may not fully represent gradual transitions between groundwater regimes.

Overall, the SOM–KNN framework advances conventional NBL assessment by integrating hydrochemical organization, statistical validation, and hydrogeological interpretation. The results highlight the importance of considering transitional or mixed groundwater units when refining background-level delineation and improving groundwater quality management. However, the limited sample size of Unit VI ($n = 6$) may reduce the reliability of local NBL estimates and classification results. Future studies should incorporate additional sampling, higher-resolution spatial information, and complementary tracers (e.g., stable isotopes) to further refine groundwater zonation and extend the framework to other groundwater quality indicators in complex aquifer systems.

CRediT authorship contribution statement

Zhiqiang Gong: Supervision, Resources, Project administration, Writing - review & editing, Funding acquisition. **Xun Zhang:** Writing - original draft, Validation, Visualization, Methodology, Investigation. **Zonglong Zheng:** Conceptualization, Project administration. **Zhiwen Dang:** Methodology, Visualization, Software. **Zhisen Ma:** Visualization, Formal analysis.

Funding

This work was financially supported by grants from the “Hebei Key Laboratory of Geological Resources and Environment Monitoring and Protection open project (JCKYT20230303)”.

Acknowledgments

This work was financially supported by grants from the “Hebei Key Laboratory of Geological Resources and Environment Monitoring and Protection open project (JCKYT20230303)”. This funding was instrumental in supporting the sampling and data collection efforts central to this research. Sincere thanks are extended to the field technicians and laboratory analysts whose meticulous work ensured the high quality of the samples and the subsequent data. Their expertise and dedication were crucial to the success of this study.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

- Adem, K., Yilmaz, E. K., Alaboz, P., Dengiz, O., & Saygin, F. (2026). Comparative Analysis of Regression and Classification-Based Deep Learning and Machine Learning Models for Soil Erodibility Prediction. *Land Degradation & Development*, 37(11), 6361-6376. <https://doi.org/10.1002/ldr.70518>
- Amiri, V., & Nakagawa, K. (2021). Using a linear discriminant analysis (LDA)-based nomenclature system and self-organizing maps (SOM) for spatiotemporal assessment of groundwater quality in a coastal aquifer. *Journal of Hydrology*, 603, 127082. <https://doi.org/10.1016/j.jhydrol.2021.127082>
- Amiri, V., Nakhaei, M., Lak, R., & Li, P. (2021). An integrated statistical-graphical approach for the appraisal of the natural background levels of some major ions and potentially toxic elements in the groundwater of Urmia aquifer, Iran. *Environmental Earth Sciences*, 80(12), 432. <https://doi.org/10.1007/s12665-021-09733-0>
- Ayed, B., Jmal, I., Sahal, S., Mokadem, N., Saidi, S., Boughariou, E., & Bouri, S. (2017). Hydrochemical characterization of groundwater using multivariate statistical analysis: the Maritime Djeffara shallow aquifer (Southeastern Tunisia). *Environmental Earth Sciences*, 76(24), 821. <https://doi.org/10.1007/s12665-017-7168-6>
- Basharat, U., Zhang, W., Han, C., Khan, S. H., Abbasi, A., Mahroof, S., & Li, S. (2025). Optimizing machine learning methods for groundwater quality prediction: Case study in District Bagh, Azad Kashmir, Pakistan. *Ecotoxicology and Environmental Safety*, 302, 118610. <https://doi.org/10.1016/j.ecoenv.2025.118610>
- Bayat, N., Nassery, H. R., & Sadeghi, E. (2025). Groundwater hydrochemistry interpretation based on self-organizing maps and hierarchical clustering in Qorveh-Dehgolan plain, Western Iran. *Environmental Geochemistry and Health*, 48(1), 33. <https://doi.org/10.1007/s10653-025-02920-z>

