https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

Hydrochemistry and modeling nitrate concentration in farmland groundwater under different hydrological seasons by integrating hybrid quantum-classical ML, virtual sample generation and AlphaEarth Foundation

Junjie Xu [a,b,c,†], Xin Wei [d,†], Yilei Yu [a, e,*], Lihu Yang [b], Yuanzheng Zhai [f], Cuicui Lv [g], Xianfang Song [b]

a College of Life Sciences, Hebei University, Baoding, Hebei, 071000, China

b Key Laboratory of Land Water Cycle and Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

c University of Chinese Academy of Sciences, Beijing 100049, China

d College of Ecology and Environment, Institute of Disaster Prevention Science and Technology, Sanhe, Hebei, 065201, China

e Engineering Research Center of Groundwater Pollution Control and Remediation, Ministry of Education of China, Beijing Normal University, Beijing 100875, China

f College of Water Sciences, Beijing Normal University, 100875, Beijing, China

g Xiong'an Institute of Innovation, Xiong'an 071899, China

* *Correspondence: yileiyu@hbu.edu.cn*

† *These authors contributed equally to this work.*

Abstract

Precise seasonal prediction of groundwater nitrate concentrations in intensive agricultural areas faces challenges such as data sparsity, strong spatiotemporal heterogeneity, and complex hydro-biogeochemical processes. To address these issues, this study proposes an integrated prediction framework combining hybrid quantum-classical machine learning, advanced virtual sample generation (t-SNE-GMM-KNN), and remote sensing foundation model semantic embedding (AEF). Modeling was conducted across the 2022-2023 normal, dry, and wet seasons in Xiong'an New Area. Hydrochemical types were dominated by Ca-Mg-HCO_3^-, controlled by mineral dissolution and evaporation. Nitrate concentrations were highest in the dry season (mean

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

42.93 mg L^{-1}), driven by evaporative concentration. Spatially, high-value zones shifted: southeast (normal), central (dry), and northwest (wet). MixSIAR modeling based on isotopes indicated domestic sewage and livestock manure (74.1%) as dominant sources. The t-SNE-GMM-KNN strategy mitigated small-sample bias while preserving nonlinear structure. When virtual samples were augmented to 10-fold, the Random Forest R^2 in the dry season increased from 0.284 to >0.85. Furthermore, a hybrid quantum-classical Random Forest exhibited superior robustness for data sparsity, achieving peak performance in the normal season (R^2=0.962, RMSE=5.73 mg L^{-1}). Additionally, using only AEF embeddings achieved screening-level accuracy (R^2 up to 0.860), providing a feasible rapid survey scheme for extensive unmonitored regions. Correlation analysis identified TDS and EC as persistent top predictors (r>0.8). This comprehensive framework offers a robust solution for seasonal nitrate prediction and sustainable water management.

Keywords: Groundwater nitrate concentration; Hydrological seasons; Virtual sample generation; Hybrid quantum-classical machine learning; AlphaEarth Foundation (AEF) embeddings; Nitrate source apportionment.

1. Introduction

Nitrate (NO_3^-) contamination in groundwater poses a serious threat to drinking water safety and ecosystem health, particularly in intensively managed agricultural regions (Wang et al., 2021). In China, groundwater nitrate pollution is a growing concern, national monitoring data from 2013 to 2017 revealed a nitrate exceedance rate exceeding 10%, with Hebei Province reporting an alarming rate of 31.66% in 2017 (Li et al., 2019). Over recent decades, escalating nitrate concentrations in surface and groundwater have been driven by intensified fertilizer use in agriculture, along with discharges of industrial and domestic wastewater (Zhang et al., 2018). Severe nitrate exceedances are especially prevalent in northern and northwestern China (Gu et al., 2013), where key contributors include domestic and industrial effluents, nitrification of soil organic nitrogen, and synthetic fertilizer application (Han et al., 2016). For instance, in the North China Plain, shallow groundwater nitrate exceedance rates range from 9.5% to 34.1%, and a rising trend persists at the regional scale, particularly in agricultural areas (Wang et al., 2018). In monsoonal temperate regions, seasonal shifts in precipitation, evapotranspiration, and groundwater recharge profoundly influence the transport, dilution, and accumulation of nitrate,

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

leading to pronounced intra-annual variability in its concentration and spatial distribution (Gao et al., 2023; Zhu et al., 2025). Consequently, understanding and forecasting nitrate dynamics across hydrological seasons is essential for informed groundwater management and pollution mitigation, but remains a formidable challenge due to the nonlinearity, high dimensionality, and data scarcity inherent in such systems (Deng et al., 2023).

Traditional monitoring and modeling approaches face three critical limitations. First, field sampling campaigns though providing high-fidelity hydrochemical data are inherently sparse in space and time, especially for large-scale or rapidly changing environments (Viswanathan et al., 2022), which are time-consuming, labor-intensive, and costly, limiting the spatial and temporal coverage of data (Cai et al., 2025). Second, while process-based models incorporate physical mechanisms, they require extensive parameterization and are computationally prohibitive for dynamic, multi-season forecasting at farm-to-regional scales (Feng et al., 2022). Hydrological seasonal variations (normal, dry, and wet seasons) significantly influence the migration and transformation of nitrogen in the soil-groundwater system (Chen et al., 2025). For instance, concentrated rainfall during the wet season (accounting for 60%-80% of annual precipitation) can promote the leaching of surface nitrogen into groundwater, leading to a 25-fold increase in stream nitrate concentrations during storm events compared to baseflow (Sebestyen et al., 2014), meanwhile, intense evaporation in the dry season leads to the accumulation of nitrate in shallow aquifers, where concentrations can exceed the US EPA drinking water standard of 10 mg L^{-1} by 2-3 times (Liu et al., 2025; Cox et al., 2016). These seasonal differences result in distinct hydrochemical characteristics and nitrate concentration distributions, increasing the complexity of prediction models (Wu et al., 2025). Third, even advanced machine learning (ML) techniques such as Random Forest (RF), despite their robustness to nonlinearity and multicollinearity, still rely heavily on sufficient representative samples to capture the multi-modal distribution and tail behavior of environmental variables, particularly for heavy-tailed pollutants like NO_3^- (Luo et al., 2022). Moreover, the small sample sizes obtained from discrete sampling often lead to data sparsity and skewed distributions, reducing the model's generalization ability by 30%-50% when applied to unmonitored areas and compromising the robustness and generalization ability of machine learning (ML) models trained on such data (Thunyawatcharakul et al., 2025; Wang et al., 2024).

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

To overcome these bottlenecks, recent efforts have explored virtual sample augmentation and hybrid modeling. Gaussian Mixture Models (GMM) and deep generative frameworks (e.g., VAEs, GANs) have shown promise in enriching training data, with GMM achieving an average similarity of 83.0% between unmixed chemical spectra and ground truth in geochemical analysis (Farnia et al., 2023; Tung et al., 2023), however, they often fail to preserve the non-linear manifold structure of high-dimensional geochemical space or require large training sets, precisely what is lacking (Zhou et al., 2025). Non-linear dimensionality reduction methods, such as t-SNE, excel at revealing latent clusters corresponding to distinct hydrological processes, with a classification accuracy of 92% for annual daily hydrograph clustering in mountainous watersheds, yet lack explicit generative mechanisms (Wang et al., 2025; Tang et al., 2022). Meanwhile, the rise of foundation models in Earth observation exemplified by Google's AlphaEarth Foundation (AEF), offers unprecedented opportunities: its 64-dimensional semantic embeddings, derived from multi-sensor satellite time series (including Sentinel-2, Landsat, and Sentinel-1), implicitly encode land use, vegetation phenology, soil moisture, and anthropogenic footprints at 10 m resolution (Tollefson, 2025). These features have been successfully applied in land use classification and crop monitoring, but their potential for predicting groundwater nitrate concentrations, especially across different hydrological seasons remains underexplored (Li et al., 2025). Quantum machine learning (QML) further opens a new frontier. Parameterized Quantum Circuits (PQCs) can map classical inputs into exponentially high-dimensional quantum Hilbert spaces, generating entangled feature representations that reveal complex, non-linear patterns inaccessible to classical kernels (Hong et al., 2025). For ozone concentration forecasting, a hybrid QML model achieved an R^2 of 94.12% for 1-hour forecasts and 75.62% for 6-hour forecasts, outperforming classical persistence models by a forecast skill of 31.01-57.46% (Oliveira et al., 2025). Crucially, analytical quantum feature extraction via Pauli-Z expectation values avoids the noisy sampling overhead of near-term quantum hardware, reducing computational latency by ~80% compared to sampling-based methods and making it viable for small-sample environmental modeling (Gujju et al., 2024; Oliveira et al., 2025).

Furthermore, identifying the sources and controlling factors of nitrate pollution is crucial for improving prediction accuracy and guiding targeted pollution control measures. Isotopic analysis ($\delta^{15}N\text{-}NO_3^-$ and $\delta^{18}O\text{-}NO_3^-$) combined with the MixSIAR model has proven effective in

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

quantitatively apportioning nitrate sources (Tian et al., 2025). Meanwhile, Bayesian models and SHapley Additive exPlanations (SHAP) analysis can reveal the key environmental variables driving nitrate concentration changes, enhancing the interpretability of prediction models (Alam et al., 2025). Despite these advancements, several gaps persist in the current research: (1) Few studies have integrated hybrid quantum-classical ML with virtual sample augmentation to address small-sample challenges in seasonal nitrate prediction; (2) The potential of AEF remote sensing semantic features for groundwater nitrate prediction remains untested, particularly in comparison with in-situ measured parameters; (3) The combined effects of hydrological seasonal variations, nitrate source apportionment, and key environmental drivers on prediction model performance require systematic investigation.

The North China Plain, an important agricultural production region in China, is characterized by high nitrogen input intensity and significant seasonal hydrological variations, making it an area prone to groundwater nitrate pollution (Liu et al., 2025). Conducting field-scale research on nitrate pollution in this region is of great significance for the protection of regional water resources. The Xiong'an New Area in China is a typical study area at the farmland scale in the North China Plain. As a major agricultural area with high nitrogen input intensities and distinct seasonal hydrological cycles, it faces significant groundwater nitrate pollution risks. This region's unique climatic regime characterized by a dry spring, wet summer with concentrated precipitation, and a cold, dry winter, creates marked seasonal disparities in groundwater recharge, evaporation, and pollutant migration (Xu et al., 2022). A mechanistic understanding of how nitrate concentrations vary across these hydrological seasons (normal, dry, wet) and their controlling factors is crucial for regional water resource management.

To fill these gaps, this study aims to: (1) propose a novel virtual sample generation method (t-SNE-GMM-KNN) to enhance small-sample datasets while preserving the non-linear structure and multi-modal distribution of original data; (2) construct a hybrid quantum-classical random forest model by integrating quantum feature encoding with classical random forest, improving the model's ability to capture complex environmental relationships; (3) evaluate the predictive performance of two input datasets (on-site measured water quality parameters and AEF remote sensing semantic features) across normal, dry, and wet seasons under Leave-One-Out Cross-Validation (LOOCV); (4) identify the dominant nitrate sources and key environmental

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

controlling factors using isotopic analysis, MixSIAR modeling, Bayesian analysis, and SHAP interpretation; (5) establish a comprehensive and accurate prediction framework for groundwater nitrate concentrations in intensive agricultural regions, providing scientific support for groundwater pollution control and sustainable water resource management in the North China Plain. The novelty of this study lies in the integration of hybrid quantum-classical machine learning, advanced virtual sample augmentation, and remote sensing semantic features to address the challenges of small-sample, high-dimensional, and seasonally variable nitrate prediction. The findings are expected to advance the state-of-the-art in groundwater quality prediction and offer a scalable approach for large-scale environmental monitoring in unmonitored areas.

2. Materials and Methods

2.1 Study area

The North China Plain is one of China's most important agricultural production bases. This study focuses on the Xiong'an New Area, situated in the central part of Hebei Province, as a representative research site within this plain. Located in the core region defined by Beijing, Tianjin, and Baoding, it boasts an advantageous geographical position, with straight-line distances of 105 km to both Beijing and Tianjin, and 30 km to Baoding. Its geographical coordinates range from 38°43′ to 39°10′ N latitude and from 115°38′ to 116°20′ E longitude, covering an area of approximately 1770 km^2 (Xiong'an New Area Official Website, 2023). The specific study area is an unmanned farm located in Xieyeqiao Village, Nanzhang Town, Rongcheng County, within the Xiong'an New Area (Fig.1). The farm covers an area of 3000 hectares and primarily cultivates two main grain crops: wheat and corn. As the first mechanized unmanned farm in Xiong'an, it has achieved full mechanization and intellectualization, enabling unmanned, precise, and standardized operations throughout all stages of tillage, sowing, management, and harvesting.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

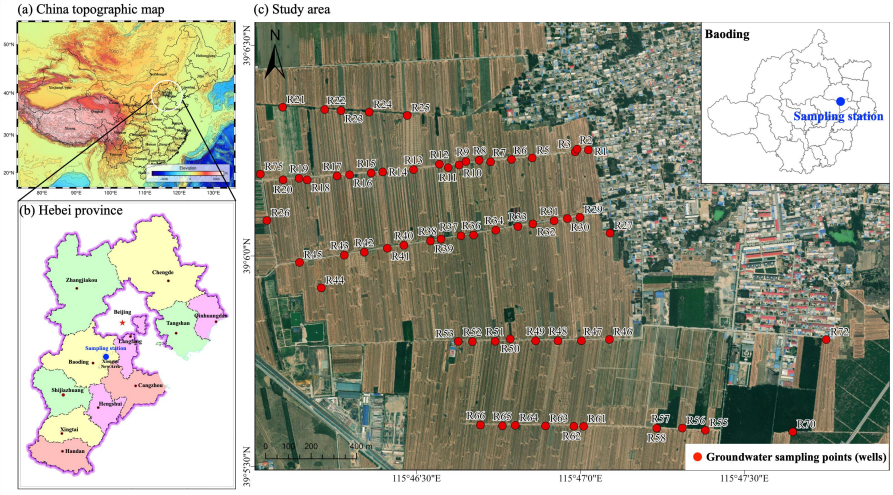


Fig.1. Study area map showing the sampling location. ((b) Based on the standard map (Approval No. Ji S (2025) 009) from the Department of Natural Resources of Hebei Province; base map is unmodified.)

Cultivated land constitutes a large proportion of the total area in the Xiong'an New Area and is predominantly dryland. Traditional fertilization in the region involves high application rates of nitrogen and manure. As a representative farm within this area, the study site also follows this conventional practice, making it susceptible to the impacts of high fertilization intensity. The annual nitrogen fertilizer application rate at the study site ranges from 540 to 660 kg (N) ha^{-1} yr^{-1}, primarily supplied as urea (46% N). The extensive application of chemical fertilizers and manure consequently elevates the risk of nitrogen pollution in groundwater. Furthermore, the rural population is relatively densely distributed, contributing to pollution from domestic sewage discharge in the vicinity. The climate is classified as a warm-temperate, monsoonal, continental semi-humid climate. Springs are dry and rainless, summers are humid with abundant precipitation, autumns are cool and dry, and winters are cold with minimal snowfall. The mean annual air temperature in Xiong'an New Area is 12.6°C, exhibiting relatively minor inter-annual fluctuations. The mean annual precipitation is 480.8 mm, which is highly concentrated from June to September. The average annual sunshine duration is 2335.2 hours, with longer periods in spring and summer and shorter ones in autumn and winter. The average frost-free period lasts 204 days. The mean annual wind speed is 1.7 m s^{-1}, with the highest average occurring in April and the lowest in January, August, and December. The multi-year average evaporation is 1661.1 mm (Liao et al.,

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

2020). The soil texture is dominated by silty loam, and the 2-8.5 m soil layer contains interlayers with high clay content such as clay and silty clay, reflecting the characteristics of vadose zone sediments in the central plain under geomorphic sedimentation. Nitrogen in the thick vadose zone is dominated by organic nitrogen, accounting for approximately 97% of the total nitrogen content. The shallow vadose zone at 3-6 m stores the largest amount of nitrate, accounting for about half of the total nitrate reserves in the North China Plain (Li et al., 2025; Zhang et al., 2007). Groundwater in the study area is primarily hosted in Quaternary unconsolidated porous aquifers, with sampled wells ranging from 70 to 120 m in depth (Bai et al., 2023). The primary source of groundwater recharge in the study area's farmland is atmospheric precipitation, while the main discharge pathway is artificial extraction for agricultural irrigation. Irrigation followed crop phenological stages. Wheat underwent muddy water irrigation at pre-sowing, overwintering, regreening, and jointing stages, and maize received a single post-sowing muddy water irrigation.

