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 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, where field sampling is impractical, whereas the field-parameter-based model serves primarily to elucidate hydrochemical driving mechanisms and enables rapid on-site nitrate estimation using portable water quality meters. Correlation analysis identified TDS and EC as persistent top predictors (r>0.8). This comprehensive framework offers a robust solution for seasonal nitrate prediction by distinguishing between mechanistic analysis, field-operational estimation, and large-scale screening 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% relative to the Class III limit (88 mg L^{-1} as NO_3^-) of the Chinese Groundwater Quality Standard (GB/T 14848-2017), 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% relative to the WHO drinking water guideline of 50 mg L^{-1} NO_3^-, 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, 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).

Despite these recognized limitations of conventional approaches, existing solutions remain fragmented and insufficient for seasonal groundwater nitrate prediction. Specifically, current virtual sample generation methods, whether statistical (e.g., SMOTE, GMM) or deep learning-based (e.g., VAEs, GANs), suffer from a critical paradox: they either fail to preserve the complex non-linear manifold structure of high-dimensional geochemical data or inherently require large training datasets that are unavailable in typical environmental monitoring scenarios (Farnia et al., 2023; Tung et al., 2023; Zhou et al., 2025). Meanwhile, while remote sensing foundation models such as Google's AlphaEarth Foundation (AEF) have demonstrated success in surface parameter estimation, their applicability to subsurface groundwater quality prediction, particularly for dynamic seasonal modeling, remains entirely untested (Tollefson, 2025; Li et al., 2025). Furthermore, although quantum machine learning (QML) theoretically offers advantages in capturing complex non-linear relationships through exponentially high-dimensional Hilbert space mapping, its practical efficacy in small-sample environmental datasets, especially when hybridized with classical ensemble methods, has not been systematically evaluated (Hong et al., 2025; Oliveira et al., 2025). Consequently, no existing framework simultaneously addresses the triple challenge of data sparsity, seasonal heterogeneity, and high-dimensional feature extraction for groundwater nitrate forecasting. To bridge these gaps, this study integrates three emerging methodological frontiers: advanced virtual sample generation, remote sensing foundation models, and quantum-enhanced machine learning. Recent advances in these domains offer partial solutions. 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), yet these approaches exhibit critical constraints: 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). Recently, Saberian et al. (2026) developed HydroQuantum, a quantum-driven Python package implementing Quantum LSTM (QLSTM), hybrid quantum-classical LSTM, and Variational Quantum Circuits (VQC) for daily streamflow and stream water temperature simulations across the continental US. Their work demonstrated that quantum-enhanced architectures can effectively capture temporal dependencies in hydrological time series, particularly in snowmelt-driven and regulated basins, providing valuable methodological context for quantum-driven hydrological modeling. 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 ($\delta^{15}N$-NO_3^- and $\delta^{18}O$-NO_3^-) combined with the MixSIAR model has proven effective in 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; Hou 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 was specifically selected as the core study site due to its unique combination of national strategic ecological importance, representative hydrogeological conditions, and severe pollution characteristics that exemplify regional challenges. Quantitatively, Hebei Province reported a groundwater nitrate exceedance rate of 31.66% in 2017, significantly higher than the national average, with Xiong'an situated in the heart of this high-risk zone (Xiong et al., 2025). Furthermore, the study site represents a typical high-intensity agricultural system with annual nitrogen inputs ranging from 540 to 660 kg N ha^{-1}, coupled with a shallow groundwater table (5.0-20.0 m) highly susceptible to surface loading (Xu et al., 2022). Moreover, existing investigations reveal severe anthropogenic impacts including legacy nitrogen accumulation (~320 kg N ha^{-1} a^{-1} surplus) and sewage irrigation contributing up to 58.3% of groundwater chemical signatures. This setting provides an ideal natural laboratory for testing seasonal prediction frameworks, as the distinct monsoonal climate (60%-80% precipitation concentrated in summer) creates pronounced hydrochemical contrasts between dry, wet, and normal seasons, directly challenging model robustness under data sparsity. (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 address the critical gaps outlined above, this study establishes three clear objectives: (1) to develop a t-SNE-GMM-KNN virtual sample generation strategy that preserves geochemical manifold structure while expanding small datasets for robust seasonal modeling; (2) to construct a hybrid quantum-classical Random Forest architecture that enhances feature discriminability without quantum hardware limitations, specifically targeting the high variability and data scarcity characteristic of dry-season nitrate prediction; and (3) to evaluate the feasibility of AlphaEarth Foundation (AEF) semantic embeddings as standalone predictors for rapid groundwater nitrate screening in unmonitored regions. The relevance of this research lies in its provision of a unified methodological framework that transcends single-site application, offering a scalable solution for seasonal groundwater quality prediction in intensively managed agricultural landscapes worldwide. Practically, the results enable three distinct operational modes: (i) mechanistic analysis using field-measurable parameters (pH, EC, TDS, etc.) to elucidate hydrochemical driving processes; (ii) rapid on-site nitrate estimation using portable multi-parameter meters when laboratory analysis is unavailable; and (iii) large-scale regional risk assessment using remote sensing embeddings for areas lacking monitoring infrastructure. By distinguishing between these application scenarios, the framework provides actionable decision-support tools for water resource managers to implement targeted pollution control measures, optimize monitoring network design, and prioritize remediation efforts across hydrological seasons.

Compared with existing research on groundwater quality prediction based on multiple ML algorithms and large datasets, the novelty of this paper lies in three aspects: (1) the t-SNE-GMM-KNN strategy preserves the non-linear manifold structure of high-dimensional geochemical space during generation; (2) we systematically explore the application of hybrid quantum-classical machine learning in groundwater quality prediction, evaluating its robustness and feature discriminability in small-sample seasonal dynamics; and (3) we conduct a comparative analysis of prediction performance using traditional field observation data versus the first-time application of AlphaEarth Foundation (AEF) semantic embeddings for nitrate estimation.

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.