- Bi, P., Liu, R., Huang, G., & Li, D. (2023). Evaluating natural background levels of heavy metals in shallow groundwater of the Pearl River Delta via removal of contaminated groundwaters: Comparison of three preselection related methods. *Environmental Pollution*, 335, 122382. <https://doi.org/10.1016/j.envpol.2023.122382>
- Bondu, R., Humez, P., Mayer, B., Chaste, E., Naumenko-Dèzes, M. O., Cloutier, V., Rosa, E., & Kloppmann, W. (2022). Estimating natural background concentrations for dissolved constituents in groundwater: A methodological review and case studies for geogenic fluoride. *Journal of Geochemical Exploration*, 233, 106906. <https://doi.org/10.1016/j.gexplo.2021.106906>
- Busico, G., Bordbar, M., Rufino, F., Sarracino, A., & Tedesco, D. (2024). Assessment of NO₃⁻, As, and F⁻ background levels in groundwater bodies: A methodological review and case study utilizing sequential Gaussian simulation (SGS). *Groundwater for Sustainable Development*, 26, 101211. <https://doi.org/10.1016/j.gsd.2024.101211>
- Cai, X., Chen, L., Hou, X., Zhang, M., Hu, Y., Ou, Q., Song, J., Wu, M., Shi, X., & Yin, X. (2025). The Mechanisms Controlling the Characteristics of Groundwater in the Qinan Coal Mine. *Mine Water and the Environment*, 44(1), 78-92. <https://doi.org/10.1007/s10230-025-01022-7>
- Chen, Y., Zhao, Q., & Lu, L. (2021). Combining the outputs of various k-nearest neighbor anomaly detectors to form a robust ensemble model for high-dimensional geochemical anomaly detection. *Journal of Geochemical Exploration*, 231, 106875. <https://doi.org/10.1016/j.gexplo.2021.106875>
- Costanzo, V. L., Izarra-Jacome, A., Costanzo-Alvarez, V., Jacome, M., Trigo-Ferre, R., Jeffrey, M., Cheng, Y., Aldana, M., & Amon, C. H. (2026). Machine-Learning–Derived Hydrogeochemical Facies to Support Geoelectrical Interpretation of Contaminant Plumes. *Environmental Processes*, 13(1), 3. <https://doi.org/10.1007/s40710-025-00811-0>
- D’Urso, P., De Giovanni, L., & Massari, R. (2020). Smoothed self-organizing map for robust clustering. *Information Sciences*, 512, 381-401.

<https://doi.org/10.1016/j.ins.2019.06.038>

Dang, Z., Zhang, X., Zhou, Y., Zheng, Z., Wang, K., Shi, S., Xin, Y., Gong, Z., & Mi, Y. (2025). Controls of Sedimentary Architecture on Groundwater Hydrochemistry and Human Health Risks in the Piedmont Alluvial Plain, Eastern Hebei. *Water Environment Research*, 97(12), e70225. <https://doi.org/10.1002/wer.70225>

Dang, Z., Zheng, Z., Tian, Y., Gao, P., Zhang, X., Gong, Z., Shi, S., & Ban, A. (2026). Unraveling the hydrochemical characteristics and assessing health risk of groundwater in the Luanhe River Piedmont Alluvial Fan Complex, an industrial-agricultural transition zone in North China. *Environmental Geochemistry and Health*, 48(2), 78. <https://doi.org/10.1007/s10653-025-02972-1>

Dhapre, M., Jadhav, S., Das, D., Khan, J., Kim, Y., Chiao, S., & Danielson, T. (2025). A systematic review of machine learning in groundwater monitoring. *Environmental Modelling & Software*, 192, 106549. <https://doi.org/10.1016/j.envsoft.2025.106549>

Du, A., Liao, F., Zan, Y., Zhang, Z., Yang, B., You, X., Zeng, Q., & Wang, G. (2026). Impact of isolated mountains on groundwater hydrochemical evolution: insight from isotopic and geochemical analysis. *Journal of Hydrology*, 674, 135494. <https://doi.org/10.1016/j.jhydrol.2026.135494>

Elumalai, V., Brindha, K., & Lakshmanan, E. (2017). Human Exposure Risk Assessment Due to Heavy Metals in Groundwater by Pollution Index and Multivariate Statistical Methods: A Case Study from South Africa. *Water*, 9(4), 234.

Fan, W., Zhou, J., Zheng, J., Guo, Y., Hu, L., & Shan, R. (2024). Hydrochemical characteristics, control factors and health risk assessment of groundwater in typical arid region Hotan Area, Chinese Xinjiang. *Environmental Pollution*, 363, 125301. <https://doi.org/10.1016/j.envpol.2024.125301>

Filipović, M., Terzić, J., Lukač Reberski, J., & Vlahović, I. (2024). Utilizing a multi-tracer method to investigate sulphate contamination: Novel insights on hydrogeochemical characteristics of groundwater in intricate karst systems. *Groundwater for Sustainable Development*, 27, 101350.

821 <https://doi.org/10.1016/j.gsd.2024.101350>

822 General Administration of Quality Supervision, I. a. Q. o. t. P. R. C., & Standardization
823 Administration of the P.R.C. (2017). *Standard for Groundwater Quality (GB/T*
824 *14848* -2017).