2.2 Data collection and measurements

2.2.1 Field sampling data and laboratory analysis

Field investigations and the collection of hydrochemical and isotopic samples were conducted in the study area from 2022 to 2023. A total of 66, 65, and 50 groundwater samples were collected in October 2022, April 2023, and August 2023, respectively. All groundwater samples were obtained from existing agricultural irrigation wells within the study area. Prior to sample collection, each well was purged by pumping. Sampling commenced only after the pumped volume exceeded three times the well's casing volume and on-site parameters had stabilized (i.e., showing minor fluctuations around a constant value rather than a continuous rising or falling trend), a procedure implemented to ensure the representativeness of the samples. At each sampling point, one 1000 mL and two 100 mL samples were collected. Before final collection, the sample bottles were rinsed three times with the water to be sampled. Immediately after collection, the samples were sealed and stored in a portable cooler for transport to the laboratory for subsequent analysis. Furthermore, the precise geographical location of each sampling point was recorded using a GPS device.

In-situ physicochemical parameters were measured using a Hach HQ400 multi-parameter water quality meter (Li et al., 2022). The measured parameters included water temperature (T, °C),

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

pH, total dissolved solids (TDS, mg L^{-1}), dissolved oxygen (DO, mg L^{-1}), electrical conductivity (EC, μS cm^{-1}), and oxidation-reduction potential (ORP, mV). The concentration of HCO_3^- was determined within 24 hours of sample collection using the dilute sulfuric acid-methyl orange titration method (Huang et al., 2012). Prior to the determination of cations and anions, water samples were filtered through 0.45 μm membrane filters. Major cations (K^+, Ca^{2+}, Na^+, Mg^{2+}) were analyzed using an inductively coupled plasma optical emission spectrometer (Avio 500). Major anions (NO_3^-, Cl^-, SO_4^{2-}) were analyzed using an ion chromatograph (ICS-2100). The analytical precision for cations and anions was controlled within ±0.2 mg L^{-1}, and the charge balance error was maintained within 5% to ensure reliability. The concentrations of nitrite nitrogen and ammonia nitrogen were determined using a flow injection analyzer (Smartchem 200, AMS Alliance) and measured using dual wavelength spectrophotometry and the indophenol blue method (Kim et al., 2019; Sun et al., 2022). The limits of detection for nitrite nitrogen and ammonium nitrogen were both 0.01 mg L^{-1}. For the analysis of stable hydrogen and oxygen isotopes, water samples were filtered through 0.22 μm membrane filters and measured using an LGR liquid water isotope analyzer (TIWA-45-EP). The analytical precisions for δ^2H, $\delta^{17}O$, and $\delta^{18}O$ were ±0.15‰, ±0.02‰, and ±0.02‰, respectively (Hamidi et al., 2023). The isotopic compositions of nitrate ($\delta^{18}O$-NO_3^- and $\delta^{15}N$-NO_3^-) were determined using a MAT-253 mass spectrometer coupled with an elemental analyzer (Li et al., 2022). To ensure analytical precision, standard references, reagent blanks, and duplicate samples were employed. Furthermore, international standards USGS 34 and USGS 35 were used for $\delta^{18}O$ quality control, while USGS 32 and USGS 34 were used for $\delta^{15}N$ quality control. All isotope results are reported in per mil (δ, ‰).

2.2.2 Google AlphaEarth Foundation

To facilitate comparisons with predictions based on in-situ field sampling data and to validate the accuracy of predicting groundwater nitrate concentration using remote sensing data, this study incorporates the Google AlphaEarth Foundation (AEF) dataset. AEF is a collection of high-dimensional surface semantic embedding features generated via pre-training on multi-source remote sensing data (Brown et al., 2025). By fusing imagery from Sentinel-2, Landsat, and other Earth observation satellites, this dataset constructs a 64-dimensional vector representation

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

(denoted as A00-A63) at a global scale with an annual temporal resolution and a 10 m spatial resolution (Alvarez et al., 2025). These embeddings implicitly encode complex environmental semantics, such as land cover types, vegetation dynamics, soil moisture, and the intensity of human activity, and have been successfully applied in tasks including land use classification, crop monitoring, and environmental risk modeling (Tollefson et al., 2025).

The primary processing workflow involved spatially sampling the 64-dimensional AEF vectors at a 10 m resolution using the GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL product on the Google Earth Engine (GEE) platform, based on the geographic coordinates of the field sampling points. To ensure data quality, only samples exhibiting exact matches between the GEE extraction and the actual field data points were retained. Given the redundancy within the initial 64-dimensional AEF features, Principal Component Analysis (PCA) based on Singular Value Decomposition (SVD) was employed for feature compression. Specifically, SVD was performed on the centered feature matrix to select the minimum number of principal components accounting for at least 95% of the cumulative explained variance (Ilyas et al., 2025). The orthogonalized, low-dimensional principal component scores were subsequently used as model inputs. This approach preserves the vast majority of the semantic information from the original embeddings while significantly mitigating the risk of overfitting. Ultimately, the PCA-reduced AEF features served as the input variables for the model.

2.3 MixSIAR model and isotopic composition of nitrate sources

The MixSIAR model uses prior information such as the number of end - members, errors, and distribution characteristics, and iterates based on the Markov Chain Monte Carlo (MCMC) method to quantitatively restore the contribution fraction of each end-member to the mixed sample (Stock et al., 2018). At present, this analytical method has been widely applied in the quantitative analysis of nitrate pollution sources in water bodies. The calculation principle of the model is as follows:

$$X_{ij} = \sum_{k=1}^{k} P_k (S_{jk} + C_{jk}) + \varepsilon_{ij} \quad (1)$$

$$S_{jk} \sim N(\mu_{jk}, \omega_{jk}^2) \quad (2)$$

$$C_{jk} \sim N(\lambda_{jk}, \tau_{jk}^2) \quad (3)$$

$$\varepsilon_{ij} \sim N(0, \sigma_j^2) \quad (4)$$

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

In the formula, X_{ij} is the value of isotope j in the i-th sample (i = 1, 2, …, 20, j = 1, 2); P_k is the contribution rate of the k-th pollution source; S_{jk} is the value of isotope j in the k-th pollution source, where μ is the mean and ω is the variance of the normal distribution; C_{jk} is the fractionation coefficient, where λ is the mean and τ is the variance of the normal distribution; ε_{ij} is the residual, with 0 as the mean and σ as the variance of the normal distribution.

In this study, the MixSIAR model is used to calculate the five potential sources of NO_3^- in water bodies, namely precipitation (NP), soil organic nitrogen (SON), synthetic NH_4^+ fertilizer (NHF), synthetic NO_3^- fertilizer (NOF), and domestic sewage & manure (DSM). The end-member values of the five sources are selected as shown in Table 1 (Mao et al., 2023; Gao et al., 2023; Torres-Martínez et al., 2021).

Table 1. Summary statistics of $\delta^{18}O$ and $\delta^{15}N$ for potential nitrate sources.

Sources	$\delta^{18}O$-NO_3^-		$\delta^{15}N$-NO_3^-	
	Mean	SD	Mean	SD
NP	57.2	6.9	0.6	1.5
NHF	-4.1	2.7	-2.1	0.7
NOF	21.7	2.9	0.2	2.3
SON	-2.7	4.4	3.8	1.8
DSM	6.1	1.6	17.4	3.9

2.4 t-SNE-GMM-KNN: based on nonlinear structure modeling in feature space

To address the challenges of overfitting and poor generalization performance in small-sample modeling, which arise from data sparsity and skewed distributions, this study proposes a three-stage virtual sample generation strategy termed t-SNE-Gaussian Mixture Sampling with KNN Inverse mapping. This method aims to preserve the non-linear manifold structure and multi-modal distribution characteristics of the original high-dimensional feature space while generating physically plausible and statistically consistent synthetic samples. The specific workflow is as follows:

1. Data standardization

All input features are standardized using Z-score standardization to eliminate scale

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

differences and enhance the stability of the subsequent dimensionality reduction (Jamshidi et al., 2022).

2. t-SNE non-linear dimensionality reduction

t-Distributed Stochastic Neighbor Embedding (t-SNE) is employed to map the high-dimensional feature space into a low-dimensional latent space (d=2) (Islam et al., 2023). To balance the preservation of local and global structures, the perplexity is set to 10, and PCA initialization is used to ensure reproducibility. t-SNE effectively reveals the clustered structure of samples on the low-dimensional manifold, reflecting the differentiation of underlying environmental processes within hydrological seasons (Liu et al., 2021).

3. GMM clustering and optimal component selection

In the t-SNE-reduced low-dimensional space, a Gaussian Mixture Model (GMM) is constructed to characterize the probability density distribution of the data (Jia et al., 2022). The GMM assumes that the data are generated from a linear combination of several Gaussian distributions. The weights, means, and covariance matrices of each Gaussian component are estimated via the Expectation-Maximization (EM) algorithm, thereby accurately capturing the complex distribution patterns of the data (Yan et al., 2023). To avoid subjectively setting the number of clusters, the Bayesian Information Criterion (BIC) is used to automatically optimize the number of components, K, within the range (Ghodba et al., 2025):

$$BIC(K) = -2\log L + p_K \log n \tag{5}$$

where L is the model's likelihood, k is the total number of free parameters for a K-component model, and n is the sample size. The value of K corresponding to the minimum BIC is selected as the optimal number of components, ensuring a balance between goodness-of-fit and model complexity.

4. Virtual sample generation and inverse mapping

Based on the optimal GMM, a specified number of virtual points are randomly sampled from its joint probability distribution. This generation process naturally inherits the multi-modality and covariance structure of the original data. To reconstruct the low-dimensional virtual samples back into the original feature space, a k-Nearest Neighbors regression model is trained (Niu et al., 2025). This model uses the t-SNE coordinates as input and the standardized original features as output, approximating the inverse of the non-linear t-SNE mapping. Finally, the virtual sample set in its

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

original physical units is obtained by applying inverse standardization.

5. Physical constraints and quality control

For the target variable, NO_3^-, a non-negativity constraint ($NO_3^- \geq 0$ mg L^{-1}) is imposed to prevent non-physical solutions that may arise from the regression approximation. Other variables, such as pH and ORP, are allowed to fluctuate within reasonable ranges without hard clipping to retain the model's flexibility. The consistency between the virtual and measured samples is validated by comparing their statistical characteristics, including mean, standard deviation, coefficient of variation, range of extreme values, and boxplot distributions. This comparison confirms that the generated data are highly consistent with the original data in terms of statistical properties, without introducing systematic bias or outliers.

The advantages of this method are as follows: ① t-SNE excels at capturing local neighborhood relationships, effectively separating implicit subgroups under different hydrological conditions. ② The GMM provides a probabilistic generative framework, supporting reasonable extrapolation for heavy-tailed distributions and extreme values. ③ The KNN-based inverse mapping circumvents the need for large training datasets, which is a limitation of traditional autoencoders, making it particularly suitable for small-sample scenarios (Tang et al., 2022; Razavi-Termeh et al., 2024).

2.5 Machine learning methods

2.5.1 Random forest

In this study, Random Forest (RF) was adopted as the baseline model. As an ensemble learning method that leverages bootstrap sampling and random feature selection, RF builds numerous decision trees and integrates their predictions (Abderzak et al., 2025). This approach effectively suppresses overfitting and improves generalization performance, making it especially well-suited for environmental data modeling scenarios involving small samples, high dimensionality, non-linearity, and multicollinearity (Boddu et al., 2025). Hyperparameters were configured based on a preliminary grid search and domain expertise: n_estimators=100, max_depth=5, min_samples_split=6, min_samples_leaf=3, and max_features= $\sqrt{p}$. For the interpretation of driving mechanisms, feature importance was quantified by the mean decrease in Gini impurity (Gini Importance) to identify the critical hydrogeochemical indicator factors (Kaur

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

et al., 2025).

2.5.2 Hybrid quantum-classical random forest

Based on the random forest, a Hybrid Quantum-Classical Random Forest (QCRF) model was constructed, integrating quantum feature enhancement with classical random forests. The core idea of the model is: utilizing a Parameterized Quantum Circuit (PQC) to perform quantum feature encoding on standardized input features, generating high-dimensional quantum features with non-linear entanglement properties (Naresh et al., 2025). These are then concatenated with the original features to construct an enhanced hybrid feature space, which is finally fed into a random forest regressor for modeling (Lamichhane et al., 2025).

(1) Quantum Feature Encoding

Quantum state transformation of classical data is achieved based on the Z-feature map in quantum computing (Vedavyasa et al., 2025). The ZFeatureMap maps the classical high-dimensional feature space into a quantum Hilbert space through single-qubit Z-gate operations and two-qubit CZ-gate entanglement operations (Khalil et al., 2025). Its core advantage lies in obtaining quantum features via analytical calculation of quantum state vectors, thereby avoiding noise interference introduced by quantum sampling and ensuring feature stability. The ZFeatureMap provided by Qiskit is used as the feature encoding circuit, with its Hamiltonian form given by (Tehrani et al., 2024):

$$U_{ZMap}(x) = \prod_{k=1}^{R} \left[\bigotimes_{i=1}^{d} H_i \cdot \exp\left(-i \sum_{S \subseteq \{1,\ldots,d\}} \phi_S(x) \bigotimes_{j \in S} Z_j \right) \right] \quad (6)$$

where d is the selected number of principal factors, R is the number of repetition layers, and ϕ_S represents data-dependent rotation angles (using linear embedding).

For each sample x, construct the corresponding quantum state $|\psi(x)\rangle$, and analytically calculate the Pauli-Z expectation value for each qubit (Liao et al., 2024):

$$\langle Z_i \rangle = \langle \psi(\boldsymbol{x}) | Z_i | \psi(\boldsymbol{x}) \rangle = P(q_i=0) - P(q_i=1) \quad (7)$$

Here, the number of i is equal to the number of predictor variables. This method requires no quantum hardware sampling, completely avoiding the interference of measurement noise and shot noise on small-sample modeling, thus ensuring the determinism and reproducibility of feature generation.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

(2) Feature Fusion and Modeling

Concatenate the original n-dimensional raw features with the n-dimensional quantum ⟨Z⟩ features to form a 2n-dimensional hybrid feature vector x_aug = [x_raw; ⟨Z⟩] (Cowlessur et al., 2025). That is, original features+quantum encoded features. Using this as input, construct a random forest regression model:

$$\hat{y}=\frac{1}{M}\sum_{m=1}^{M}\mathrm{Tree}_m(x_{aug}) \tag{8}$$

The hyperparameter settings are the same as those for the classical random forest method described above.

2.6 Evaluation methods and prediction process

2.6.1 SHAP analysis

This study adopts the SHapley Additive exPlanations (SHAP) method for local and global explainability analysis (Merabet et al., 2025). Through three typical visualization methods, namely summary plot, dependence plot, and waterfall plot, the following are revealed respectively: (1) The overall ranking and distribution of feature importance across all samples (global perspective); (2) The nonlinear relationship or interaction effects between a single predictor variable and the predicted nitrate concentration (conditional dependence); (3) The contribution decomposition of each feature in the prediction result of a representative sample (local attribution) (Alam et al., 2025).

The SHAP value is mathematically defined as: the marginal contribution of feature j to the model output offset from the baseline mean (Li et al., 2024), and its form is:

$$\phi_j=\sum_{S\subseteq F\setminus\{j\}}\frac{|S|!\,(|F|-|S|-1)!}{|F|!}[f(S\cup\{j\})-f(S)] \tag{9}$$

where F is the set of all features, S is a subset not containing feature j, and f is the model output. By averaging the absolute SHAP values $|\phi_j|$ over all samples, a feature importance measure with a game-theoretic foundation, unbiased and robust, can be obtained (Hollmannet al., 2025).

2.6.2 Leave-One-Out Cross-Validation (LOOCV) and model evaluation indicators

Given the limited sample size in each hydrological season, this study adopts Leave-One-Out Cross-Validation (LOOCV) for model performance evaluation to maximize the use of training

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

data and reduce evaluation bias (Austin et al., 2025). The LOOCV process is: each time, one sample is left out as the validation set, and the remaining n-1 samples are used for training. After repeating nn times, the average of the evaluation indicators is taken as the final result (Ren et al., 2021).