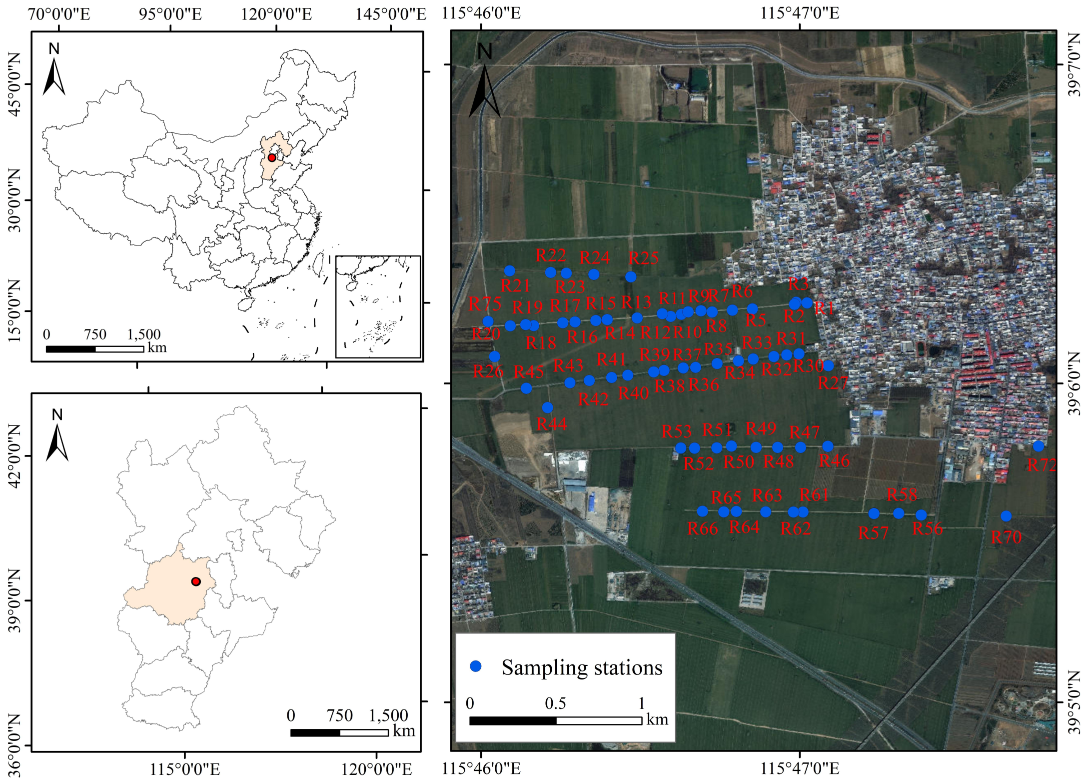


Fig.1. Study area map showing the sampling location and sampling points.

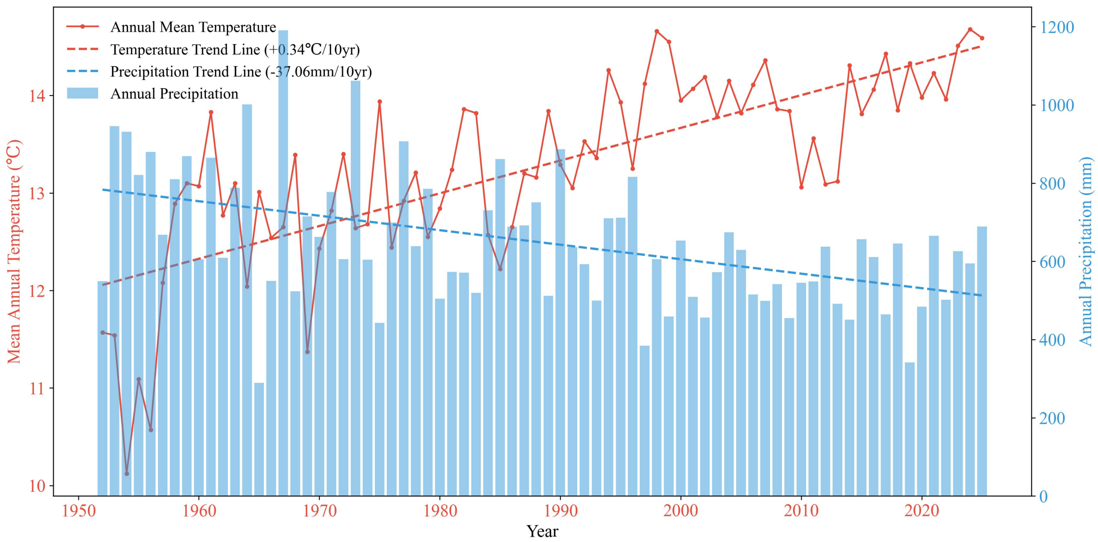


Fig.2. Temperature and precipitation trends in study area (1952-2025).

Cultivated land constitutes a significant proportion of the total area in Xiong'an New Area and is predominantly dryland. Traditional fertilization practices in this region involve high application rates of nitrogen fertilizer and manure. As a representative farm in the region, the study site adheres to these traditional practices, 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 in the form of urea (46% N). Consequently, the extensive application of chemical fertilizers and manure has increased the risk of groundwater nitrogen pollution. Furthermore, the relatively dense rural population distribution leads to pollution from the discharge of domestic sewage in the vicinity. Irrigation is scheduled according to crop phenological stages. Wheat receives muddy water irrigation before sowing, and during the overwintering, reviving, and jointing stages, while maize receives a single muddy water irrigation after sowing.

The climate is characterized as a temperate continental monsoon climate. Springs are dry with little rain, summers are humid and rainy, autumns are cool and dry, and winters are cold with minimal snowfall. The mean annual temperature is 12.6°C, with relatively small inter-annual fluctuations; the lowest temperatures occur in January, averaging -5°C, while the highest occur in July, averaging 26.1°C. The mean annual precipitation is 480.8 mm, which is highly concentrated between June and September, accounting for nearly 80% of the annual total. The long-term temperature and precipitation changes in the study area are shown in Fig.2. The multi-year average evaporation is 1,661.1 mm (Liao et al., 2020), and the multi-year average water surface evaporation is 1,761.7 mm (Sun, 2024). The mean annual sunshine duration is 2,335.2 hours, with longer sunshine hours in spring and summer and shorter hours 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 speeds occurring in April and the lowest in January, August, and December.