825 [https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=F745E3023BD5B10B9F](https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=F745E3023BD5B10B9FB5314E0FFB5523)
826 [B5314E0FFB5523](https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=F745E3023BD5B10B9FB5314E0FFB5523)

827 Gholami, V., & Sahour, H. (2024). A Hybrid Approach of Supervised Self-organizing
828 Maps and Genetic Algorithms for Predictive Mapping of Arsenic Pollution in
829 Groundwater Resources. *Exposure and Health*, 16(3), 775-790.
830 <https://doi.org/10.1007/s12403-023-00592-7>

831 Haggerty, R., Sun, J., Yu, H., & Li, Y. (2023). Application of machine learning in
832 groundwater quality modeling - A comprehensive review. *Water Research*, 233,
833 119745. <https://doi.org/10.1016/j.watres.2023.119745>

834 Hartmann, A., Goldscheider, N., Wagener, T., Lange, J., & Weiler, M. (2014). Karst
835 water resources in a changing world: Review of hydrological modeling approaches.
836 *Reviews of Geophysics*, 52(3), 218-242. <https://doi.org/10.1002/2013RG000443>

837 He, B., He, J., Zeng, Y., Sun, J., Peng, C., & Bi, E. (2022). Coupling of multi-
838 hydrochemical and statistical methods for identifying apparent background levels
839 of major components and anthropogenic anomalous activities in shallow
840 groundwater of the Liujiang Basin, China. *Science of The Total Environment*, 838,
841 155905. <https://doi.org/10.1016/j.scitotenv.2022.155905>

842 Huang, Y., Wang, C., Wang, Y., Lyu, G., Lin, S., Liu, W., Niu, H., & Hu, Q. (2023).
843 Application of machine learning models in groundwater quality assessment and
844 prediction: progress and challenges. *Frontiers of Environmental Science &*
845 *Engineering*, 18(3), 29. <https://doi.org/10.1007/s11783-024-1789-2>

846 Jalali, R., Tishehzan, P., & Hashemi, H. (2024). A machine learning framework for
847 spatio-temporal vulnerability mapping of groundwaters to nitrate in a data scarce
848 region in Lenjanat Plain, Iran. *Environmental Science and Pollution Research*,
849 31(29), 42088-42110. <https://doi.org/10.1007/s11356-024-33920-8>

850 Jones, P. J., James, M. K., Davies, M. J., Khunti, K., Catt, M., Yates, T., Rowlands, A.

- V., & Mirkes, E. M. (2020). FilterK: A new outlier detection method for k-means clustering of physical activity. *Journal of Biomedical Informatics*, 104, 103397. <https://doi.org/10.1016/j.jbi.2020.103397>
- Joseph, L., & Bhaumik, B. K. (1997). Improved Estimation of the Box–Cox Transform Parameter and Its Application to Hydrogeochemical Data. *Mathematical Geology*, 29(8), 963-976. <https://doi.org/10.1023/A:1022353401858>
- Khadra, W. M., Elias, A. R., & Majdalani, M. A. (2022). A systematic approach to derive natural background levels in groundwater: Application to an aquifer in North Lebanon perturbed by various pollution sources. *Science of The Total Environment*, 847, 157586. <https://doi.org/10.1016/j.scitotenv.2022.157586>
- Lazarev, I. D., Narozniak, M., Byrnes, T., & Pyrkov, A. N. (2025). Hybrid quantum-classical unsupervised data clustering based on the self-organizing feature map. *Physical Review A*, 111(1), 012416. <https://doi.org/10.1103/PhysRevA.111.012416>
- Li, J., Shi, Z., Liu, M., Wang, G., Liu, F., & Wang, Y. (2021). Identifying anthropogenic sources of groundwater contamination by natural background levels and stable isotope application in Pinggu basin, China. *Journal of Hydrology*, 596, 126092. <https://doi.org/10.1016/j.jhydrol.2021.126092>
- Li, J., Xie, X., & Wang, Y. (2024). Effects of the groundwater flowing and redox conditions on arsenic mobilization in aquifers of the Datong Basin, Northern China. *Journal of Hydrology*, 636, 131256. <https://doi.org/10.1016/j.jhydrol.2024.131256>
- Li, N., He, J., He, B., Chu, Y., & Chen, Z. (2025). Comparative study of hydrochemical and machine learning methods for Fe anomaly identification and NBLs estimation in shallow groundwater of representative hydrogeological units in the North China plain. *Groundwater for Sustainable Development*, 31, 101505. <https://doi.org/10.1016/j.gsd.2025.101505>
- Liu, N., Chen, M., Gao, D., Wu, Y., & Wang, X. (2025). Identification of hydrogeochemical processes in shallow groundwater using multivariate statistical analysis and inverse geochemical modeling. *Environmental Monitoring and*