The coefficient of determination R^2, root mean square error (RMSE), and mean absolute error (MAE) are used to quantitatively describe the model accuracy and error characteristics (Gul et al., 2025):

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (10)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (11)$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (12)$$

where y_i is the measured value of the i-th sample, $\hat{y}_i$ is the model's predicted value, $\bar{y}$ is the mean of the measured values, and n is the number of samples in the test set.

2.6.3 Standardized prediction workflow

To systematically evaluate the nitrate concentration prediction capabilities of different input variables and modeling strategies across various hydrological seasons, and to validate the effectiveness of virtual sample augmentation for small-sample modeling, this study established a standardized prediction pipeline (Fig.2). The specific steps are as follows: (1) Data Preprocessing and Grouping: Observed samples were partitioned by seasons. Z-score normalization was applied separately to two types of input features: field water quality parameters and AlphaEarth Foundation (AEF) features reduced via Principal Component Analysis (PCA). (2) Virtual Sample Generation and Validation: A t-SNE-GMM-KNN strategy was employed to generate virtual samples. Their physical plausibility and distribution consistency were rigorously verified using statistical indicators, box plots, and histograms. (3) Model Training: Under unified hyperparameters, classical Random Forest (RF) and quantum-enhanced RF models were constructed. The latter generates <Z> quantum features via Parameterized Quantum Circuits (PQC) encoding, which are concatenated with original features to form 2*n input features. Models were trained using two distinct input datasets and combinations of "original samples + 1~10× virtual samples." (4) Model Evaluation: The Leave-One-Out Cross-Validation (LOOCV) strategy was

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

adopted to calculate R^2, RMSE, and MAE. Visual diagnostics were performed using observed-predicted scatter plots, residual plots, and box plots. (5) Interpretability Analysis: Multi-scale interpretation was conducted based on the SHAP framework, including summary plots (global importance ranking), dependence plots (nonlinear response and interaction effects), and waterfall plots (local attribution). The driving mechanisms were cross-verified with results from Bayesian models and Pearson correlation analysis. This workflow encompasses the full process from data augmentation, modeling, and evaluation to attribution, providing a reproducible and highly transparent solution for precise groundwater nitrate prediction under conditions of small samples, multiple seasons, and multi-source inputs.

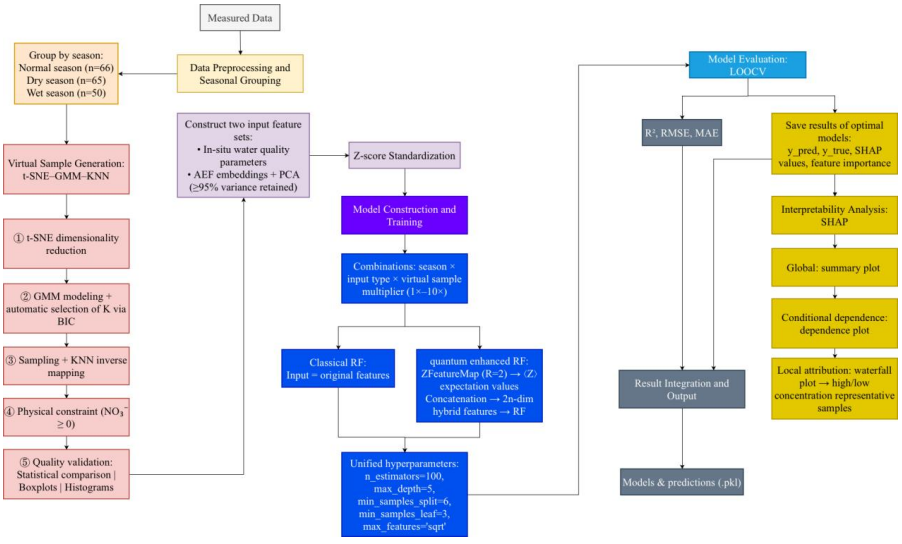


Fig.2. Process diagram for constructing prediction framework.

3. Results

3.1 Seasonal hydrochemical controls of nitrate distribution in farmland groundwater

3.1.1 Hydrochemical parameters

Regarding the basic physical parameters, the pH value was weakly alkaline during the normal water period with minimal variation, whereas it was near-neutral in the dry and wet seasons (Table 2). A minimum value of 5.77 occurred in the wet season, indicating the presence of acidic water bodies. Temperature (T) exhibited significant seasonal variation but remained relatively stable within each season (CV≈0.06). EC, salinity, and TDS showed consistent patterns, all peaking

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

during the dry season and reaching their lowest levels in the wet season. The redox indicators displayed high volatility. The mean DO was slightly higher in the wet season, while the mean ORP was consistently low across all seasons, with extremely large standard deviations and coefficients of variation. In terms of ionic composition, the average concentrations of Ca^{2+} and Mg^{2+} were highest during the dry season, and Na^{+} also peaked in this period. The concentration of K^{+} was relatively low. Among the anions, HCO_3^{-} concentration was highest in the wet season, while the average concentrations of Cl^{-} and SO_4^{2-} were both at their maximum during the dry season. The average concentration of NO_3^{-} was higher in the dry season than in other periods and lowest in the wet season. The concentrations of nitrite (NO_2^{-}) and ammonium (NH_4^{+}) were much lower than that of nitrate. Concerning the distribution of the indicators, most variables were right-skewed. Notably, extreme values were present for NO_3^{-} in the dry season (maximum=358.58 mg L^{-1}, mean=42.93 mg L^{-1}), Cl^{-} in the dry season (maximum=241.36 mg L^{-1}, mean=24.90 mg L^{-1}), and F^{-} in the normal period (maximum=13.17, mean=3.70).

Table 2. Statistical summary of chemical and field measurement parameters.

Periods		pH	T	EC	DO	ORP	Salt	TDS	Depth	K^{+}	Ca^{2+}
	Unit		℃	μs cm^{-1}	mg L^{-1}	mv	ppt	mg L^{-1}	m	mg L^{-1}	mg L^{-1}
Normal season	Max	8.60	16.70	1110.00	10.57	369.60	0.53	724.00	20.64	36.66	69.73
n=66	Min	7.61	13.40	349.00	2.20	-58.30	0.11	225.00	18.26	1.04	13.38
	Mean	8.16	14.90	549.32	6.31	4.57	0.21	357.21	18.91	2.92	37.28
	SD	0.11	0.71	173.56	1.94	56.27	0.09	113.34	0.62	4.45	13.06
	CV	0.01	0.05	0.32	0.31	12.32	0.44	0.32	0.03	1.52	0.35
Dry season	Max	8.21	18.80	1134.00	8.82	144.20	0.54	737.00	18.95	3.52	43.86
n=65	Min	6.97	14.10	343.00	1.85	-110.40	0.11	227.00	17.88	0.67	4.81
	Mean	7.35	15.59	658.68	6.34	4.55	0.27	427.65	18.43	1.93	16.89
	SD	0.45	0.92	185.20	1.64	42.10	0.10	120.55	0.27	0.56	8.35
	CV	0.06	0.06	0.28	0.26	9.26	0.36	0.28	0.01	0.29	0.49
Wet season	Max	8.97	21.00	977.00	9.61	195.00	0.41	635.00	18.87	3.37	51.54
n=50	Min	5.77	15.50	371.00	3.34	-112.20	0.12	243.00	17.61	1.14	17.27
	Mean	7.34	17.01	535.24	6.98	17.57	0.20	347.92	18.27	1.75	25.70

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

	SD	0.52	0.94	153.65	1.32	47.59	0.08	100.39	0.40	0.34	6.83
	CV	0.07	0.06	0.29	0.19	2.71	0.42	0.29	0.02	0.19	0.27
		Na^+	Mg^{2+}	HCO_3^-	Cl^-	SO_4^{2-}	F	NO_3^-	NO_2^-	NH_4^+	
	Unit	mg L^{-1}	mg L^{-1}	mg L^{-1}	mg L^{-1}	mg L^{-1}	mg L^{-1}	mg L^{-1}	mg L^{-1}	mg L^{-1}	
Normal season	Max	98.85	92.78	192.15	87.09	40.49	13.17	161.17	0.45	0.20	
n=66	Min	3.81	3.80	22.88	1.03	1.79	0.12	2.39	0.04	0.00	
	Mean	24.01	31.78	96.65	20.53	15.53	3.70	33.67	0.11	0.04	
	SD	12.46	17.18	44.51	17.84	10.02	2.02	35.83	0.08	0.04	
	CV	0.52	0.54	0.46	0.87	0.64	0.55	1.06	0.72	1.00	
Dry season	Max	123.20	138.90	289.29	241.36	123.75	0.57	358.58	10.38	0.84	
n=65	Min	16.28	17.97	5.10	1.40	2.00	0.21	0.10	0.57	0.05	
	Mean	48.52	58.17	42.19	24.90	15.91	0.33	42.93	3.35	0.16	
	SD	16.48	25.86	71.11	35.13	18.67	0.08	56.35	1.98	0.11	
	CV	0.34	0.44	1.69	1.41	1.17	0.26	1.31	0.59	0.69	
Wet season	Max	36.21	53.72	207.40	51.93	34.15	0.45	98.36	4.31	0.41	
n=50	Min	10.95	12.76	83.88	1.56	5.88	0.09	4.15	1.51	0.02	
	Mean	21.78	22.57	132.49	15.60	15.16	0.20	27.14	2.62	0.10	
	SD	4.90	8.13	27.12	13.60	7.47	0.08	23.86	0.61	0.08	
	CV	0.23	0.36	0.20	0.87	0.49	0.39	0.88	0.23	0.83	

3.1.2 Type of water

During the dry season, the data points are highly concentrated in the zone of calcium-type cations and bicarbonate-type anions, indicating that the groundwater is primarily controlled by the dissolution of carbonate rocks (Fig.3). In the wet season, although the Ca-Mg-HCO_3^- type remains dominant, some samples shift towards the sulfate and chloride types, reflecting the leaching input effect of surface pollutants (such as agricultural fertilizers and domestic sewage) brought by rainfall infiltration. By the normal season, the hydrochemical types exhibit the widest distribution, presenting a mixed type with coexisting bicarbonate and chloride types. Overall, the groundwater hydrochemical characteristics in the study area are jointly controlled by precipitation-evaporation dynamics and carbonate weathering.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

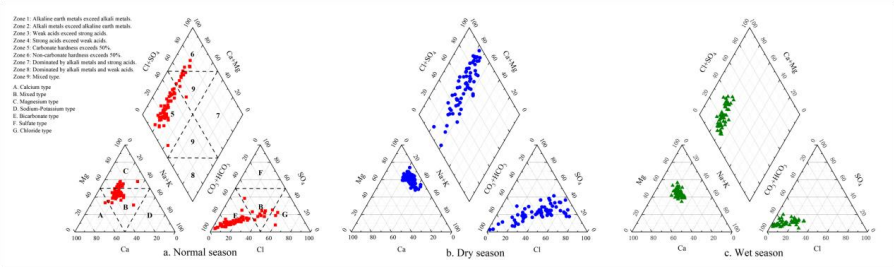


Fig.3. Piper diagram classifying the hydrochemical facies of the analyzed groundwater.

3.1.3 Sources and controlling factors of ions in groundwater

The Gibbs diagram shows that the groundwater in the study area is primarily controlled by rock weathering during the normal, dry, and wet seasons, indicating the dominance of water-rock interaction (Fig. 4). The ratio of $\gamma(Na^+ + K^+)$ to γCl^- (Fig. 5a) shows that the vast majority of sample points plot above the 1:1 line, indicating that Na^+ and K^+ are primarily sourced from the dissolution of evaporite rocks. In the relationships between $\gamma(Ca^{2+} + Mg^{2+})$ and γHCO_3^-, and between $\gamma(Ca^{2+} + Mg^{2+})$ and $\gamma(HCO_3^- + SO_4^{2-})$ (Fig. 5b-c), samples from all periods plot above the 1:1 line, confirming that Ca^{2+} and Mg^{2+} mainly originate from the dissolution of carbonate minerals. Furthermore, the γCa^{2+}-γMg^{2+} relationship (Fig. 5d) helps identify the types of mineral dissolution. Samples from the dry season are concentrated below the 1:2 line, indicating a dominance of magnesium-poor mineral dissolution, with cation exchange causing a relative depletion of Ca^{2+}. Samples from the normal and wet seasons are stably distributed between the 1:1 and 1:2 lines, reflecting that dolomite dissolution has reached equilibrium while calcite remains in a state of non-equilibrium dissolution, continuously supplying Ca^{2+}. In the relationship between $\gamma(SO_4^{2-} + Cl^-)$ and γHCO_3^- (Fig. 5e), the distribution of sample points on both sides of the 1:1 line suggests that groundwater ions have dual contributions from both evaporite and carbonate rocks. Conversely, in the γCa^{2+} versus γSO_4^{2-} relationship (Fig. 5f), samples generally plot above the 1:1 line, which excludes gypsum as a primary source of Ca^{2+} and indicates that Ca^{2+} is mainly derived from the dissolution of carbonate minerals. Therefore, the chemical composition of groundwater in the study area is primarily controlled by the dissolution of carbonate minerals, and is also influenced by hydrological seasonal variations and cation exchange processes.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

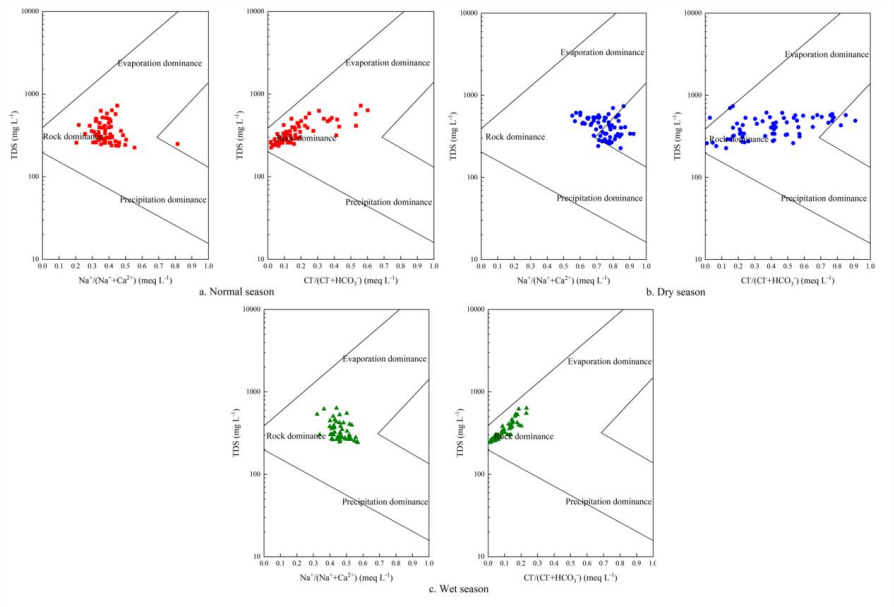


Fig.4. Gibbs diagrams of the groundwater samples.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

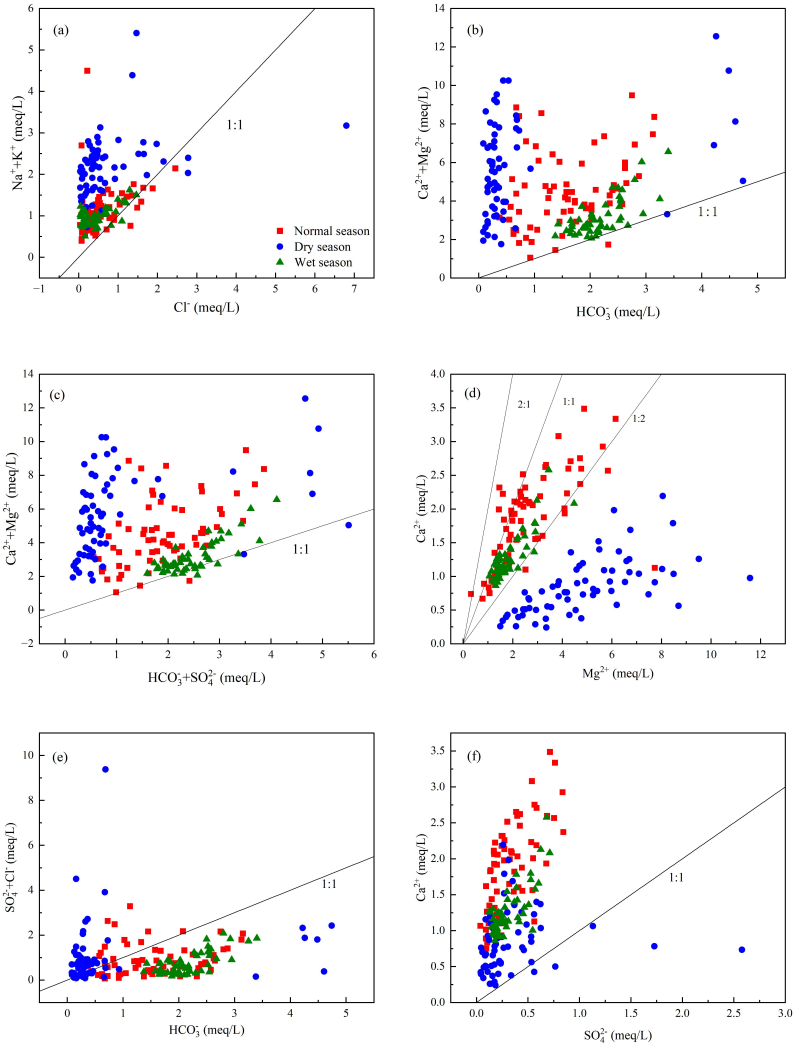


Fig.5. Plots of ion ratio relationship.