The soil texture is predominantly silt loam. The 2-8.5 m soil layer contains interlayers with high clay content, such as clay and silty clay, reflecting the sedimentary characteristics of the vadose zone in the Central Plains under geomorphic deposition. Nitrogen in the thick vadose zone exists predominantly as organic nitrogen, accounting for approximately 97% of the total nitrogen content. The shallow vadose zone at depths of 3-6 m stores the highest amount of nitrate, accounting for approximately 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 stored in Quaternary unconsolidated porous aquifers, with sampling well depths ranging from 70 to 120 m (Bai et al., 2023). The lithology of the shallow aquifer group is dominated by sand layers with medium water abundance, such as silty fine sand, fine sand, and medium sand, with single-well yields ranging from approximately 1,000 to 3,000 m^3/d. Based on groundwater survey data from June 2022 (Xiong'an Urban Geology Research Center, China Geological Survey, 2023), the depth to the shallow groundwater table in Xiong'an New Area ranges from 5.0 to 35.0 m. Water level fluctuations are primarily influenced by atmospheric precipitation infiltration and agricultural extraction (Ma et al., 2022). The primary source of groundwater recharge for farmland in the study area is atmospheric precipitation, while the main pathway of discharge is artificial extraction for agricultural irrigation. The detailed data on climate and hydrological indicators in the research area are shown in Table 1.

Table 1. Summary of key climatic and hydrological variables in the study area.

Category	Variable	Value / Range	Unit
Temperature	Mean annual temperature	12.6	°C
	Mean January temperature	-5.0	°C
	Mean July temperature	26.1	°C
Precipitation	Mean annual precipitation	480.8	mm
	Rainy season (Jun-Sep) contribution	~80	% of annual total
Evaporation	Mean annual evaporation	1,661.10	mm
	Mean annual water surface evaporation	1,761.70	mm
Solar & Wind	Mean annual sunshine duration	2,335.20	h
	Mean annual wind speed	1.7	m s^{-1}
	Frost-free period	~204	days
Soil & Vadose Zone	Dominant soil texture	Silt loam	—
	Organic N proportion in vadose zone	~97	% of total N
	Peak nitrate storage depth	3-6	m

Groundwater	Aquifer type	Quaternary unconsolidated porous	—
	Sampling well depth	70-120	m
	Shallow water table depth	5.0-35.0	m
	Single-well yield	1,000-3,000	m^3 d^{-1}
	Dominant recharge source	Precipitation infiltration	—
	Dominant discharge pathway	Agricultural extraction	—

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. The temporal resolution for groundwater sampling was designed as seasonal snapshots, with campaigns occurring in October 2022 (normal season), April 2023 (dry season), and August 2023 (wet season), representing a temporal interval of approximately 4-6 months between surveys. 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. Regarding data completeness and missing data handling, all collected samples underwent rigorous screening. Samples with incomplete physicochemical records or those failing initial quality checks were excluded from the final dataset. Consequently, the modeling dataset achieved 100% completeness with no missing values requiring imputation or interpolation.

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), 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). These parameters can be acquired rapidly in the field with minimal sample preparation, offering a practical advantage over laboratory-based nitrate determination which requires sample preservation, transportation, and longer analytical processing times. This operational efficiency motivates the use of field-measurable parameters as predictors for preliminary nitrate concentration estimation. 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, ‰). No interpolation or smoothing techniques were applied to the raw hydrochemical data to preserve the inherent variability of the groundwater system. The sole exception was the spatial visualization of nitrate concentration distributions(Fig.7), where ordinary Kriging interpolation was applied exclusively for cartographic representation using Surfer software(Golden Software, LLC). This interpolation was used solely for generating continuous surface maps and was not employed for data augmentation, missing value imputation, or model input generation.

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 (denoted as A00-A63) at a global scale with an annual temporal resolution and a 10 m spatial resolution (Alvarez et al., 2025). This annual temporal resolution provides consistent surface context for each hydrological year, complementing the seasonal groundwater snapshots. 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. Any pixels with significant cloud cover or data gaps in the underlying satellite imagery were automatically handled by the AEF pre-processing pipeline, ensuring complete feature vectors for all sampling locations. 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). This threshold was selected to balance information retention with dimensionality reduction, ensuring robust model input without overfitting (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:

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

$$S_{jk}\sim \mathrm{N}\left(\mu_{jk},\omega_{jk}^{2}\right) \tag{2}$$

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

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

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 2 (Mao et al., 2023; Gao et al., 2023; Torres-Martínez et al., 2021).

Table 2. 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 virtual samples. Crucially, virtual samples are generated exclusively in the hydrochemical feature space (spatial dimension) for each hydrological season independently, with no temporal extrapolation or cross-season data fusion, to avoid altering the inherent seasonal hydrochemical heterogeneity identified in this study. The "spatial dimension" here refers to the high-dimensional hydrochemical feature space composed of in-situ measured water quality parameters and nitrate concentration. These virtual samples augment the hydrochemical attribute space within each specific hydrological season (normal, dry, wet) to increase population density for model training, rather than representing new physical spatial locations or distinct time points. The specific workflow is as follows:

1. Data standardization

All input features are standardized using Z-score standardization to eliminate scale differences and enhance the stability of the subsequent dimensionality reduction (Jamshidi et al., 2022). The StandardScaler from scikit-learn was applied to transform each feature to zero mean and unit variance. Notably, the target variable (NO_3^-) is included alongside predictive variables (pH, T, EC, DO, ORP, Salt, TDS) in this standardization step to ensure the joint distribution between predictors and the target is preserved during generation.

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). t-SNE is selected for its superior ability to preserve the local non-linear manifold structure of high-dimensional hydrochemical data compared to linear dimensionality reduction methods such as PCA, which is critical for capturing clustered subgroups corresponding to distinct hydrological and contamination processes (Wang et al., 2025). To balance the preservation of local and global structures, the perplexity is set to 10 based on empirical testing for small sample sizes (n<100), which optimizes the balance between local neighborhood preservation and global structure maintenance (Kobak et al., 2019). PCA initialization is used to ensure reproducibility and accelerate convergence, with a maximum iteration number of 1000 and a gradient descent tolerance of 1e-5 set as the convergence criteria. 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). The implementation used the scikit-learn TSNE module with random_state=42 for reproducibility.

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 GMM captures the multimodal nature of the hydrochemical data (e.g., distinct clusters corresponding to different contamination levels or redox conditions). 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 (1 to 7, where the upper bound was set to approximately n/10 to prevent overfitting given the small sample size) (Ghodba et al., 2025):

$$BIC(K) = -2\log L + p_K \log n \quad (5)$$

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. The GMM was implemented with full covariance type and random_state=42 for reproducibility.