Assessment, 197(2), 135. <https://doi.org/10.1007/s10661-024-13528-8>

Liu, Q., Liang, K., Xia, F., Yun, Z., Deng, S., Han, X., Yang, Y., & Jiang, Y. (2026). A Review of the Application of Machine Learning Models in Groundwater Resources Management and Quality Assessment. *Sustainability*, 18(11), 5261.

Liu, R., Li, Z., Xie, X., Hou, Q., Han, D., Yuan, Z., & Huang, G. (2026). A multi-method for identifying and spatial modeling of groundwater nitrate natural background levels in an urbanized delta in China: Insight from Cl/Br ratios combining with KNN algorithm and EBK. *Water Research*, 289, 124953. <https://doi.org/10.1016/j.watres.2025.124953>

Liu, S., Li, H., Yang, J., Ma, M., Shang, J., Tang, Z., & Liu, G. (2023). Using Self-Organizing Map and Multivariate Statistical Methods for Groundwater Quality Assessment in the Urban Area of Linyi City, China. *Water*, 15(19), 3463.

Liu, Z., Feng, S., Zhangsong, A., Han, Y., & Cao, R. (2023). Long-term evolution of groundwater hydrochemistry and its influencing factors based on self-organizing map (SOM). *Ecological Indicators*, 154, 110697. <https://doi.org/10.1016/j.ecolind.2023.110697>

Ma, F., Wei, A., Deng, Q., & Zhao, H. (2014). Hydrochemical characteristics and the suitability of groundwater in the coastal region of Tangshan, China. *Journal of Earth Science*, 25(6), 1067-1075. <https://doi.org/10.1007/s12583-014-0492-9>

Ma, Y., Zhao, X., Zhang, C., Zhang, J., & Qin, X. (2021). Outlier detection from multiple data sources. *Information Sciences*, 580, 819-837. <https://doi.org/10.1016/j.ins.2021.09.053>

Masciale, R., Amalfitano, S., Frollini, E., Ghergo, S., Melita, M., Parrone, D., Preziosi, E., Vurro, M., Zoppini, A., & Passarella, G. (2021). Assessing Natural Background Levels in the Groundwater Bodies of the Apulia Region (Southern Italy). *Water*, 13(7), 958.

Mia, M. Y., Haque, M. E., Islam, A. R. M. T., Jannat, J. N., Jion, M. M. M. F., Islam, M. S., Siddique, M. A. B., Idris, A. M., Senapathi, V., Talukdar, S., & Rahman, A. (2023). Analysis of self-organizing maps and explainable artificial intelligence to identify hydrochemical factors that drive drinking water quality in Haor region.

Science of The Total Environment, 904, 166927.
<https://doi.org/10.1016/j.scitotenv.2023.166927>

Ministry of Ecology and Environment, P. (2023). *Guidelines for statistical characterization of groundwater environmental background values (Trial version)*. Ministry of Ecology and Environment of the People's Republic of China.
https://www.mee.gov.cn/xxgk/xxgk/xxgk06/202310/t20231027_1044123.html

Parrone, D., Frollini, E., Masciale, R., Melita, M., Passarella, G., Preziosi, E., & Ghergo, S. (2022). A double pre-selection method for natural background levels assessment in coastal groundwater bodies. *Environmental Pollution*, 313, 120076.
<https://doi.org/10.1016/j.envpol.2022.120076>

Parrone, D., Ghergo, S., & Preziosi, E. (2019). A multi-method approach for the assessment of natural background levels in groundwater. *Science of The Total Environment*, 659, 884-894. <https://doi.org/10.1016/j.scitotenv.2018.12.350>

Preziosi, E., Rotiroti, M., Condesso de Melo, M. T., & Hinsby, K. (2021). Natural Background Levels in Groundwater. *Water*, 13(19), 2770.