The Chloro-Alkaline Index method was employed to analyze the cation exchange and adsorption between groundwater and sediments. A CAI value less than zero indicates the occurrence of cation exchange, with more negative values reflecting stronger exchange intensity. . Furthermore, the relationship between [$\gamma(Ca^{2+}) + \gamma(Mg^{2+}) - \gamma(HCO_3^-) - \gamma(SO_4^{2-})$] and [$\gamma(Na^+) - \gamma(Cl^-)$] can be used to further investigate the cation exchange processes in the groundwater. During

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

the dry season, the slope was 0.55, suggesting the presence of extremely weak cation exchange in the water body (Fig. 6a). With the exception of samples from the dry season, sampling points from the normal and wet seasons were plotted near a line with a slope of -1, with respective slopes of -1.52 and -1.36. This trend is consistent with the conclusions drawn from the Chloro-Alkaline Index, providing further evidence that cation exchange and adsorption occurred in the groundwater during the normal and wet seasons. The ion exchange process was more active during the rainy season (R^2=0.41), leading to the enrichment of Na^+ in the groundwater of the area. In contrast, most groundwater samples from the dry season showed no evidence of cation exchange and adsorption.

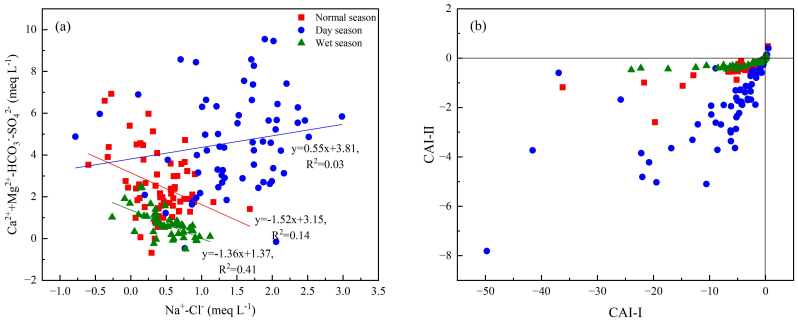


Fig.6. Relationship diagram of groundwater (Ca^{2+}+Mg^{2+}-SO_4^{2-}-HCO_3^-) and (Na^+-Cl^-) along with CAI-1 and CAI-2 correlation diagrams.

3.1.4 Spatial distribution dynamics of groundwater depth and nitrate driven by seasonal hydrological processes

The spatial distribution of groundwater depth reflects the regional hydraulic gradient and groundwater flow direction, whereas the spatial variability of nitrate concentration is closely associated with flow paths, pollution source inputs, and hydrological processes (Fig.7). During the normal season, the groundwater depth distribution is relatively uniform. The eastern region of the farm, characterized by shallower depths, serves as a recharge zone, with groundwater flowing towards the deeper western region. At this time, nitrate concentration are relatively dispersed, with high-concentration zones located in the southeastern part of the farm. In the dry season, the groundwater depth becomes shallower, and the flow direction shifts from the eastern and western

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

sides towards the central area. During this period, nitrate concentration reach their annual peak (mean: 42.93 mg L^{-1}). The distribution of nitrate exhibits a higher degree of spatial coincidence with the groundwater flow direction, indicating that enhanced evaporative concentration during the dry season leads to the further accumulation of flow-transported pollutants in the discharge zone. In the wet season, the groundwater depth further decreases, and groundwater flows from the southeastern region towards the northwestern region. nitrate concentration drop to their annual minimum (mean: 27.14 mg L^{-1}). High-concentration areas are distributed in the northwest, overlapping with regions of deeper groundwater depth. It is inferred that precipitation infiltration during the rainy season dilutes the groundwater nitrate; as dilution is the dominant process during infiltration, the nitrate concentration exhibits a decreasing trend along the groundwater flow direction.

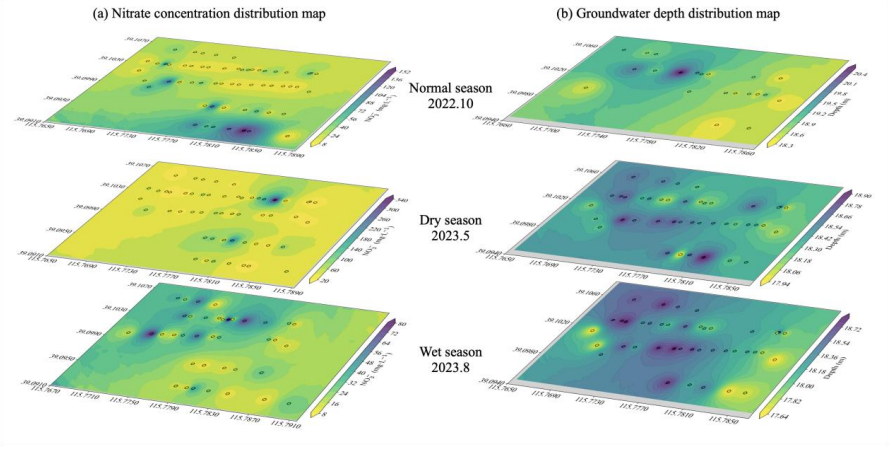


Fig. 7. Spatial distribution of nitrate concentration and groundwater depth in different seasons.

3.2 Groundwater recharge sources and pollution source identification

3.2.1 Stable hydrogen and oxygen isotope composition of water

During the normal season, the mean values of groundwater δD and $\delta^{18}O$ were -61.31‰ and -7.31‰, with ranges of -73.40‰ to -53.52‰ and -10.25‰ to -2.82‰, respectively. The d-excess values ranged from -38.07‰ to 17.30‰, with a mean of -2.80‰. In the dry season, the mean groundwater δD and $\delta^{18}O$ values were -71.05‰ and -9.74‰, with ranges of -76.93‰ to -60.55‰ and -10.75‰ to -7.86‰, respectively. The $\delta^{17}O$ values ranged from -5.60‰ to -2.79‰, averaging

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

-5.01‰, while the d-excess varied from 0.01‰ to 13.69‰, with a mean of 6.89‰. During the wet season, the mean groundwater δD and $\delta^{18}O$ were -74.43‰ and -9.99‰, with ranges of -76.84‰ to -70.91‰ and -10.70‰ to -8.73‰, respectively. The $\delta^{17}O$ values were between -5.72‰ and -4.72‰, with a mean of -5.28‰, and the d-excess ranged from -3.67‰ to 9.80‰, averaging 5.50‰. The d-excess during the dry season was the highest among the three periods, while it was the lowest during the normal period, indicating significant variations in d-excess across different seasons. The $\delta^{17}O$ values in the dry season were higher than those in the wet and normal periods, which is a direct reflection of the impact of precipitation variations on the isotopic composition of the water body.

The isotopic values of precipitation δD and $\delta^{18}O$ ranged from -97.78‰ to -20.22‰ and from -13.48‰ to -1.96‰, with mean values of -55.36‰ and -7.60‰, respectively. Overall, the $\delta^{18}O$ and δD values of precipitation in the study area fall within the global ranges of -50‰ to 10‰ and -350‰ to 50‰. The Local Meteoric Water Line (LMWL) for the study area is defined by the equation: $\delta D = 6.2\ \delta^{18}O - 8.2$ (Fig.8). Specifically, the equations for the normal, dry, and wet seasons are: $\delta D = 1.07\ \delta^{18}O - 53.50$, $\delta D = 4.20\ \delta^{18}O - 30.09$, and $\delta D = 1.95\ \delta^{18}O - 54.97$, respectively. The slope of the annual LMWL is lower than that of the Global Meteoric Water Line (GMWL) proposed by Craig in 1964 ($\delta D = 8\ \delta^{18}O + 10$), as well as lower than the China Meteoric Water Line (CMWL) ($\delta D = 7.9\ \delta^{18}O + 8.2$). The stable hydrogen and oxygen isotopic characteristics of groundwater samples from the three periods all exhibit a discrete, linear distribution and plot below both GMWL and LMWL. This phenomenon reveals that the water isotopes have undergone strong fractionation during evaporation in the normal, wet, and dry seasons. Furthermore, the stable hydrogen and oxygen isotope data for the dry and normal seasons are mainly concentrated in the lower-left region of the plot, indicating relative isotopic depletion during these two periods. During the normal period, the δD and $\delta^{18}O$ values exhibit a high degree of dispersion and are widely distributed in the upper-central part of the scatter plot. This reflects that the stable hydrogen and oxygen isotopes are relatively enriched and have a wide range of variation during the normal season.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

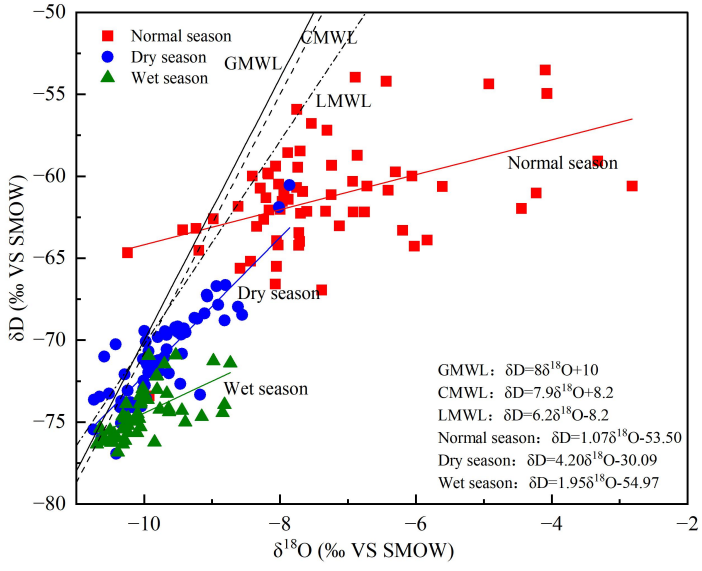


Fig.8. $\delta^{18}O/\delta D$ relationship of groundwater samples.

3.2.2 Identification of nitrate sources using isotopes and MixSIAR model

During the normal water period, the nitrogen and oxygen isotopic compositions in groundwater exhibit a wide range of variation. The $\delta^{15}N\text{-}NO_3^-$ values range from 5.6‰ to 24.52‰ (average: 18.22‰), while the $\delta^{18}O\text{-}NO_3^-$ values range from -6.33‰ to 6.23‰ (average: 0.22‰). In the low water period, the range of $\delta^{15}N\text{-}NO_3^-$ values expands to 3.2‰-21.96‰ (average: 12.19‰), and the $\delta^{18}O\text{-}NO_3^-$ values range from -9.58‰ to 8.04‰ (average: 0.65‰). Previous studies have established characteristic $\delta^{18}O\text{-}NO_3^-$ ranges for different nitrate sources: atmospheric deposition (23‰-75‰), nitrate fertilizers (18‰-24‰), and products of nitrification (-10‰-10‰). The data points are predominantly concentrated within the zone of animal manure and domestic wastewater, indicating that nitrate is primarily derived from these sources, with soil nitrogen as a secondary contributor.

The MixSIAR model was employed to quantitatively apportion the sources of groundwater nitrate nitrogen. According to the average contributions from each source, the five pollution sources in the study area were ranked as follows: DSM (74.1%) > SON (20.9%) > NHF (4.2%) >

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

NOF (0.6%) > NP (0.2%) (Fig.9). This indicates that the primary contributor to groundwater NO_3^--N in the study area was manure and sewage, followed by soil nitrogen. The influences of atmospheric precipitation and chemical fertilizers on groundwater nitrate were negligible. The quantitative results from the MixSIAR analysis are consistent with the qualitative findings, confirming that manure and sewage, along with soil nitrogen, are the dominant sources of nitrate pollution in the study area.

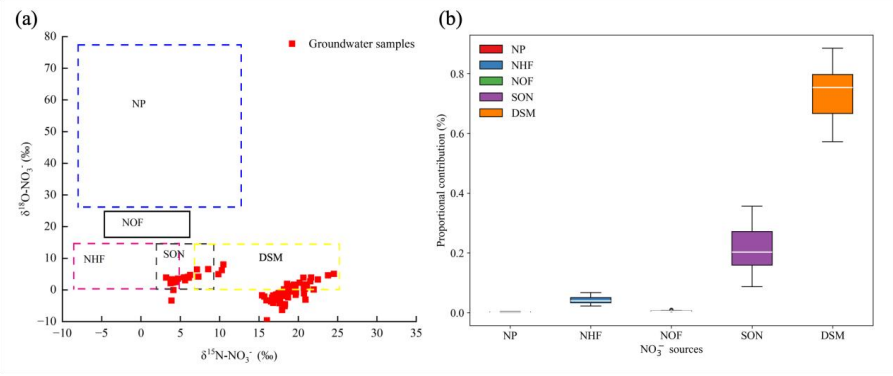


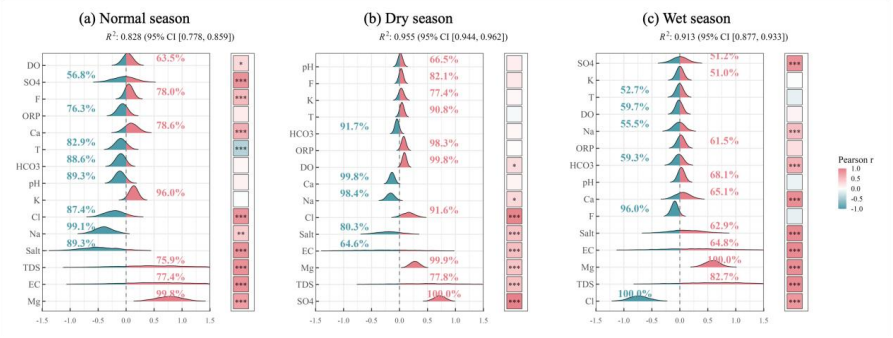
Fig.9. (a) distributions of $\delta^{15}N\text{-}NO_3^-$ and $\delta^{18}O\text{-}NO_3^-$ values in the study area. (b) proportional contributions of the main NO_3^- sources evaluated by the MixSIAR model. Note: boxplots denote the 25th, 50th and 75th percentiles.