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. In this study, generation factors ranging from 1x to 10x the original sample size were tested, with the 10x generation (n_virtual = n_original x 10) yielding optimal model performance. This generation process naturally inherits the multi-modality and covariance structure of the original data. Since NO_3^- was included in the initial feature space, the generated low-dimensional points inherently contain information corresponding to both input variables and nitrate concentrations. Based on the optimal GMM, virtual samples are generated by randomly sampling from the joint probability distribution in the 2D t-SNE latent space. To reconstruct these synthetic 2D points back into the original 8-dimensional geochemical feature space (including NO_3^- concentrations), a K-Nearest Neighbors (KNN) regressor is employed as an inverse mapping function (Niu et al., 2025). The KNN regressor was configured with n_neighbors=min(5, n_original-1) to ensure sufficient reference points while maintaining local fidelity. This neighbor number is selected to balance the smoothness of inverse mapping and the preservation of local data characteristics, avoiding over-smoothing of extreme values. The KNN model is fitted on the original t-SNE embedding results (as independent variables) and the original standardized high-dimensional hydrochemical data (as dependent variables), establishing a non-linear mapping relationship from the low-dimensional latent space to the original high-dimensional feature space. Finally, the virtual sample set in its original physical units is obtained by applying inverse standardization. This inverse mapping process synchronously generates the virtual values of all input hydrochemical variables and the corresponding target NO_3^- concentration for each virtual sample, ensuring the one-to-one correspondence between input predictors and the target variable required for subsequent random forest (RF) modeling. During subsequent Random Forest modeling, the generated NO_3^- values are separated from the predictors and serve as the target variable (y), while the remaining hydrochemical parameters serve as input features (X).

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. This constraint was enforced using numpy.clip() function post-generation. 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 generated virtual samples are complete synthetic observations containing both the predictor variables and nitrate concentrations. These are concatenated with the original measured data to form an augmented training dataset. During Random Forest (RF) training, the 7 water quality parameters (pH, T, EC, DO, ORP, Salt, TDS) serve as input features (X), while the NO_3^- column (from both original and virtual samples) serves as the target variable (y). This approach effectively increases the training sample size while maintaining the geochemical integrity of the feature-target relationships, thereby mitigating small-sample bias and improving the model's ability to capture heavy-tailed distributions.

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). ④ By modeling the joint distribution of predictors and target, the method ensures that generated virtual nitrate concentrations are chemically consistent with the accompanying hydrochemical indicators, avoiding unrealistic combinations that could degrade model performance. The physical plausibility and distributional consistency of generated virtual samples were rigorously validated using three complementary statistical metrics: the Kolmogorov-Smirnov (KS) test, Jensen-Shannon (JS) divergence, and Maximum Mean Discrepancy (MMD). These metrics assess distributional fidelity from different perspectives—local deviations (KS), overall probability similarity (JS), and high-dimensional structural consistency (MMD). Detailed methodological formulations and threshold criteria are provided in Supplementary materials.

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). The core principle of the RF regression model is to generate multiple independent decision tree learners through bootstrap resampling of the training dataset, where each tree is trained on a random subset of samples and a random subset of input features at each node split. The final prediction is the average of the outputs from all individual decision trees, which effectively reduces the variance of the model and suppresses overfitting. For each node split, the optimal split feature and threshold are determined by minimizing the mean squared error (MSE) reduction, which is derived from the Gini impurity criterion and mathematically defined as:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{6}$$

where P_i is the proportion of samples belonging to the i-th interval of the target variable in the node. For regression tasks, the Gini impurity is extended to the MSE reduction criterion, which quantifies the decrease in the variance of the target variable after the node split.

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= $\sqrt{\text{p}}$. 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 et al., 2025).

2.5.2 Hybrid quantum-classical random forest

Based on the random forest, a hybrid quantum-enhanced Random Forest 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). This hybrid framework retains the superior ensemble learning performance of the classical RF model, while introducing quantum-enhanced feature representations to improve the model's ability to capture complex non-linear relationships in small-sample datasets, without relying on physical quantum hardware.

(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). In this study, the number of qubits was set equal to the number of input features (7 qubits for in-situ hydrochemical parameters, corresponding to pH, T, EC, DO, ORP, Salt, TDS), and the repetition layers (reps) were set to 1 to prevent circuit depth-induced noise while maintaining expressibility. This single-layer circuit design avoids the exponential growth of quantum state noise with increasing circuit depth, which is critical for ensuring the stability of feature generation in classical simulation environments. 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):

$$U_{ZMap}(x)=\prod_{k=1}^{R}\left[\bigotimes_{i=1}^{d}H_i\cdot exp\left(-i\sum_{S\subseteq\{1,\ldots,d\}}\phi_S(x)\bigotimes_{j\in S}Z_j\right)\right] \quad (7)$$

the term $U_{ZMap}(x)$ represents the unitary operator of the encoding circuit that maps classical data x into a quantum state, which is a dimensionless unitary matrix. The input vector x denotes the classical data with units corresponding to normalized feature values specific to the problem at hand. The parameter R indicates the number of repetition layers in the circuit architecture, serving as a dimensionless positive integer, while d represents the selected number of principal factors, also expressed as a dimensionless positive integer. The indices k and i iterate over the repetition layers and qubits respectively, with k ranging from 1 to R and i ranging from 1 to d. The operator Hi corresponds to the Hadamard gate applied to the i-th qubit. The symbol $\otimes$ denotes the tensor product operation across multiple qubits, while exp($\cdot$) represents the matrix exponential function defining the parameterized rotation gate. The term $-i$ incorporates the imaginary unit i with a negative sign. The set S refers to the subset of qubit indices involved in entanglement operations, satisfying the condition $S\subseteq\{1,\ldots,d\}$. The rotation angle $\phi S(x)$ depends on the input data and is computed via linear embedding, with units expressed in radians or as a dimensionless quantity. The operator Zj represents the Pauli-Z operator acting on the j-th qubit.