Rahman, A. T. M. S., Kono, Y., & Hosono, T. (2022). Self-organizing map improves understanding on the hydrochemical processes in aquifer systems. *Science of The Total Environment*, 846, 157281. <https://doi.org/10.1016/j.scitotenv.2022.157281>

Ramos Marinho, A. P., Nunes, G. S., Benetti, M., & de Noronha, M. (2020). Cross-cultural adaptation and measurement properties of the Brazilian–Portuguese version of the Cincinnati Knee Rating System. *Disability and Rehabilitation*, 42(8), 1183-1189. <https://doi.org/10.1080/09638288.2018.1516818>

Rhyu, J., Bozinovski, D., Dubs, A. B., Mohan, N., Cummings Bende, E. M., Maloney, A. J., Nieves, M., Sangerman, J., Lu, A. E., Hong, M. S., Artamonova, A., Ou, R. W., Barone, P. W., Leung, J. C., Wolfrum, J. M., Sinskey, A. J., Springs, S. L., & Braatz, R. D. (2024). Automated outlier detection and estimation of missing data. *Computers & Chemical Engineering*, 180, 108448.
<https://doi.org/10.1016/j.compchemeng.2023.108448>

Sako, A., & Ouangaré, C. A. C. (2023). Hydrogeochemical characterization and natural

background level determination of selected inorganic substances in groundwater from a semi-confined aquifer in Midwestern Burkina Faso, West Africa. *Environmental Monitoring and Assessment*, 195(4), 519. <https://doi.org/10.1007/s10661-023-11127-7>

Shang, Y., Fu, C., Zhang, W., Li, X., & Li, X. (2024). Groundwater hydrochemistry, source identification and health assessment based on self-organizing map in an intensive mining area in Shanxi, China. *Environmental Research*, 252, 118934. <https://doi.org/10.1016/j.envres.2024.118934>

Tian, J., Yang, J., Liu, W., Zhang, M., Daskalopoulou, K., Zou, Y., Xu, N., Liao, Z., Huo, Y., Zhu, Y., Cao, Y., Xu, S., & Liu, J. (2025). Assessing groundwater quality for drinking and irrigation using hydrogeochemistry and machine learning in Northern China. *Agricultural Water Management*, 322, 109975. <https://doi.org/10.1016/j.agwat.2025.109975>

Urresti-Estala, B., Carrasco-Cantos, F., Vadillo-Pérez, I., & Jiménez-Gavilán, P. (2013). Determination of background levels on water quality of groundwater bodies: A methodological proposal applied to a Mediterranean River basin (Guadalhorce River, Málaga, southern Spain). *Journal of Environmental Management*, 117, 121-130. <https://doi.org/10.1016/j.jenvman.2012.11.042>

Vieira, L. D., Bertolo, R. A., Gouvêa Júnior, J. C. R., Hart, S. T., & Huguen, L. A. (2026). A Statistical Assessment of Outlier Removal Methods for Defining Groundwater Background Levels from Historical Data in Coal Mining Regions: A Southern Brazil Case Study. *Mine Water and the Environment*. <https://doi.org/10.1007/s10230-026-01114-y>

Voutchkova, D. D., Ernstsén, V., Schullehner, J., Hinsby, K., Thorling, L., & Hansen, B. (2021). Roadmap for Determining Natural Background Levels of Trace Metals in Groundwater. *Water*, 13(9), 1267.

Wang, C., Liu, R., Han, D., Hou, Q., & Li, Z. (2025). Hydrogeochemical characteristics and groundwater quality in typical alluvial-pluvial fans of the Hebei plain based on high-density sampling. *Environmental Earth Sciences*, 84(21), 646. <https://doi.org/10.1007/s12665-025-12649-8>

971 Wang, Q., Wang, M., Li, Y., Guo, B., Li, H., Liu, Y., Zhao, L., Ma, C., & Yuan, Z. (2025).
 972 Study on Natural Background Levels and Mechanisms of Groundwater
 973 Contamination in an Overexploited Aquifer Region: A Case Study of Xingtai City,
 974 North China Plain. *Water*, 17(19), 2836.

975 Wang, R., Li, X., & Wei, A. (2022). Hydrogeochemical characteristics and gradual
 976 changes of groundwater in the Baiquan karst spring region, northern China.
 977 *Carbonates and Evaporites*, 37(3), 47. [https://doi.org/10.1007/s13146-022-](https://doi.org/10.1007/s13146-022-00794-1)
 978 00794-1

979 Wang, S., Chen, J., Chen, D., Liu, F., Zhang, S., Bai, Y., Xiao, C., & Zhang, X. (2025).
 980 Groundwater recharge sources and hydrochemical evolution based on self-
 981 organizing maps (SOM) in the southeast of Nansi Lake Basin, North China.
 982 *Journal of Geochemical Exploration*, 277, 107815.
 983 <https://doi.org/10.1016/j.gexplo.2025.107815>