3.3 Bayesian model analysis and correlation analysis

During the normal season, Bayesian model indicated the central role of Mg^{2+}, which is consistent with its strong positive correlation (r=0.75) in the correlation matrix (Fig.10). Na^+ exhibited a significant negative effect, whereas it only showed a weak positive correlation (r=0.39) in the correlation matrix. This suggests that variations in Na^+ concentration are more reflective of hydrological processes, such as evaporative concentration, rather than direct involvement in the chemical transformation of NO_3^-. Although the correlation matrix revealed strong correlations between NO_3^- and both TDS and EC (r > 0.8), their respective probabilities of direction (pd) in the Bayesian model were both below 80%. This further confirms that their influence is primarily manifested indirectly through collinearity with other ions. In the dry season, Bayesian model identified SO_4^{2-} as the primary positive driver of NO_3^-, a finding that is in strong agreement with the high positive correlation (r=0.96) between SO_4^{2-} and NO_3^- observed in the correlation matrix.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

642 Concurrently, Bayesian model indicated significant negative effects for both Na^+ and Ca^{2+}, which
643 contrasts with their weak positive correlations with NO_3^- in the correlation matrix. This
644 discrepancy likely arises because the elevated concentrations of Na^+ and Ca^{2+} are attributed to
645 evaporative concentration, whereas the increase in NO_3^- stems from anthropogenic inputs,
646 indicating no direct causal relationship between them. During the wet season, Bayesian model
647 identified Cl^- and Mg^{2+} as the most critical driving factors, with clear directional effects and high
648 confidence. This aligns with the trends observed in the correlation matrix, where NO_3^- correlated
649 negatively with Cl^- (r=-0.78) and positively with Mg^{2+} (r=0.84), thereby validating their direct
650 influence on NO_3^- concentrations during the wet season. While the correlation matrix also showed
651 high positive correlations between NO_3^- and both TDS and EC, their posterior distributions in the
652 Bayesian model were wider and their pd values were lower. This suggests that their impact is
653 likely a result of high collinearity with key variables such as Mg^{2+} and Cl^-, rather than an
654 independent effect.



655

656 **Fig.10**. Factor effects and Pearson coefficients of physicochemical variables on NO_3^- at different
657 periods. The left part of each subgraph shows the relative importance and posterior distribution of
658 each environmental variable to NO_3^- after the Bayesian model operation. The red area represents
659 the probability density of the positive effect, and the blue area represents the probability density of
660 the negative effect. The percentage values beside the distribution represent the Probability of
661 Direction (pd). The right part of each subgraph is the heat map of the correlation analysis.

662

663 3.4 Model performance evaluation

664 3.4.1 Virtual data analysis

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

To address the modeling bias arising from limited measured samples, this study constructed virtual datasets at 1-10 times the original scale based on a strategy combining t-SNE dimensionality reduction, GMM clustering sampling, and KNN inverse mapping, to enhance the robustness of model training. Taking the 10x virtual dataset as an example, the statistical characteristics (Table 3) show that the virtual samples effectively reproduced the central tendency and dispersion of the original data. For the normal season, the mean NO_3^- concentration was 30.41 mg L^{-1} (vs. observed mean of 33.67), with a standard deviation of 28.78 (vs. 35.83) and a coefficient of variation (CV) of 0.95 (vs. 1.06). In the dry season, the maximum value of the virtual samples reached 178.09 mg L^{-1}, while this did not fully replicate the extreme high values (observed maximum of 358.58 mg L^{-1}), it effectively expanded the range of the heavy-tailed distribution. For the wet season, although the CV for all indicators was slightly lower than the measured values, their ranges (8.47-80.37 vs. 4.15-98.36 mg L^{-1}) still showed a high degree of overlap, indicating that no systematic distortion was introduced.

Table 3. Statistical characteristics of different virtual samples.

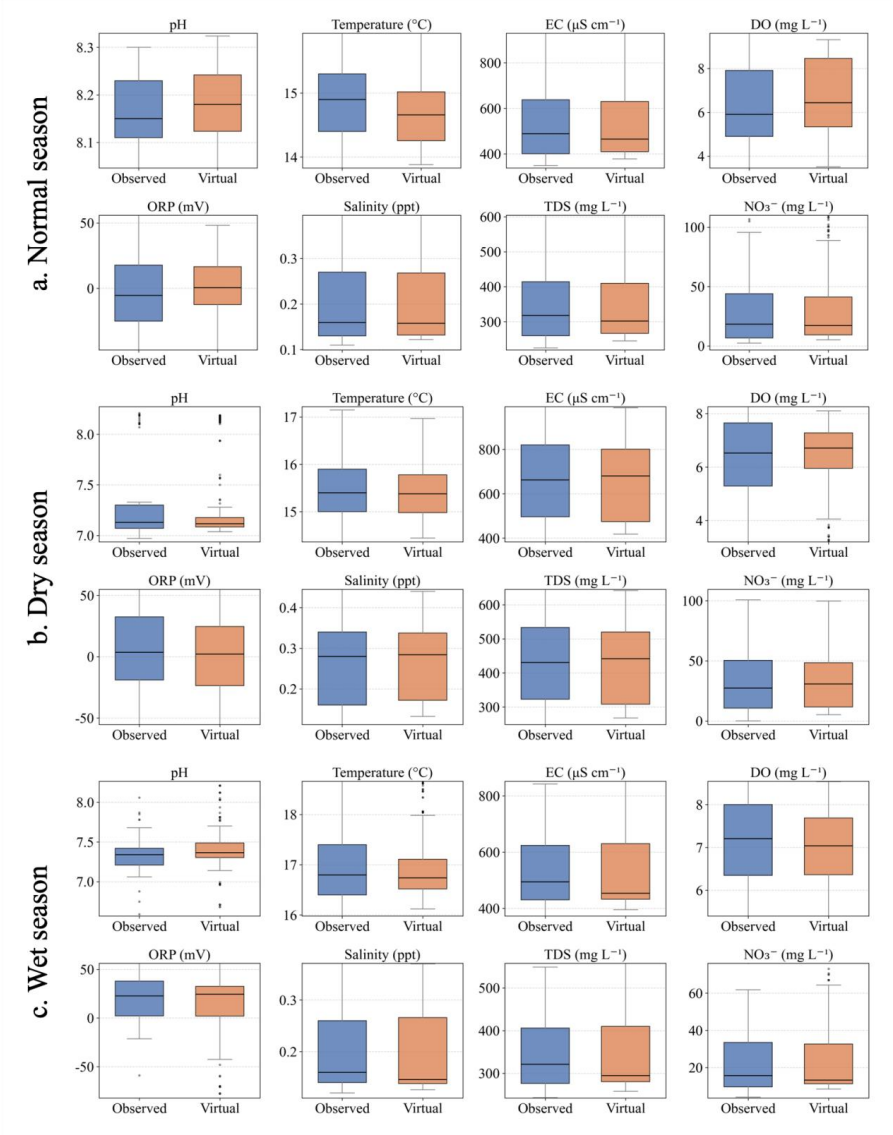
Periods		pH	T	EC	DO	ORP	Salt	TDS	NO_3^-
	Unit		℃	μs cm^{-1}	mg L^{-1}	mv	ppt	mg L^{-1}	mg L^{-1}
Normal season	Max	8.32	16.2	963.6	9.31	101.02	0.42	628.6	124.03
n=660	Min	7.95	13.88	378.2	3.52	-48.28	0.12	245	5.10
	Mean	8.18	14.72	527.31	6.66	2.09	0.20	342.72	30.41
	SD	0.07	0.54	148.31	1.66	25.97	0.08	96.92	28.78
	CV	0.01	0.04	0.28	0.25	12.44	0.41	0.28	0.95
Dry season	Max	8.19	17.59	987.8	8.10	69.18	0.44	641.8	178.09
n=650	Min	7.04	14.44	417.6	2.68	-59.14	0.132	267.8	5.19
	Mean	7.35	15.49	661.79	6.414	0.52	0.27	429.90	37.75
	SD	0.44	0.69	170.65	1.24	29.61	0.09	111.16	33.13
	CV	0.06	0.04	0.26	0.19	57.21	0.34	0.26	0.88
Wet season	Max	8.21	18.88	872.8	8.54	56.4	0.37	569	80.37
n=500	Min	6.31	16.12	395.8	5.34	-77.64	0.13	257.6	8.47
	Mean	7.38	16.93	535.7	7.06	13.75	0.20	348.24	25.85

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

SD	0.33	0.62	131.02	0.78	28.85	0.08	86.05	19.88
CV	0.04	0.04	0.24	0.11	2.10	0.38	0.25	0.77

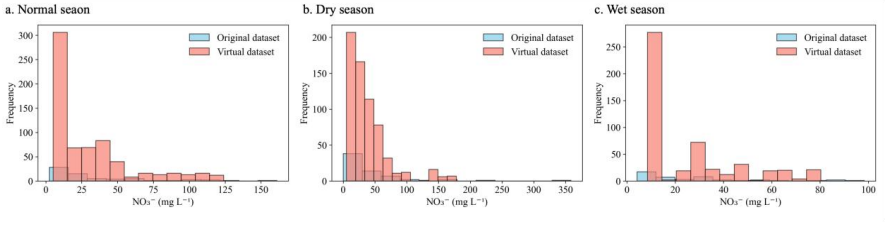
Standardized multivariate boxplots (Fig.11) visually confirm that the median, interquartile range (IQR), whisker length, and outlier distribution of the virtual data for each period were highly similar to the measured data, demonstrating that the central tendency and dispersion characteristics were well-preserved. Hydrological seasonal characteristics, such as high EC/TDS/Cl^-/NO_3^- in the dry season and low, concentrated NO_3^- in the wet season, were also accurately preserved. Although the number of some newly added outliers slightly increased, they all fell within physically reasonable ranges, with no non-physical solutions, such as negative concentrations or out-of-bounds pH values, occurring. Fig.12 presents a comparison of nitrate concentration frequency distributions between the original and synthetic datasets across normal, dry, and wet periods. The distributional comparison indicates that the proposed t-SNE + GMM + KNN inverse mapping synthetic sample generation strategy maintains the core features of the NO_3^- distribution for each hydrological period, while simultaneously improving sample representation in sparse areas and intervals of high variability. Therefore, the t-SNE + GMM method effectively captured the non-linear structure and extreme value information of the original data, and can provide reliable data support for subsequent model training.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026



694

695 **Fig.11**. Box plots of the observed and virtual variable data at different periods.



696

697 **Fig.12**. Comparison of nitrate concentration distribution in the original and virtual datasets under

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

different periods.

3.4.2 Prediction based on on-site measured water quality data

During the normal season, the R^2 values for the baseline Random Forest and the quantum-enhanced RF models were 0.673 and 0.660, respectively, indicating high prediction errors (Table S1). As the number of virtual samples was increased from 1-fold to 10-fold the size of the original dataset, the R^2 of both models steadily improved to above 0.958, with the quantum-enhanced RF model performing better and ultimately achieving an R^2 of 0.9622. When the number of virtual samples exceeded 500, the performance gains began to plateau. For the dry season, the modeling performance with the original data was the poorest, which is correlated with the high variability of NO_3^- concentrations during this period (Table S2). This suggests that with a limited sample size, models are susceptible to interference from outliers, and a small number of measured samples is insufficient to support effective model learning. After introducing virtual samples, the model performance improved significantly. A mere 1-fold augmentation of the sample size increased the R^2 to 0.527 (RF). When augmented to 8-fold, the R^2 reached 0.854. Although the quantum-enhanced model slightly underperformed the classical RF in the initial stages (≤2-fold augmentation), their performances converged at higher augmentation levels, both achieving high accuracy. This indicates that virtual samples effectively mitigated the modeling challenges posed by data sparsity and skewed distributions. In contrast, the modeling performance with original data was optimal during the wet season, attributed to the generally lower NO_3^- concentrations and their smaller spatial variability (Table S3). The use of virtual samples further elevated the prediction accuracy to an exceptionally high level. A 4-fold augmentation yielded an R^2 of 0.962. After augmentation to 10-fold, the R^2 of the RF model stabilized at 0.977, with the RMSE dropping to as low as 3.03 mg L^{-1}. The overall performance of the quantum-enhanced RF was consistent with the classical RF, with only slight fluctuations within a very small error range, showing that when data quality is high and the relationships are more linear, the marginal gains from quantum feature encoding are limited.

During the normal season, as illustrated in Fig. 13(A1)-(A2), a deviation was observed between the predicted and observed values for both models when utilizing only the 66 original samples. The prediction results exhibited high dispersion, and the median deviated markedly from

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

the observed median. This is consistent with the low R^2 values, indicating errors inherent in small-sample modeling. With an increase in the number of virtual samples, the distribution of predicted values gradually converged towards the observed values, and the interquartile range (IQR) and whiskers of the boxplots progressively narrowed, indicating a substantial enhancement in model stability and accuracy. When the virtual samples were expanded tenfold (to 660 virtual samples), the boxplots of the predicted values highly overlapped with those of the observed values, consistent with the reduction of the RMSE to 6.02 mg L^{-1}. In the final stage, the quantum-enhanced model slightly outperformed the classical RF model, achieving an R^2 of 0.9622. In the dry season (Fig. 13(B1)-(B2)), on the original dataset, the predicted values from both models were generally overestimated due to the extremely high and skewed distribution of NO_3^- concentrations. Consequently, the predicted boxplots were positioned entirely above the actual values, yielding an R^2 of only 0.28. The introduction of virtual samples led to a substantial improvement in model performance. Starting from single-fold augmentation, the median and range of the predicted boxplots began to converge towards the observed values, at augmentation levels of 8-fold and higher, the predicted values effectively captured the distributional characteristics of the high-concentration intervals. Although the classical RF model slightly outperformed the quantum-enhanced model at low augmentation levels, their performances converged as the sample size further increased. This demonstrates that the virtual sample generation strategy effectively alleviates modeling challenges caused by data sparsity and extreme values. During the wet season (Fig. 13(C1)-(C2)), the predicted values of both models already exhibited good consistency with the observed values on the original dataset, with the predicted boxplots substantially overlapping the observed ones. With the incorporation of virtual samples, the prediction accuracy and stability of the models were further enhanced. The IQR and whiskers of the boxplots continued to narrow, and the predicted values became more concentrated within the true distribution range of the observed values. Following tenfold data augmentation, the agreement between predicted and observed values was exceptionally high, with an R^2 reaching 0.977 and an RMSE decreasing to 3.03 mg L^{-1}, demonstrating excellent predictive performance.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

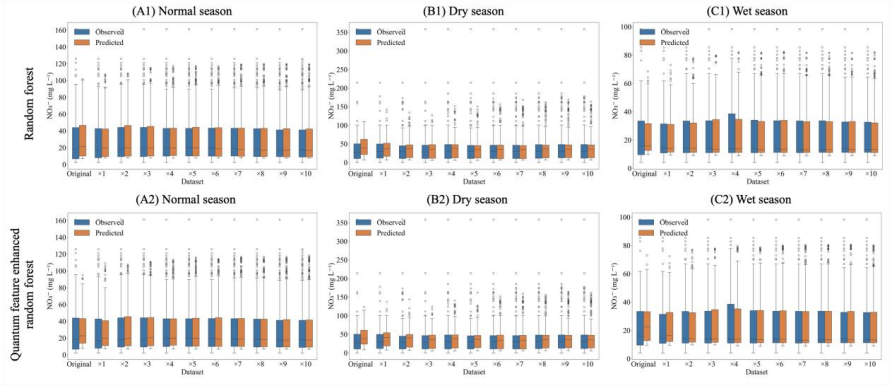


Fig.13. Comparison of observed and predicted NO_3^- concentrations across data augmentation levels for random forest and quantum feature-enhanced random forest models in normal, dry, and wet Seasons.

3.4.3 Prediction based on AlphaEarth Foundation Embeddings

To explore the potential of remote sensing semantic embedding features in predicting groundwater nitrate concentrations, this section employs the 64-dimensional surface semantic vectors derived from the Google AlphaEarth Foundation (AEF) dataset as model input variables. We reduced the variables through principal component analysis to preserve ⩾95% variance.

During the normal season, modeling based on original samples yielded poor performance. The R^2 for Random Forest and quantum-enhanced RF were 0.167 and 0.119, respectively, with RMSE values as high as 32.89 and 33.82 mg L^{-1} (Table S4). These results suggest that, given the limited sample size, relying solely on AEF embedding features is insufficient to fully capture the hydrological processes characteristic of this period. Model performance improved with the introduction of virtual samples. When the virtual samples were expanded to ten times the size of the original dataset, the R^2 of the RF model increased to 0.860, and the RMSE decreased to 10.73 mg L^{-1}. Similarly, the quantum-enhanced RF achieved an R^2 of 0.844, exhibiting a consistent overall trend. A comparison of boxplots (Fig. 14A) reveals that the initial predictions severely overestimated the low-to-medium concentration ranges while underestimating the high-value tails. As the sample size expanded, the predicted boxplots progressively converged toward the observed distribution. The agreement between the median and interquartile range (IQR) improved significantly, confirming that virtual samples effectively enhanced the capability of AEF features

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

to represent non-linear patterns.