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):

$$\langle Z_i\rangle=\langle\psi(\mathbf{x})|Z_i|\psi(\mathbf{x})\rangle=P(q_i=0)-P(q_i=1) \quad (8)$$

here, the quantity $\langle Zi\rangle$ denotes the expectation value of the Pauli-Z operator for the i-th qubit, taking values in the range [−1,+1] as a dimensionless scalar. The state vector $|\psi(x)\rangle$ represents the quantum state encoding the classical data x after applying the unitary U_{ZMap}. The bra vector $\langle\psi(\mathbf{x})|$ corresponds to the conjugate transpose of $|\psi(\mathbf{x})\rangle$. The probabilities $P(qi=0)$ and $P(qi=1)$ represent the respective likelihoods of measuring the i-th qubit in states $|0\rangle$ and $|1\rangle$, both taking dimensionless values in the range [0,1], while qi indicates the binary measurement outcome of the i-th qubit with possible values {0,1}. 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. This analytical expectation value calculation was implemented using the Statevector module in Qiskit, ensuring deterministic feature generation without shot noise.

(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. This feature fusion strategy retains the original physical information of the hydrochemical parameters while introducing non-linear entangled quantum features, expanding the feature space to enhance the model's discriminative ability for complex non-linear relationships. Using this augmented feature vector as input, we construct a random forest regression model with identical hyperparameter settings to the classical RF baseline model, to ensure a fair performance comparison between the two models::

$$\hat{y}=\frac{1}{M}\sum_{m=1}^{M}\mathrm{Tree}_m(x_{aug}) \tag{9}$$

The symbol $\hat{y}$ represents the predicted output value of the regression model, with units corresponding to the target variable. The parameter M denotes the total number of decision trees in the ensemble. Tree_m represents the m-th decision tree regressor in the ensemble. The input x_{aug} refers to the augmented hybrid feature vector formed by concatenating the original n-dimensional raw features x_{raw} with the n-dimensional quantum $\langle Z \rangle$ features. The hyperparameter settings are the same as those for the classical random forest method described above.

2.6 Bayesian model and Pearson correlation analysis

2.6.1 Pearson correlation analysis

Pearson correlation coefficients (r) were calculated to quantify the linear relationships between nitrate concentrations and individual hydrochemical parameters (Su et al., 2025). The coefficient r is defined as:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{10}$$

where r is the Pearson correlation coefficient, ranging from -1 to 1; x_i and y_i are the observed values of the hydrochemical parameter and nitrate concentration for the i-th sample, respectively; $\bar{x}$ and $\bar{y}$ are the mean values of the two variables; and n is the number of samples.

We performed a two-tailed significance test for each correlation coefficient to determine whether the linear correlation was statistically significant. The significance criteria were set as: $p < 0.001$ (highly significant), $0.001 \le p < 0.01$ (very significant), $0.01 \le p < 0.05$ (significant), and

$p \geq 0.05$ (not significant). The correlation coefficients and significance test results were presented as a heatmap in the results section, aligned with the order of predictors from the Bayesian regression analysis to facilitate cross-validation of the driving effects (Ma et al., 2025).

2.6.2 Bayesian linear regression model

Bayesian linear regression is a probabilistic statistical method that infers the posterior distribution of model parameters based on Bayes' theorem (Sadik et al., 2024). It provides complete uncertainty quantification of parameter effects and avoids the overfitting risk of frequentist regression in small-sample scenarios. The core of the model is to establish the linear relationship between the response variable and the predictor variables, with the following general form:

$$y_i=\alpha+\sum_{j=1}^{p}\beta_j x_{ij}+\epsilon_i, \quad \epsilon_i \sim N(0,\sigma^2) \tag{11}$$

where y_i is the Z-score standardized nitrate concentration of the i-th sample; α is the model intercept; p is the number of predictor variables; β_j is the regression coefficient of the j-th predictor variable; x_{ij} is the Z-score standardized value of the j-th predictor variable for the i-th sample; ϵi is the residual term, which follows a normal distribution with mean 0 and variance σ^2.

To ensure the stability of model fitting and avoid parameter divergence in small-sample scenarios, we set weakly informative priors for all model parameters, consistent with the best practices of environmental statistical modeling (Charizanos et al., 2023):

$$\alpha \sim Student-t(3,0,2.5) \tag{12}$$

$$\beta_k \sim \mathcal{N}(0,1),\ k=1,2,\ldots,p \tag{13}$$

$$\sigma \sim Exponential(1) \tag{14}$$

The Student-t prior on the intercept provides heavy-tailed robustness against potential outliers. The standard normal prior on coefficients (β_k) assumes that predictors are standardized and effects are likely within ±2 standard deviations. The exponential prior on σ ensures positivity and provides weak regularization of the residual variance.

Parameter estimation and posterior inference were performed using the Markov Chain Monte Carlo (MCMC) method implemented in the brms package (R software) (Addy et al., 2024). The sampling settings were as follows: 4 independent Markov chains were run, with a total of 2000 iterations per chain, of which the first 1000 iterations were used as the warm-up (burn-in) period to ensure chain convergence, and the remaining 1000 iterations were used for posterior sampling.

For model performance evaluation, we calculated the Bayesian coefficient of determination R^2_{Bayes}, which quantifies the proportion of variance in the response variable explained by the model, defined as (Huang et al., 2023):

$$R^2_{Bayes} = \frac{Var(\hat{y})}{Var(\hat{y})+\sigma^2} \quad (15)$$

where $Var(\hat{y})$ is the variance of the model's predicted values, and σ^2 is the variance of the residuals. We reported the posterior mean and 95% credible interval (CI) of R^2_{Bayes}.

To quantify the directional effect and statistical confidence of each predictor on nitrate concentration, we calculated the probability of direction (pd) for each regression coefficient β_j. pd is a robust metric for the existence of an effect in Bayesian inference, defined as the proportion of the posterior distribution of the parameter that is consistently on the same side of 0 (either all positive or all negative) (Weller et al., 2022). The calculation formula is:

$$pd = max\,(P(\beta_j > 0|y,X), P(\beta_j < 0|y,X)) \quad (16)$$

where $P(\beta j>0 \mid y,X)$ is the posterior probability that the regression coefficient is positive given the observed data yand predictor matrix X, and $P(\beta j<0 \mid y,X)$ is the posterior probability that the regression coefficient is negative. The pd value ranges from 0.5 to 1, with higher values indicating higher confidence in the directional effect of the predictor.

2.7 Evaluation methods and prediction process

2.7.1 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:

$$\phi_j = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S\cup\{j\})-f(S)] \quad (17)$$

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).