984 Wang, X., Xie, G.-J., Tian, N., Dang, C.-C., Cai, C., Ding, J., Liu, B.-F., Xing, D.-F.,
 985 Ren, N.-Q., & Wang, Q. (2022). Anaerobic microbial manganese oxidation and
 986 reduction: A critical review. *Science of The Total Environment*, 822, 153513.
 987 <https://doi.org/10.1016/j.scitotenv.2022.153513>

988 Wang, Y., Zhang, Y., Shao, S., Gong, Z., Wang, Y., & Qu, Z. (2025). Analysis of the
 989 environmental background values and health risk assessment of shallow
 990 groundwater in the Jidong Plain. *Human and Ecological Risk Assessment: An*
 991 *International Journal*, 31(3-4), 348-366.
 992 <https://doi.org/10.1080/10807039.2025.2464907>

993 Wen, S., Liang, S., Pang, G., Shan, Q., Ye, Y., Zhang, J., Dong, M., Fu, L., & Wen, M.
 994 (2025). Hydrochemical Characteristics and Evolution Mechanisms of Shallow
 995 Groundwater in the Alluvial–Coastal Transition Zone of the Tangshan Plain, China.
 996 *Water*, 17(19), 2810.

997 Wu, C., Zhou, H., Lu, C., Zhao, Y., Liu, R., Zhan, L., & Shi, Z. (2025). Groundwater
 998 nitrate responses to extreme rainfall in alluvial–diluvial plain aquifers: Evidence
 999 from hydrogeochemistry and isotopes. *Journal of Contaminant Hydrology*, 273,
 1000 104584. <https://doi.org/10.1016/j.jconhyd.2025.104584>

- Xiao, S., Shen, B., Tang, Q., Kaufman, A. J., Yuan, X., Li, J., & Qian, M. (2014). Biostratigraphic and chemostratigraphic constraints on the age of early Neoproterozoic carbonate successions in North China. *Precambrian Research*, 246, 208-225. <https://doi.org/10.1016/j.precamres.2014.03.004>
- Zanotti, C., Caschetto, M., Bonomi, T., Parini, M., Cipriano, G., Fumagalli, L., & Rotiroti, M. (2022). Linking local natural background levels in groundwater to their generating hydrogeochemical processes in Quaternary alluvial aquifers. *Science of The Total Environment*, 805, 150259. <https://doi.org/10.1016/j.scitotenv.2021.150259>
- Zhang, J., He, L., Ma, T., Yuan, H., Liu, W., Sun, L., & Yan, Z. (2026). Hydrochemical evolution and formation mechanisms of groundwater in the Daihai Lake Basin plain under ongoing ecological water replenishment. *Scientific reports*, 16(1), 20303. <https://doi.org/10.1038/s41598-026-50698-6>
- Zhang, T., Wang, P., He, J., Liu, D., Wang, M., Wang, M., & Xia, S. (2023). Hydrochemical Characteristics, Water Quality, and Evolution of Groundwater in Northeast China. *Water*, 15(14), 2669.
- Zhang, X., Zhang, Y., Gui, Y., Sun, R., Li, J., Wu, Q., Ding, Y., & Chen, K. (2025). Chemical characteristics of groundwater and surface water affected by human activities in the upper Jinzi River Basin, China. *Scientific reports*, 15(1), 9294. <https://doi.org/10.1038/s41598-025-93318-5>
- Zhang, Y., Zhang, Q., Chen, W., Shi, W., Cui, Y., Chen, L., & Shao, J. (2023). Hydrogeochemical analysis and groundwater pollution source identification based on self-organizing map at a contaminated site. *Journal of Hydrology*, 616, 128839. <https://doi.org/10.1016/j.jhydrol.2022.128839>
- Zhong, C., Wang, H., & Yang, Q. (2022). Hydrochemical interpretation of groundwater in Yinchuan basin using self-organizing maps and hierarchical clustering. *Chemosphere*, 309, 136787. <https://doi.org/10.1016/j.chemosphere.2022.136787>
- Zhou, Y., Wu, J., Gao, X., Guo, W., & Chen, W. (2023). Hydrochemical Background Levels and Threshold Values of Phreatic Groundwater in the Greater Xi'an Region, China: Spatiotemporal Distribution, Influencing Factors and Implication to Water

1031 Quality Management. *Exposure and Health*, 15(4), 757-771.
1032 <https://doi.org/10.1007/s12403-022-00521-0>
1033