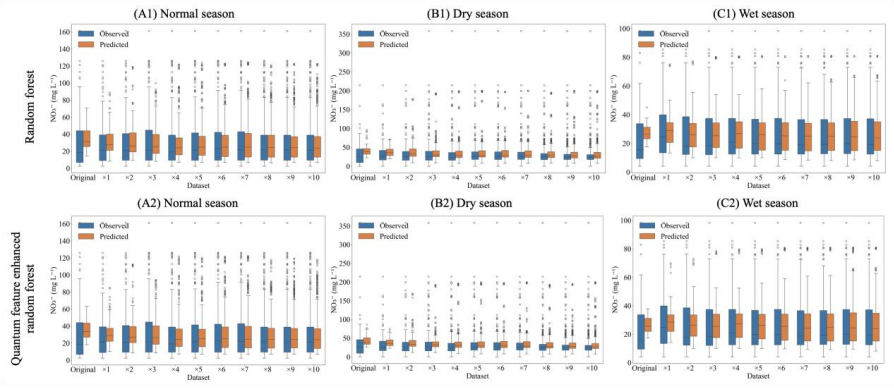


Fig.14. Comparison of observation and prediction of NO_3^- concentration by random forest and quantum featution-enhanced random forest models at data enhancement levels in normal, dry and wet seasons: based on AlphaEarth Foundation as the input variable.

In the dry season, modeling with the original dataset resulted in a negative R^2, reflecting the extremely weak generalization ability of AEF features in scenarios characterized by high variability and heavy-tailed distributions (Table S5). Following the introduction of a single-fold of virtual samples, the R^2 rose to 0.039. Upon an 8-fold expansion, the RF R^2 reached 0.641 (RMSE=21.05 mg L^{-1}), at a 10-fold expansion, it further improved to 0.674 (RMSE=19.86 mg L^{-1}). The quantum-enhanced RF slightly outperformed the standard RF at high expansion levels, indicating that quantum encoding offers certain advantages in mitigating the influence of extreme values and enhancing model robustness (Fig. 14B). The prediction distribution plots indicate that the initial model failed entirely to identify the high-concentration clustering characteristics of NO_3^- during the dry season. After data augmentation, the predicted boxplots progressively covered the true high-value intervals, and the trend of tail extension gradually aligned with the observed data.

In the wet season, although modeling with the original data still yielded a negative R^2, the performance improvement was the most rapid (Table S6). An R^2 of 0.5 was achieved with only a 2-fold expansion of virtual samples. At 5-fold expansion, it reached 0.685, and after a 10-fold expansion, the RF R^2 stabilized at 0.784 (RMSE=8.27 mg L^{-1}), while the quantum-enhanced RF reached 0.781. The boxplots show that the predicted values, initially severely dispersed and systematically biased, rapidly converged to the dense intervals of the observed values, with the final IQR and whisker ranges showing a high degree of overlap.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

Compared to modeling results based on in-situ observation data, the predictive performance based on AlphaEarth Foundation embedding features was generally lower. Under the same virtual sample augmentation multiplier, the maximum R^2 for the normal, dry, and wet seasons were approximately 10.27%, 17.37%, and 19.33% lower, respectively. This indicates that measured water quality parameters more directly reflect the key processes of nitrogen migration and transformation. However, given that AEF can be obtained globally without the need for field sampling, it offers a feasible alternative for the rapid screening of groundwater nitrate risks in large-scale unmonitored areas.

3.5 Feature importance analysis

Fig.15 illustrates the feature importance rankings of the RF and quantum-enhanced RF models when using in-situ measured water quality parameters as inputs across different hydrological seasons. The dominant predictive factors vary across different seasons, and the virtual sample augmentation strategy influences both the stability of feature importance and model performance. There are distinct differences in the key driving factors for each season, which aligns with the results of the Bayesian models and correlation analysis. In the normal season, TDS, EC, Salt, and DO are the most important predictive variables, with their importance significantly higher than that of other parameters. In the dry season, TDS, EC, Salt, and pH exhibit the highest importance. In the wet season, the importance of TDS, EC, Salt, and ORP is most prominent. With the increase in the number of virtual samples, the ranking of feature importance tends to stabilize. For instance, in the dry season, when the sample size increased from the original 65 to 715, the importance of TDS and EC continued to rise and eventually stabilized. Comparing the RF and quantum-enhanced models, quantum enhancement did not fundamentally alter the ranking of feature importance; however, it slightly increased the importance of certain variables or made them more stable, demonstrating the effectiveness of quantum feature encoding as a means of information enhancement.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

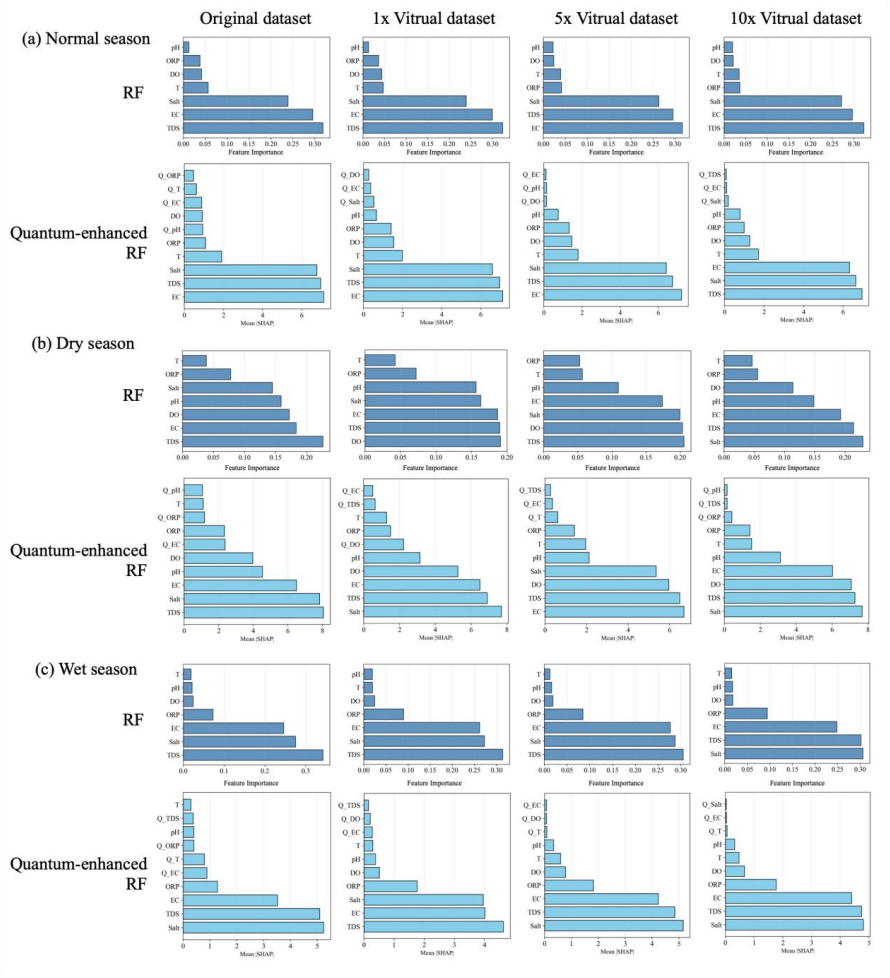


Fig.15. Input feature importance of classical and quantum-enhanced Random Forest in seasonal nitrate prediction under different data augmentation strategies: based on Gini index and SHAP Model.

Fig.16 presents the feature importance of the RF and quantum-enhanced RF models when using only the 64-dimensional AEF semantic embedding vectors as inputs. Since the AEF features themselves are highly abstract, we cannot assign them specific physical meanings; however, we can infer which remote sensing semantic information is critical for predicting nitrate concentration through their importance rankings. Compared to in-situ measured parameters, the importance of AEF features fluctuates significantly more across different seasons and data volumes, lacking a consistent core feature set. This reflects that although AEF embeddings contain rich

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

environmental semantic information, their direct correlation with groundwater nitrate concentration is relatively weak, necessitating the learning of large amounts of data to establish a robust mapping relationship. In the normal season, features such as A05, A07, and A00 exhibit relatively high importance. These features may encode seasonal information related to land use types, soil moisture, or vegetation cover. In the dry season, features such as A08, A06, and A05 are relatively more important. These features may be associated with surface dryness, bare surface area, or the intensity of human activity, correlating with the spatial distribution of high-concentration NO_3^- pollution sources during the dry season. In the wet season, the importance of features like A02, A03, and A05 is prominent. These features may be related to surface runoff, vegetation growth status, or soil water content, reflecting the driving effect of rainfall on pollutant migration during the wet season.

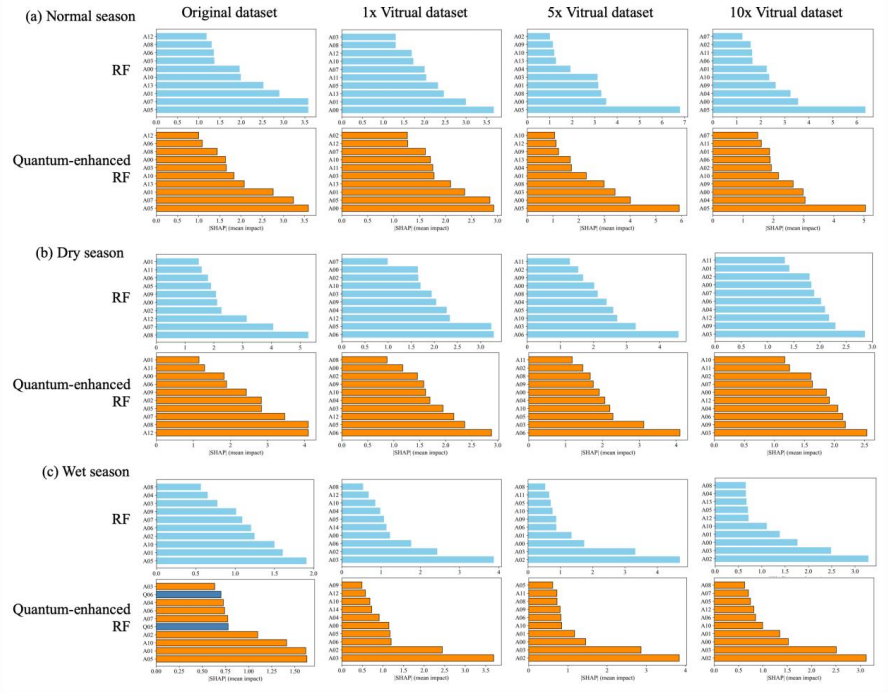


Fig.16. Importance of AEF input features in seasonal nitrate prediction using classical and quantum enhanced Random Forest under different data augmentation strategies.

The introduction of virtual samples is crucial for stabilizing the importance of AEF features. On the original dataset, the feature importance ranking was chaotic and unstable; as virtual samples increased, the ranking gradually became clearer, and the importance of certain core

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

features was highlighted. This once again demonstrates the effectiveness of the virtual sample generation strategy for small-sample modeling. When using AEF features, the feature importance distribution of the quantum-enhanced model is similar to that of the RF model, but it occasionally assigns slightly higher weights to certain features . This suggests that quantum feature encoding may assist the model in extracting more discriminative information from the high-dimensional, complex remote sensing semantic space, thereby slightly optimizing the feature selection process.

Fig.17 presents a local feature attribution analysis for representative samples predicting the highest and lowest NO_3^- concentrations using SHAP waterfall plots. Regardless of whether the classical RF or the quantum-enhanced RF model is used, samples predicting high NO_3^- concentrations are driven by a set of features with positive contributions (red bars). In the normal season, for the highest NO_3^- sample with a predicted value of 161.17 mg L^{-1}, features A05, A09, and A06 contributed the highest positive values, with A05 making the largest contribution and serving as the key factor driving the prediction to a high level. In the dry season, for the sample with a predicted value as high as 358.58 mg L^{-1}, features A12, A00, and A07 were the main positive driving factors, with A12 contributing most prominently. In the wet season, for the sample with a predicted value of 98.36 mg L^{-1}, features A03, A02, and A04 provided the main positive contributions, with A03 contributing the most. For samples predicting low NO_3^- concentrations, model decisions mainly rely on features with negative contributions (blue bars). The role of these features is to pull the predicted value down from the baseline (E[f(X)]). In the normal season, for the lowest NO_3^- sample with a predicted value of 2.39 mg L^{-1}, features A05, A04, and A00 exhibited strong negative contributions, with A05 showing the largest negative contribution. In the dry season, for the sample with a predicted value of only 0.10 mg L^{-1}, features A09, A12, and A06 were the main negative driving factors, with A09 contributing the most negatively. In the wet season, for the sample with a predicted value of 4.15 mg L^{-1}, features A03, A06, and A00 provided the main negative contributions, with A03 contributing the most negatively.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

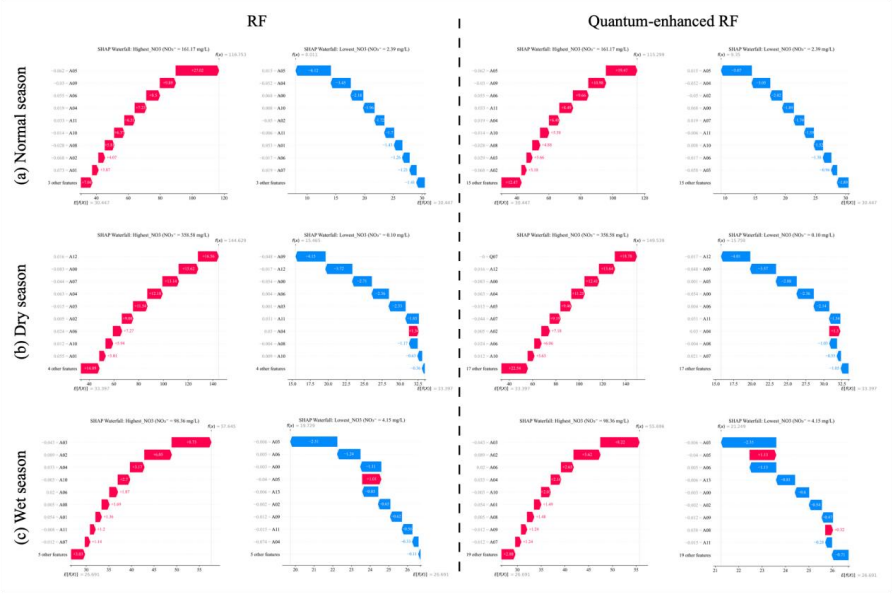


Fig.17. SHAP waterfall-based feature attribution comparison between classical and quantum-enhanced Random Forest across different seasons.

4. Discussion

4.1 Nitrogen sources, migration, and transformation

The Piper diagram indicates that the hydrochemical type is predominantly Ca-Mg-HCO_3^-. The Gibbs diagram and ion ratios confirm that the hydrochemical background is dominated by carbonate rock dissolution, with weak cation exchange. Concentrated precipitation during the rainy season leads to dilution and infiltration, reducing the NO_3^- concentration to 27.14 mg L^{-1}. This indicates that surface manure leaches into the groundwater with rainfall (Sun et al., 2024). During this period, cation exchange is enhanced, improving the aquifer's temporary retention capacity for NO3- (Wang et al., 2025). Isotopic evidence and MixSIAR source apportionment consistently indicate that the primary sources of current nitrate pollution are domestic sewage and manure (DSM, 74.1%) and soil organic nitrogen (SON, 20.9%), whereas the contributions from chemical fertilizers and precipitation are minimal. This suggests that in the study area, the direct leaching of fertilizer nitrogen is not the dominant pathway, rather,fertilizer nitrogen remains in the soil-vadose zone and enters groundwater through long-term water drive (Wang et al., 2025). Given that the vadose zone thickness in the North China Plain generally exceeds 10 m, the

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

currently elevated NO_3^- levels are more likely derived from historical fertilizer residues and the long-term infiltration of manure, particularly as the farmlands in the study area are situated near rural residential areas (Wu et al., 2024). The mean $\delta^{15}N\text{-}NO_3^-$ values range from 12.2‰ to 18.2‰, which far exceeds the typical range for chemical fertilizers but closely matches that of manure and soil organic nitrogen. This confirms that the nitrogen has undergone microbial mineralization and nitrification processes (Li et al., 2022). SON is converted to NO_3^- through ammonification followed by nitrification under aerobic conditions, while ammonium from DSM also enters the groundwater via nitrification (Liu et al., 2023).