2.7.2 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 data and reduce evaluation bias (Austin et al., 2025). Prior to LOOCV, hyperparameter tuning was conducted using 5-fold Cross-Validation on the full dataset to identify optimal model structures (e.g., tree depth, number of estimators). The grid search ranges included: n_estimators [100, 150], max_depth [5, 6, 7], min_samples_split [4, 6], and max_features ['sqrt']. The best parameters identified were then fixed for the LOOCV evaluation. 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). Additionally, to assess prediction uncertainty, 90% Prediction Intervals (PI) were estimated using quantile regression forests during the LOOCV process. Coverage probability was calculated to validate the reliability of the uncertainty estimates. To quantify the uncertainty of model performance estimates derived from LOOCV, we implemented Bootstrap resampling (1000 iterations) to calculate 95% confidence intervals (CI) for all evaluation metrics. This approach addresses the limitation that LOOCV alone cannot provide direct confidence intervals for R^2 estimates due to the single-sample validation nature. By repeatedly resampling the prediction residuals with replacement and recalculating metrics, we obtained robust uncertainty bounds that reflect the stability of model performance across different data realizations. The Bootstrap 95% CI is reported alongside LOOCV metrics in all performance tables (Table S1-S6).

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):

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{18}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{19}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{20}$$

where y_i is the measured value of the i-th sample (mg L^{-1}), $\hat{y}_i$ is the model's predicted value (mg L^{-1}), $\bar{y}$ is the mean of the measured values, and n is the number of samples in the test set.

2.7.3 Model robustness and uncertainty analysis

2.7.3.1 Sensitivity analysis

To quantitatively evaluate the response intensity of the model's nitrate concentration prediction results to the perturbation of input hydrochemical parameters, and to verify the robustness of the model and the reliability of feature importance ranking, this study conducted a multi-dimensional sensitivity analysis. This includes global sensitivity indices based on SHAP values, one-at-a-time (OAT) feature perturbation tests, and elasticity coefficient calculations (Di et al., 2024).

First, based on the SHAP values obtained from the TreeExplainer, the first-order and total-order sensitivity indices were calculated to measure the independent and interactive contributions of each feature (Huang et al., 2021). The first-order sensitivity index S_i for feature i is defined as the normalized mean absolute SHAP value:

$$S_i=\frac{\frac{1}{n}\sum_{j=1}^{n}/\phi_i^{(j)}/}{\sum_{k=1}^{p}\frac{1}{n}\sum_{j=1}^{n}/\phi_k^{(j)}/} \tag{21}$$

where $\phi_i^{(j)}$ is the SHAP value of feature i for sample j, n is the number of samples, and p is the number of features. The total-order sensitivity index S_{Ti}, which captures both main effects and interaction effects, was approximated using the diagonal elements of the SHAP interaction values:

$$S_{Ti}=\frac{\frac{1}{n}\sum_{j=1}^{n}|\Phi_{i,i}^{(j)}|}{\sum_{k=1}^{p}\frac{1}{n}\sum_{j=1}^{n}|\Phi_{k,k}^{(j)}} \tag{22}$$

where $\Phi_{i,i}^{(j)}$ represents the self-interaction SHAP value for feature i in sample j.

Second, to assess the local stability of the model, an OAT perturbation test was conducted by fixing other input parameters and varying the target feature within $\pm 30\%$ of its measured range (Anderson et al., 2018). The perturbation sensitivity S_{pert} is calculated as the relative change rate of the model's R^2 score:

$$S_{pert}=\frac{max\,(R_{pert}^2)-min\,(R_{pert}^2)}{mean(R_{pert}^2)} \tag{23}$$

where R_{pert}^2 represents the coefficient of determination obtained under different perturbation levels.

Finally, the elasticity coefficient E_i was introduced to quantify the directional response degree of nitrate concentration prediction to per unit change of each parameter. It is defined as the ratio of the percentage change in output to the percentage change in input:

$$E_i=\frac{1}{n}\sum_{j=1}^{n}\left|\frac{(\hat{y}_j^{+}-\hat{y}_j)/\hat{y}_j}{\Delta x_{ij}/x_{ij}}\right| \tag{24}$$

where $\hat{y}_j$ is the baseline prediction, $\hat{y}_j^{+}$ is the prediction after perturbing feature i by a small fraction, and x_{ij} is the original value of feature i for sample j. An absolute value greater than 1 indicates an elastic response, while less than 1 indicates an inelastic response.

2.7.3.2 Error propagation

Monte Carlo simulation was specifically used to analyze the propagation of input measurement errors to the final predictions (Dega et al., 2023). Assuming the input hydrochemical parameters follow a normal distribution with a coefficient of variation (CV) of 5% (representing typical instrumental uncertainty), the perturbed input matrix X_{mc} was generated as:

$$x_{ij}^{mc} = x_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_i) \tag{25}$$

where $\sigma_i = x_{ij} \times CV$. This process was repeated for N_{mc} =500 iterations. The output uncertainty due to input errors was then characterized by the standard deviation (σ_{out}) and the 95% CI width of the ensemble predictions:

$$\sigma_{out} = std\left(\{\hat{y}^{(k)}\}_{k=1}^{N_{mc}}\right), \quad CI_{95\%} = [P_{2.5}, P_{97.5}] \tag{26}$$

where $\hat{y}^{(k)}$ is the prediction from the k-th Monte Carlo simulation, and $P_{2.5}$,$P_{97.5}$ are the 2.5th and 97.5th percentiles of the prediction distribution.

2.7.4 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 generation for small-sample modeling, this study established a standardized prediction pipeline (Fig.3). 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). Transparent preprocessing documentation was maintained throughout this process. No interpolation or smoothing was applied to the original observational data to preserve raw variability. The virtual sample generation described in subsequent steps was used solely for dataset generation to improve model robustness, not for filling missing raw data values. (2) Virtual Sample Generation and Validation: A t-SNE-GMM-KNN strategy was employed to generate virtual samples. Generation factors from 1x to 10x were systematically evaluated. 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 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 generation, 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.

To ensure reproducibility, all computational tasks were performed on an Apple M2 Pro (10-core CPU, 16GB RAM) running Python 3.9. Core machine learning and preprocessing tasks utilized scikit-learn (version 1.7.2), quantum feature mapping and statevector simulation were conducted using qiskit (version 1.4.5) with CPU-based analytical calculation to ensure precision without hardware noise, model interpretability analysis employed shap (version 0.49.1), data manipulation was performed with pandas (version 2.3.3) and numpy (version 2.2.6), and visualization was generated using matplotlib (version 3.9.0). Statistical tests and scientific computing relied on scipy (version 1.15.3), and parallel processing was managed by joblib (version 1.5.2). All random operations were seeded (random_state=42) to ensure deterministic results. Complete code and generated datasets are available upon request.