The seasonal variation in nitrate concentrations is essentially driven by scarce precipitation and strong evaporation during the dry season. This leads to a decline in groundwater levels and a reduction in flow velocity, creating a positive migration potential gradient. Consequently, NO_3^- accumulates in the discharge areas along with the groundwater flow. Furthermore, cation exchange is inhibited, weakened Na^+ adsorption and relative Ca^{2+} depletion indicate a decrease in the aquifer's retention capacity for NO_3^-, making accumulation the dominant process (Ahmed et al., 2013). In contrast, concentrated precipitation during the wet season triggers rapid infiltration, raising groundwater levels and increasing flow velocity, during which cation exchange becomes active (Zhang et al., 2023). The groundwater is generally oxidizing, as evidenced by the extremely low concentrations of NO_2^- (<0.11 mg L^{-1}) and NH_4^+ (<0.16 mg L^{-1}). This indicates that the majority of the area is an oxidative environment conducive to the stable existence of NO_3^-. The $\delta^{18}O\text{-}NO_3^-$ values range from -9.58‰ to 8.04‰, falling within the typical nitrification interval. This excludes significant denitrification, confirming that the transformation process is dominated by nitrification while denitrification is limited (Zhang et al., 2025).

4.2 Virtual sample generation effectively mitigates small-sample bias and reveals the model's sensitive response to seasonal heterogeneity

Model overfitting and insufficient generalization resulting from small-sample data are prevalent challenges in the field of environmental forecasting (Zhu et al., 2023). The t-SNE-GMM-KNN virtual sample generation strategy proposed in this study demonstrates that the generated virtual samples are highly consistent with the original data in terms of statistical characteristics, such as mean, standard deviation, and coefficient of variation, and successfully

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

reproduce hydrochemical differences across different seasons. The substantial improvement in model performance following virtual sample expansion clearly confirms that data sparsity, rather than insufficient model capacity, is the core bottleneck in seasonal nitrate modeling (Saha et al., 2023). Furthermore, even with the incorporation of generated virtual samples, the magnitude of prediction performance gains exhibits seasonal divergence. During the dry season, characterized by highly right-skewed NO_3^- concentrations, the model benefits most significantly. With 10-fold expansion, the R^2 value surges from 0.28 to over 0.85. Conversely, in the wet season, although the absolute performance gain is smaller due to dominant dilution effects and low concentrations, excellent predictive accuracy is still achieved. This phenomenon aligns with fundamental hydrological principles: strong evaporation and concentration during the dry season intensify the spatial heterogeneity of pollutant accumulation and process nonlinearity, necessitating richer samples to characterize tail behaviors (Li et al., 2025). In contrast, the dilution effects caused by rainfall leaching during the wet season tend to homogenize the system, reducing its dependency on sample size (Bigler et al., 2024). The t-SNE–GMM–KNN strategy proposed in this study outperforms traditional oversampling methods (e.g., SMOTE) or deep generative models (e.g., VAE) in preserving multimodal structures and heavy-tailed covariance; the latter often ignore the manifold geometric properties of high-dimensional geochemical spaces or inherently rely on large amounts of training data (Udu et al., 2025), which is precisely what is lacking in the scenario of this study. Compared to common methods such as Gaussian Mixture Models (GMM) and Generative Adversarial Networks (GANs), the core advantages of this strategy are reflected in three aspects: first, t-SNE dimensionality reduction accurately captures sample clustering structures driven by different hydrological processes, providing a reliable foundation for subsequent distribution modeling; second, the number of GMM clusters is automatically optimized based on the Bayesian Information Criterion (BIC), avoiding biases arising from subjective settings; and third, KNN inverse mapping enables reconstruction from low-dimensional to high-dimensional space without the need for large-scale training data, making it more suitable for small-sample scenarios (Silva et al., 2023; Kurniawan et al., 2024; Peng et al., 2025).

4.3 Performance analysis of hybrid quantum-classical model

Quantum Machine Learning offers a novel approach to capturing complex non-linear

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

relationships through feature mapping in high-dimensional quantum Hilbert spaces. The hybrid quantum-classical Random Forest yields slight performance improvements in scenarios where original data is scarce or the distribution is highly skewed (Lamichhane et al., 2025). When classical feature representation capacity approaches saturation, the Z-feature mapping based on Parameterized Quantum Circuits (PQC) can expose entangled non-linear patterns in the Hilbert space, thereby enhancing feature discriminability. The gain from this enhancement tends to converge after sufficient virtual sample expansion. In this study, quantum features were generated by analytically calculating the Pauli-Z expectation value ⟨Z⟩, completely circumventing hardware noise interference associated with quantum sampling. This renders the quantum-enhanced RF practically feasible for small-sample environmental tasks. However, the performance improvement of the quantum-enhanced RF is not absolute; in scenarios with high data quality and significant linear relationships during the wet season, the marginal gain of quantum features is limited. Conversely, in the dry season, characterized by sparse data and numerous extreme values, quantum encoding demonstrates stronger stability by reducing measurement noise interference (Ranga et al., 2024). This phenomenon indicates that the advantages of hybrid quantum-classical modeling are concentrated in scenarios with data complexity and limited information. Its essence lies in expanding the model's representational capacity through quantum feature enhancement, rather than replacing the core logic of classical models. This exploration verifies the potential value of quantum machine learning in addressing small-sample problems in earth sciences, even if its absolute advantage may not be as pronounced as in pure quantum algorithms (Adhikari, 2022).

4.4 Potential and limitations of using AlphaEarth Foundation Embeddings for large-scale monitoring

Modeling performance using only AEF embeddings as input generally yields an R^2 approximately 10-20% lower than that achieved using measured water quality parameters. The core reason for this discrepancy lies in the fact that water quality parameters directly reflect the immediate state of the groundwater chemical environment and are directly related to nitrate transport and transformation processes, whereas surface remote sensing semantics provide only an indirect characterization (Alam et al., 2025). After 10-fold virtual expansion, the AEF model still achieves R^2 values of >0.67 in the dry season, >0.85 in the normal season, and >0.78 in the wet

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

season, proving its feasibility as a rapid large-scale screening tool, particularly in unmonitored areas. Seasonal shifts in feature importance (dominated by A05/A00 in the normal season, A08/A06 in the dry season, A02/A03 in the wet season) suggest potential physical interpretations. A05/A00 may encode crop residue or soil organic matter information, A08/A06 may characterize the degree of bare soil exposure, and A02/A03 may reflect vegetation growth status or surface runoff potential. These inferences align highly with MixSIAR source apportionment and Bayesian driving factors. Although causal inference remains indirect, the global coverage and annual update characteristics of AEF make it a powerful supplement rather than a substitute for large-scale monitoring.

5. Conclusions

This study develops an integrated prediction framework combining hybrid quantum-classical machine learning, advanced virtual sample augmentation (t-SNE–GMM–KNN), and remote sensing foundation model embedding (AlphaEarth Foundation, AEF). The framework is designed to systematically address three core challenges in predicting groundwater nitrate concentrations in agricultural areas across different hydrological seasons: small sample bias, seasonal heterogeneity, and input data scarcity.

Hydrological seasonality acts as the dominant controlling factor for the spatiotemporal variability of nitrates. Nitrate concentrations peak during the dry season (mean: 42.93 mg L^{-1}), driven primarily by evaporative concentration and pollutant accumulation effects. In contrast, concentrations reach a minimum in the wet season (mean: 27.14 mg L^{-1}) due to dilution by precipitation. The groundwater hydrochemical type is consistently Ca-Mg-HCO_3^- across all seasons, controlled predominantly by carbonate mineral dissolution. TDS, EC, and salinity remain consistently top-ranked across all seasons, with additional season-specific drivers including Mg^{2+} and Na^+ (normal season), SO_4^{2-} (dry season), and Cl^- (wet season). Stable hydrogen and oxygen isotope analysis reveals strong evaporative fractionation of groundwater. MixSIAR analysis quantitatively apportioned nitrate sources: domestic sewage and manure (DSM) contribute 74.1%, soil organic nitrogen (SON) 20.9%, while synthetic fertilizers (NHF+NOF=4.8%) and atmospheric deposition (0.2%) are negligible, strongly indicating that legacy nitrogen stored in the thick vadose zone, rather than in-season fertilizer leaching, sustains current pollution.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

The proposed t-SNE-GMM-KNN virtual sample strategy effectively alleviates the bottleneck associated with small-sample modeling. By preserving the nonlinear manifold structure and multimodal distribution characteristics of the high-dimensional hydrochemical space, this method significantly enhances the model's ability to fit heavy-tailed distributions. Model performance improves significantly with virtual sample expansion. Using measured parameters as inputs, a 10-fold augmentation increased the coefficient of determination (R^2) for the dry season from 0.284 to >0.85, while stabilizing it at >0.95 for the normal and wet seasons. This confirms that data sparsity is the fundamental constraint limiting performance. Although performance gains are limited with high-quality data, the quantum-enhanced Random Forest demonstrates superior stability compared to classical models in small-sample, highly skewed scenarios, validating the feasibility and value of quantum feature enhancement strategies in environmental small-sample learning. The overall prediction performance using measured hydrochemical parameters surpasses that of AEF remote sensing semantic embeddings (R^2 is approximately 10-20% higher), as the former directly reflects subsurface nitrogen migration and transformation processes. Following 10-fold virtual sample augmentation, the AEF model also achieves usable accuracy, with feature importance exhibiting seasonal shifts.

Acknowledgement

This work is supported by the Open Project Program of Engineering Research Center of Groundwater Pollution Control and Remediation, Ministry of Education of China (GW202212).

Author contributions

Junjie Xu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing (original draft preparation), Writing (review and editing). Xin Wei: Conceptualization, Validation, Formal analysis, Investigation, Resources, Visualization, Writing (original draft preparation). Yilei Yu: Conceptualization, Funding acquisition, Project administration, Supervision, Writing-review and editing. Lihu Yang and Xianfang Song: Investigation, Resources, Supervision. Yuanzheng Zhai: Funding acquisition. CuiCui Lv: Investigation, Supervision, Project administration.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

Competing interests

The contact author has declared that none of the authors has any competing interests.

References

Wang, S., Chen, J., Zhang, S., et al., 2023. Hydrochemical evolution characteristics, controlling factors, and high nitrate hazards of shallow groundwater in a typical agricultural area of Nansi Lake Basin, North China[J]. Environmental Research, 223: 115430.

Li, S.P., Li, W.P., Yin, X.L., et al., 2019. Distribution and evolution characteristics of national groundwater quality from 2013 to 2017[J]. Hydrogeology & Engineering Geology, 2019(06). (in Chinese)

Zhang, M., Zhi, Y.Y., Shi, J.C.,et al., 2018.Apportionment and uncertainty analysis of nitrate sources based on the dual isotope approach and a Bayesian isotope mixing model at the watershed scale[J].Science of the Total Environment,639:1175-1187.

Gu, B.J., Ge, Y., Chang, S.X., et al., 2013. Nitrate in groundwater of China: Sources and driving forces[J]. Global Environmental Change, 23(5):1112-1121.

Han, D.M., Currell, M.J., Cao, G.L., 2016. Deep challenges for China's war on water pollution[J]. Environmental Pollution, 218:1222-1233.

Wang, S.Q., Zheng, W.B., Kong, X.L., 2018. Spatial distribution characteristics of nitrate in shallow groundwater of the agricultural area of the North China Plain[J]. Chinese Journal of Eco-Agriculture, 26(10):1476-1482.

Cox, S.E., Huffman, R.L., Olsen, T.D., et al., 2016. Concentration of nitrate and other water-quality constituents in groundwater from the water table beneath forage fields receiving seasonal applications of dairy manure, Whatcom County, Washington (2015)[J]. US Geological Survey (USGS) Data Release, 2016: 368.

Sebestyen, S.D., Shanley, J.B., Boyer, E.W., et al, 2014. Coupled hydrological and biogeochemical processes controlling variability of nitrogen species in streamflow during autumn in an upland forest[J]. Water Resources Research, 50(2): 1569-1591.

Thunyawatcharakul, P., Cho, K.H., Chotpantarat, S., 2025. Predicting Arsenic Speciation in Coastal Aquifers Using Machine Learning: A Case Study of the Chonburi and Rayong

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

Groundwater Basins, Thailand[J]. ACS ES&T Water, 5(9): 5011-5024.

Zhu, G., Shi, H., Zhong, L., et al., 2025. Nitrous oxide sources, mechanisms and mitigation[J]. Nature Reviews Earth & Environment, 6(9): 574-592.

Gao, H., Yang, L., Song, X., et al., 2023. Sources and hydrogeochemical processes of groundwater under multiple water source recharge condition[J]. Science of the Total Environment, 903: 166660.

Deng, Y., Ye, X., Du, X., 2023. Predictive modeling and analysis of key drivers of groundwater nitrate pollution based on machine learning[J]. Journal of Hydrology, 624: 129934.

Viswanathan, H.S., Ajo-Franklin, J., Birkholzer, J.T., et al., 2022. From fluid flow to coupled processes in fractured rock: Recent advances and new frontiers[J]. Reviews of Geophysics, 60(1): e2021RG000744.

Cai, C., Zhao, H., Zhang, H., et al., 2025. Timely assessment of maize lodging severity with limited samples using multi-temporal Sentinel-1 and Sentinel-2 data across large spatial extents[J]. Computers and Electronics in Agriculture, 237: 110671.

Feng, D., Liu, J., Lawson, K., et al., 2022. Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy[J]. Water Resources Research, 58(10): e2022WR032404.

Chen, Q., Yang, H., Cui, R., et al., 2025. Shallow groundwater table fluctuations: A driving force for accelerating the migration and transformation of phosphorus in cropland soil[J]. Water Research, 275: 123209.

Liu, X., Yue, F.J., Li, L., et al., 2025. Chronic nitrogen legacy in the aquifers of China[J]. Communications Earth & Environment, 6(1): 58.

Wu, Y., Wang, J., Liu, Z., et al., 2025. Seasonal nitrate input drives the spatiotemporal variability of regional surface water-groundwater interactions, nitrate sources and transformations[J]. Journal of Hydrology, 655: 132973.

Luo, Y., Yan, J., McClure, S.C., et al., 2022. Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model[J]. Environmental Science and Pollution Research, 29(22): 33205-33217.

Wang, C., Yang, J., Zhang, B., 2024. A fault diagnosis method using improved prototypical network and weighting similarity-Manhattan distance with insufficient noisy data[J].

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

Measurement, 226: 114171.

Tang, W., Carey, S.K., 2022. Classifying annual daily hydrographs in Western North America using t-distributed stochastic neighbour embedding[J]. Hydrological Processes, 36(1): e14473.

Tung, P.Y., Sheikh, H.A., Ball, M., et al, 2023. SIGMA: Spectral interpretation using gaussian mixtures and autoencoder[J]. Geochemistry, Geophysics, Geosystems, 24(1): e2022GC010530.

Farnia, F., Wang, W.W., Das, S., et al., 2023. Gat–gmm: Generative adversarial training for gaussian mixture models[J]. SIAM Journal on Mathematics of Data Science, 5(1): 122-146.

Zhou, C., Huang, Q., Cui, M., et al., 2025. Advances of machine learning in stable isotope geochemistry[J]. Journal of Analytical Atomic Spectrometry.

Wang, N., Zhou, Q., Gao, J., et al., 2025. Evaluating the efficacy of PCA and t-SNE in optimizing input features for groundwater level simulation using machine learning models[J]. Environmental Earth Sciences, 84(12): 336.

Tollefson, J., 2025. Google AI model creates maps of Earth 'at any place and time'[J], 2025, 644: 313.

Li, X., Yu, L., 2025. Mapping Complex Cropping Patterns in China (2018–2021) at 10 m Resolution: A Data-Driven Framework based on Multi-Product Integration and Google Satellite Embedding[J]. Earth System Science Data Discussions, 2025: 1-40.

Hong, Y.Y., Lopez, D.J.D., 2025. A Review on Quantum Machine Learning in Applied Systems and Engineering[J]. IEEE Access.