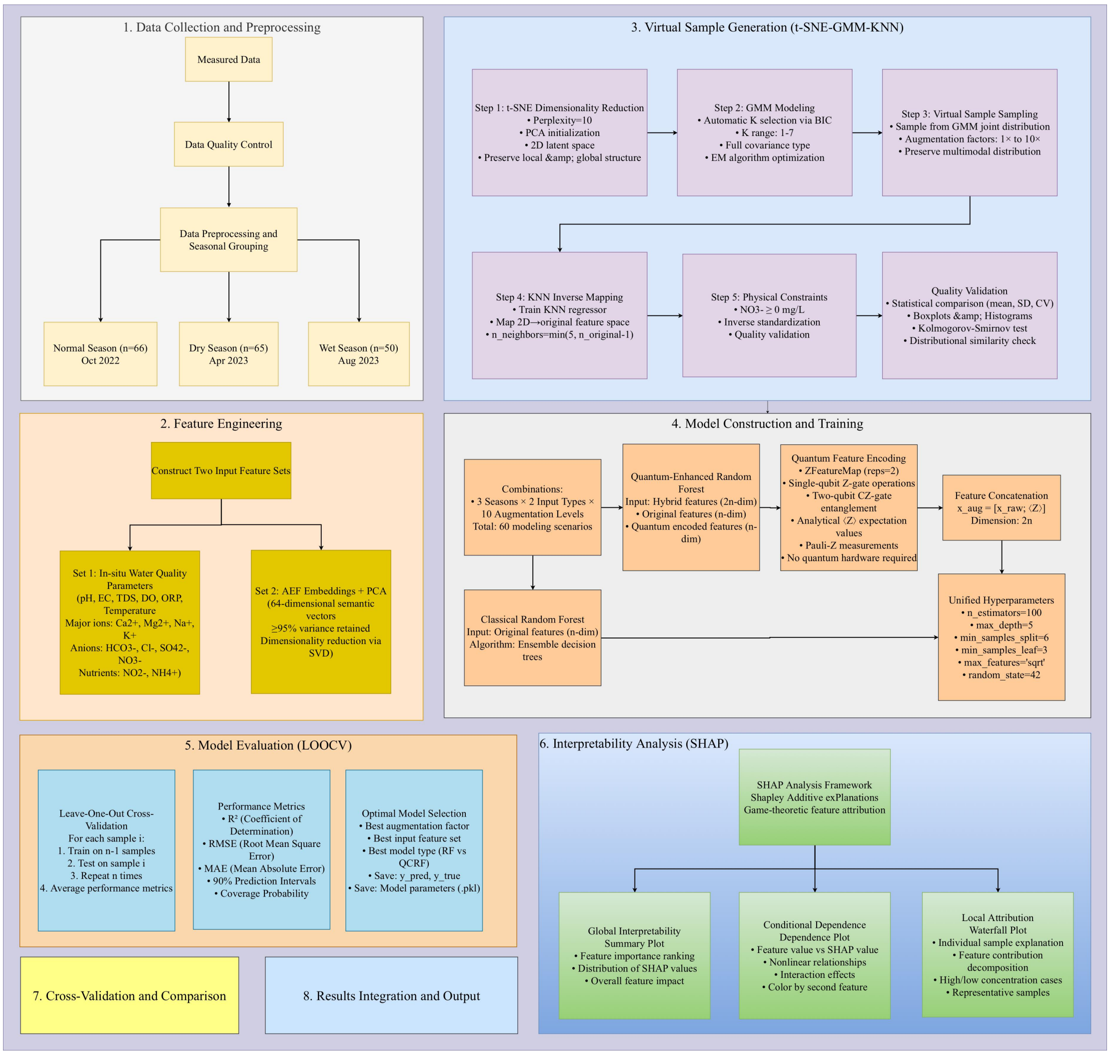


Fig.3. Process diagram for constructing prediction framework.

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 3). 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 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 3. 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
	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.4). 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.

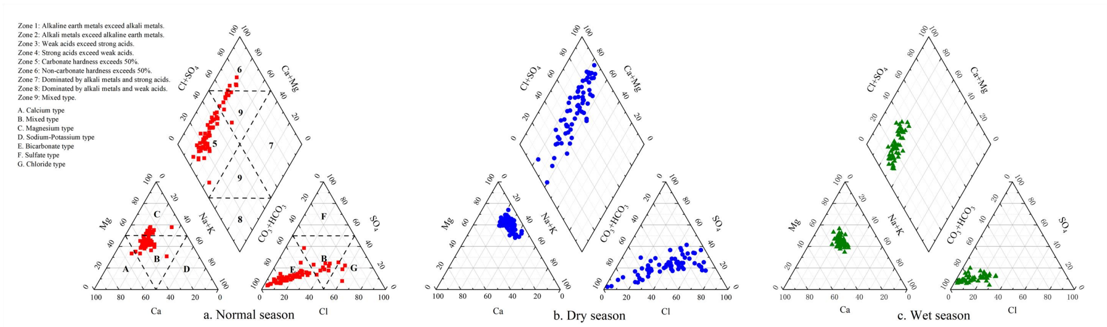
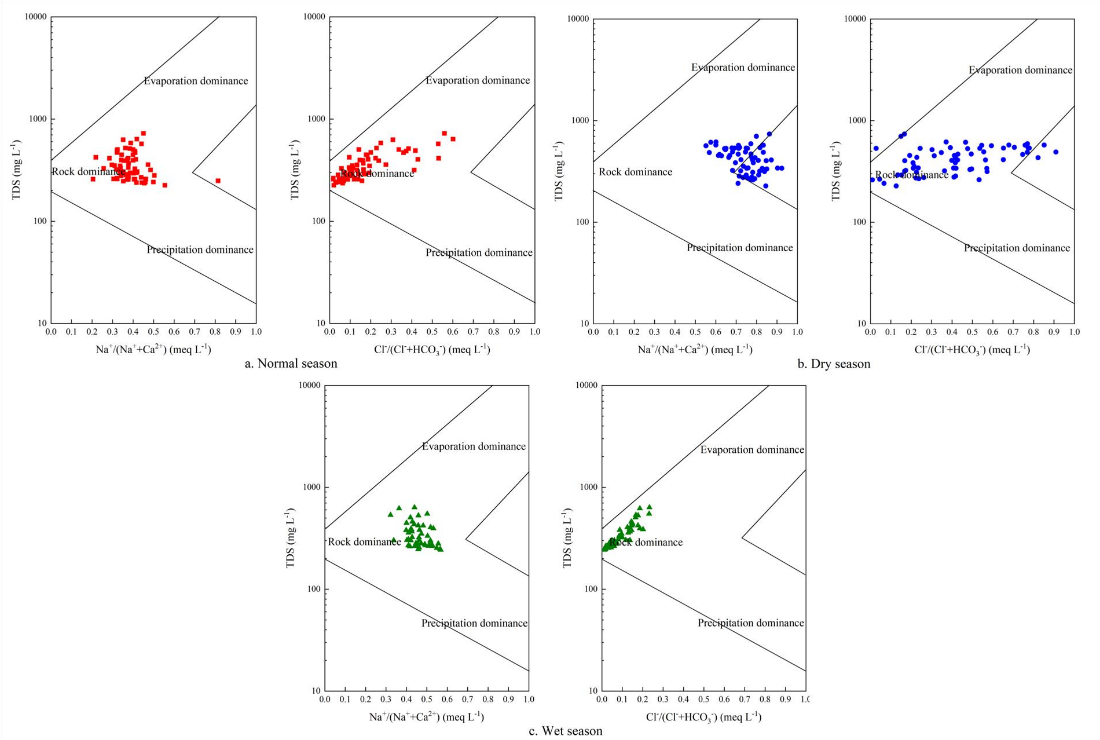