Oliveira, Santos.V., Costa, Rocha.P.A., Thé, J.V.G., et al., 2025. Optimizing the Architecture of a Quantum–Classical Hybrid Machine Learning Model for Forecasting Ozone Concentrations: Air Quality Management Tool for Houston, Texas[J]. Atmosphere, 16(3): 255.

Gujju, Y., Matsuo, A., Raymond, R., 2024. Quantum machine learning on near-term quantum devices: Current state of supervised and unsupervised techniques for real-world applications[J]. Physical Review Applied, 21(6): 067001.

Tian, D., Zhao, X., Gao, L., et al., 2025. A framework for tracing the sources of nitrate in surface water through remote sensing data coupled with machine learning[J]. npj Clean Water, 8(1): 43.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

1140 Alam, S.M.K., Li, P., Rahman, M., et al., 2025. Key factors affecting groundwater nitrate levels in
1141 the Yinchuan Region, Northwest China: Research using the eXtreme Gradient Boosting
1142 (XGBoost) model with the SHapley Additive exPlanations (SHAP) method[J].
1143 Environmental Pollution, 364: 125336.
1144 Liu, M., Geng, D., Wu, L., et al., 2025. The impact of agricultural land use change on water and
1145 nitrate fluxes in the deep vadose zone, the North China Plain[J]. Journal of Hydrology:
1146 Regional Studies, 2025, 62: 102914.
1147 Xu, G., Su, X., Yuan, Z., et al., 2022. Nitrogen behavior during artificial groundwater recharge
1148 through ponds: A case study in Xiong'an New Area[J]. Environmental Geochemistry and
1149 Health, 44(8): 2545-2561.
1150 Xiong'an New Area Official Website, 2023. "Geographical Environment and Climatic
1151 Characteristics of Xiong'an." *Xiong'an New Area Official Website*,
1152 http://www.xiongan.gov.cn/2023-02/27/c_129769113.htm. Accessed 16 Dec, 2025.
1153 Liao, Y.M., Huang, D.P., 2020. Climate Characteristics and Change Trend of Xiongan New Area.
1154 Chinese Agricultural Science Bulletin, 36(23): 99-105.
1155 Bai H, Yang H F, Meng R F, et al., 2023. Chemical characteristics and evolution of groundwater
1156 in Baoding Plain[J]. Geol. Rev., 69(06): 2216-2228.
1157 Li, X., Liu, M., Min, L., et al., 2025. Nitrogen transport and transformation processes in the
1158 typical deep vadose zone in the central North China Plain[J]. Chinese Journal of
1159 Eco-Agriculture.
1160 Zhang, Z.J., 2007. Investigation and evaluation of groundwater sustainable utilization in North
1161 China Plain[M].Shijiazhuang:Institute of Hydrogeology and Environmental Geology,Chinese
1162 Academy of Geological Sciences.
1163 Li, J., Zhu, D., Zhang, S., et al., 2022. Application of the hydrochemistry, stable isotopes and
1164 MixSIAR model to identify nitrate sources and transformations in surface water and
1165 groundwater of an intensive agricultural karst wetland in Guilin, China[J]. Ecotoxicology and
1166 Environmental Safety, 231: 113205.
1167 Huang, P., Chen, J., 2012. Recharge sources and hydrogeochemical evolution of groundwater in
1168 the coal-mining district of Jiaozuo, China[J]. Hydrogeology Journal, 20(4): 739-754.
1169 Kim, J.H., Kang, Y.J., Kim, K.I., et al., 2019. Toxic effects of nitrogenous compounds (ammonia,

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

nitrite, and nitrate) on acute toxicity and antioxidant responses of juvenile olive flounder, Paralichthys olivaceus[J]. Environmental toxicology and pharmacology, 67: 73-78.

Sun, J., Jia, Q., Li, Y., et al., 2022. Effects of arbuscular mycorrhizal fungi and biochar on growth, nutrient absorption, and physiological properties of maize (Zea mays L.)[J]. Journal of Fungi, 8(12): 1275.

Hamidi, M.D., Gröcke, D.R., Joshi, S.K., et al., 2023. Investigating groundwater recharge using hydrogen and oxygen stable isotopes in Kabul city, a semi-arid region[J]. Journal of Hydrology, 626: 130187.

Stock, B.C., Jackson, A.L., Ward, E.J., et al., 2018. Analyzing mixing systems using a new generation of Bayesian tracer mixing models[J]. PeerJ, 6: e5096.

Mao, H., Wang, G., Liao, F., et al., 2023. Spatial variability of source contributions to nitrate in regional groundwater based on the positive matrix factorization and Bayesian model[J]. Journal of Hazardous Materials, 445: 130569.

Gao, H, Yang, L., Song, X., et al., 2023. Sources and hydrogeochemical processes of groundwater under multiple water source recharge condition[J]. Science of the Total Environment, 903: 166660.

Torres-Martínez, J.A., Mora, A., Mahlknecht, J., et al., 2021. Estimation of nitrate pollution sources and transformations in groundwater of an intensive livestock-agricultural area (Comarca Lagunera), combining major ions, stable isotopes and MixSIAR model[J]. Environmental pollution, 269: 115445.

Jamshidi, E.J, Yusup, Y., Kayode, J.S., et al., 2022. Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature[J]. Ecological Informatics, 69: 101672.

Islam, M.T., Zhou, Z., Ren, H., et al., 2023. Revealing hidden patterns in deep neural network feature space continuum via manifold learning[J]. Nature Communications, 14(1): 8506.

Liu, H., Yang, J., Ye, M., et al., 2021. Using t-distributed Stochastic Neighbor Embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data[J]. Journal of Hydrology, 597: 126146.

Jia, B., Zhou, J., Tang, Z., et al., 2022. Effective stochastic streamflow simulation method based on Gaussian mixture model[J]. Journal of Hydrology, 605: 127366.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

Yan, R., Huang, J.J., 2023. Confident learning-based Gaussian mixture model for leakage detection in water distribution networks[J]. Water Research, 247: 120773.

Ghodba, A., Richelle, A., McCready, C., et al., 2025. A novel dynamic flux balance analysis for modeling CHO cell fed-batch cultures with pH and temperature shifts[J]. Journal of Biotechnology.

Niu, H., McCallum, G.B., Chang, A.B., et al., 2025. Exploring unsupervised feature extraction algorithms: tackling high dimensionality in small datasets[J]. Scientific Reports, 15(1): 21973.

Tang, W., Carey, S.K., 2022. Classifying annual daily hydrographs in Western North America using t-distributed stochastic neighbour embedding[J]. Hydrological Processes, 36(1): e14473.

Razavi-Termeh, S.V., Sadeghi-Niaraki, A., Razavi, S., et al., 2024. Enhancing flood-prone area mapping: fine-tuning the K-nearest neighbors (KNN) algorithm for spatial modelling[J]. International Journal of Digital Earth, 17(1): 2311325.

Abderzak, M., Zeghmar, A., Leila, B., et al., 2025. Ensemble learning-driven optimization of coagulant dosing for drinking water treatment plants using a scalable framework for smart and sustainable process control[J]. Environmental Research: 123229.

Boddu, Y, A, M, Jayanth, T., 2025. Enhanced environmental time-series forecasting using ICA-LSD Bayesian LSTM: a robust approach for accurate and uncertainty-aware predictions[J]. Earth Science Informatics, 18(3): 487.

Kaur, H., Bansod, B.S., Khungar, P., et al., 2025. Combining clustering and ensemble learning for groundwater quality monitoring: a data-driven framework for sustainable water management[J]. Environmental Science and Pollution Research: 1-42.

Naresh, V.S., Reddi, S., 2025. Quantum-enhanced predictive analytics in healthcare: benchmarking QSVM and QNN on medical datasets[J]. Measurement: 119099.

Lamichhane, P., Rawat, D.B., 2025. Quantum Machine Learning: Recent Advances, Challenges and Perspectives[J]. IEEE Access.

Vedavyasa, K.V., Kumar, A., 2025. Classification Analysis of Transition Metal Compounds Using Quantum Machine Learning[J]. Advanced Quantum Technologies, 8(5): 2400081.

Khalil, M., Zhang, C., Ye, Z., et al., 2025. PegasosQSVM: A Quantum Machine Learning

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

Approach for Accurate Fake News Detection[J]. Applied Artificial Intelligence, 39(1): 2457207.

Tehrani, M.G., 2024. Quantum Cybersecurity Analytics: Evaluation of Hybrid QML for DGA Botnet Detection[D]. The George Washington University.

Liao, H., Wang, D.S., Sitdikov, I., et al., 2024. Machine learning for practical quantum error mitigation[J]. Nature Machine Intelligence, 6(12): 1478-1486.

Cowlessur, H., Alpcan, T., Thapa, C., et al., 2025. A Qubit-Efficient Hybrid Quantum Encoding Mechanism for Quantum Machine Learning[J]. arXiv preprint arXiv:2506.19275.

Merabet, K., Di, Nunno.F., Granata, F., et al., 2025. Predicting water quality variables using gradient boosting machine: global versus local explainability using SHapley Additive Explanations (SHAP)[J]. Earth Science Informatics, 18(3): 1-34.

Alam, G.M.I., Arfin, Tanim.S., Sarker, S.K., et al., 2025. Deep learning model based prediction of vehicle CO2 emissions with eXplainable AI integration for sustainable environment[J]. Scientific Reports, 15(1): 3655.

Li, R., Feng, K., An, T., et al., 2024. Enhanced insights into effluent prediction in wastewater treatment plants: Comprehensive deep learning model explanation based on shap[J]. ACS ES&T Water, 4(4): 1904-1915.

Hollmann, N., Müller, S., Purucker, L., et al., 2025. Accurate predictions on small data with a tabular foundation model[J]. Nature, 637(8045): 319-326.

Austin, G.I., Pe'er, I., Korem, T., 2025. Distributional bias compromises leave-one-out cross-validation[J]. Science Advances, 11(48): eadx6976.

Ren, M., Sun, W., Chen, S., 2021. Combining machine learning models through multiple data division methods for PM2. 5 forecasting in Northern Xinjiang, China[J]. Environmental Monitoring and Assessment, 193(8): 476.

Gul, N., Khan, A.U., Ullah, B., et al., 2025. Optimizing LSTM for sediment load prediction in the Swat River Basin, Pakistan: Evaluation of optimizers and activation functions[J]. Physics and Chemistry of the Earth, Parts A/B/C, 140: 104019.

Brown, C.F., Kazmierski, M.R., Pasquarella, V.J., et al., 2025. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data[J]. arXiv preprint arXiv:2507.22291.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

1260 Alvarez, C.I., Ulloa, Vaca.C.A., Echeverria, Llumipanta.N.A., 2025. Machine learning for urban
1261 air quality prediction using Google AlphaEarth Foundations satellite embeddings: A case
1262 study of Quito, Ecuador[J]. Remote Sensing, 17(20): 3472.
1263 Tollefson, J., 2025. Google AI model creates maps of Earth 'at any place and time'[J]. Nature, 644:
1264 313.
1265 Ilyas, M., Niaz, R., Persio, L.D., et al., 2025. A spatiotemporal approach for drought monitoring
1266 using Empirical Orthogonal Function (EOF) analysis and neural networks[J]. Earth Systems
1267 and Environment: 1-20.
1268 Zhu, J.J., Yang, M., Ren, Z.J., 2023. Machine learning in environmental research: common pitfalls
1269 and best practices[J]. Environmental Science & Technology, 57(46): 17671-17689.
1270 Saha, G.K., Rahmani, F., Shen, C., et al., 2023. A deep learning-based novel approach to generate
1271 continuous daily stream nitrate concentration for nitrate data-sparse watersheds[J]. Science of
1272 the Total Environment, 878: 162930.
1273 Li, J., Liu, G., Shen, Z., 2025. Integrating spatiotemporal variability of non-point source pollution
1274 and best management practice efficiency to improve adaptive watershed management[J].
1275 Water Research: 124042.
1276 Bigler, M.C., Brusseau, M.L., Guo, B., et al., 2024. High-resolution depth-discrete analysis of
1277 PFAS distribution and leaching for a vadose-zone source at an AFFF-Impacted site[J].
1278 Environmental Science & Technology, 58(22): 9863-9874.
1279 Udu, A.G., Salman, M.T., Ghalati, M.K., et al., 2025. Emerging SMOTE and GAN-variants for
1280 Data Augmentation in Imbalance Machine Learning Tasks: A Review[J]. IEEE Access.
1281 Silva, R., Melo-Pinto, P., 2023. t-SNE: A study on reducing the dimensionality of hyperspectral
1282 data for the regression problem of estimating oenological parameters[J]. Artificial
1283 Intelligence in Agriculture, 7: 58-68.
1284 Kurniawan, R., Fitriansyah, A., Lestari, F., et al., 2024. Optimizing the Identification of Suitable
1285 Congregations for Preachers Using a GMM-PCA-BIC Hybrid Clustering Approach[C]//2024
1286 7th International Conference of Computer and Informatics Engineering (IC2IE). IEEE: 1-6.
1287 Peng, D., Gui, Z., Wei, W., et al., 2025. Sampling-enabled scalable manifold learning unveils the
1288 discriminative cluster structure of high-dimensional data[J]. Nature Machine Intelligence:
1289 1-16.

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

1290 Lamichhane, P., Rawat, D.B., 2025. Quantum Machine Learning: Recent Advances, Challenges
1291 and Perspectives[J]. IEEE Access.
1292 Ranga, D., Rana, A., Prajapat, S., et al., 2024. Quantum machine learning: Exploring the role of
1293 data encoding techniques, challenges, and future directions[J]. Mathematics, 12(21): 3318.
1294 Adhikari, M., 2022. Hybrid Computing Models Integrating Classical and Quantum Systems for
1295 Enhanced Computational Power: A Comprehensive Analysis[J]. Journal of Advanced
1296 Computing Systems, 2(12): 1-9.
1297 Alam, M.M.T., Milas, A.S., 2025. Dimensionality Optimized Machine Learning Retrieval of
1298 Canopy Chlorophyll, Nitrogen, and Phosphorus from Google Satellite Embeddings[J]. Smart
1299 Agricultural Technology: 101601.
1300 Sun, H., Zheng, W., Wang, S., et al., 2024. Variation of nitrate sources affected by precipitation
1301 with different intensities in groundwater in the piedmont plain area of alluvial-pluvial fan[J].
1302 Journal of Environmental Management, 367: 121885.
1303 Wang, J., Hao, X., Liu, X., et al., 2025. Groundwater–surface water exchange affects nitrate fate
1304 in a seasonal freeze–thaw watershed: Sources, migration and removal[J]. Journal of
1305 Hydrology, 654: 132803.
1306 Wang, J., Wang, M., Zhang, C., et al., 2025. Vertical Migration Characteristics and Effect Factors
1307 of BaP in Contaminated Soil Under Rainwater Infiltration[J]. Water, Air, & Soil Pollution,
1308 236(14): 1-19.
1309 Wu, H., Song, F., Min, L., et al., 2024. Exploring recharge mechanisms of soil water in the thick
1310 unsaturated zone using water isotopes in the North China Plain[J]. Catena, 234: 107615.
1311 Liu, S., Hao, Y., Wang, H., et al., 2023. Bidirectional potential effects of DON transformation in
1312 vadose zones on groundwater nitrate contamination: Different contributions to nitrification
1313 and denitrification[J]. Journal of Hazardous Materials, 448: 130976.
1314 Ahmed, M.A., Abdel, Samie.S.G., Badawy, H.A., 2013. Factors controlling mechanisms of
1315 groundwater salinization and hydrogeochemical processes in the Quaternary aquifer of the
1316 Eastern Nile Delta, Egypt[J]. Environmental Earth Sciences, 68(2): 369-394.
1317 Zhang, W., Xin, C., Yu, S., 2023. A review of heavy metal migration and its influencing factors in
1318 karst groundwater, Northern and Southern China[J]. Water, 15(20): 3690.
1319 Zhang, J., Zhang, L., Zheng, T., et al., 2025. Tracing Nitrate Contamination Sources and

https://doi.org/10.5194/egusphere-2026-272
Preprint. Discussion started: 29 January 2026

Transformations in a Rural-Urban Karst Groundwater System in North China Using Multiple Isotopes and Simmr Modeling[J]. Water Resources Research, 61(10): e2025WR040156.