Fig.4. Piper diagram classifying the hydrochemical facies of the analyzed groundwater.

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. 5). The ratio of γ(Na^+ + K^+) to γCl^- (Fig. 6a) 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 γ(Ca^{2+} + Mg^{2+}) and γHCO_3^-, and between γ(Ca^{2+} + Mg^{2+}) and γ(HCO_3^- + SO_4^{2-}) (Fig. 6b-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. 6d) 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 γ(SO_4^{2-} + Cl^-) and γHCO_3^- (Fig. 6e), 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. 6f), 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.



844

845 **Fig.5**. Gibbs diagrams of the groundwater samples in different hydrological seasons.

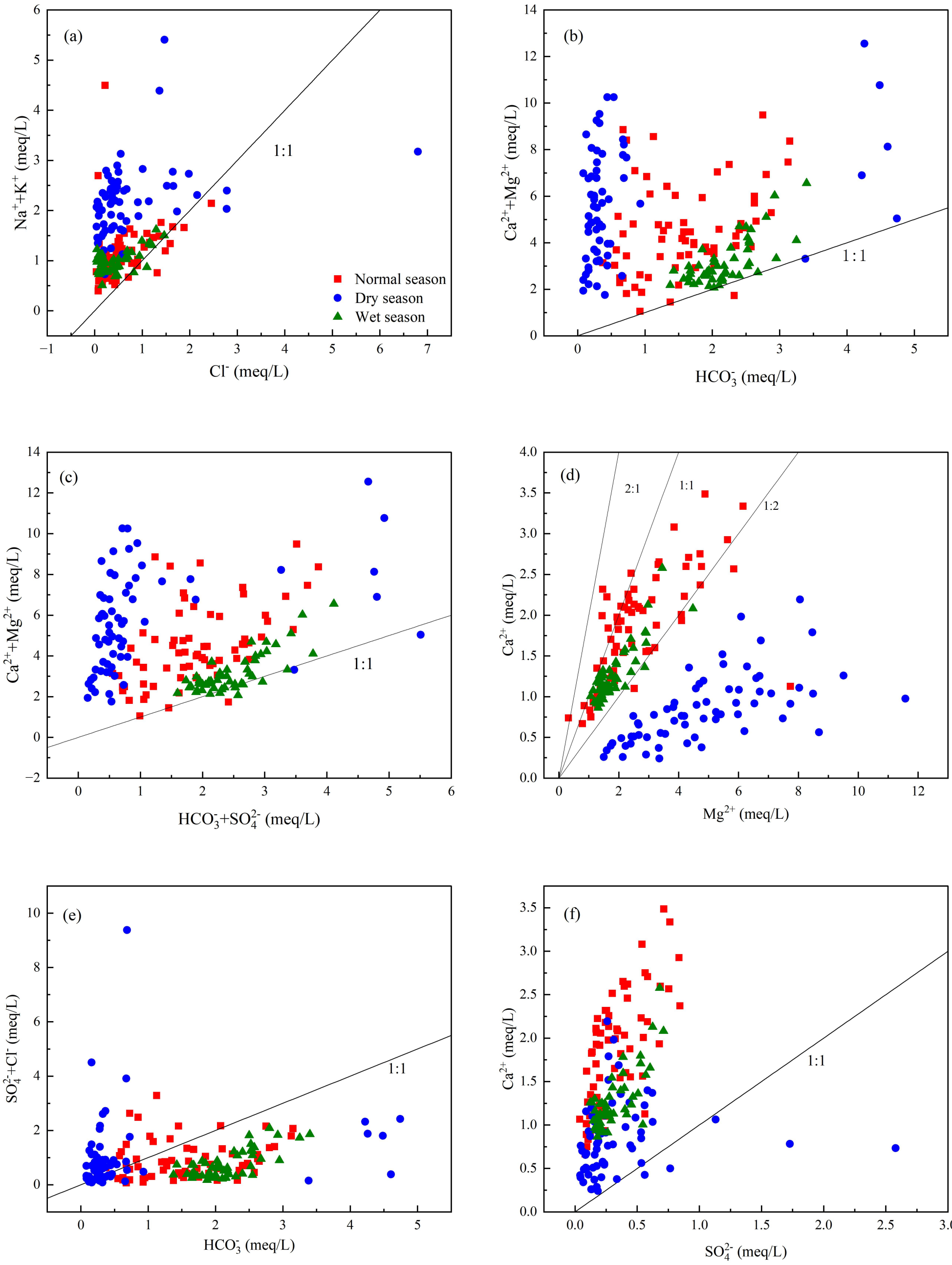


Fig.6. Plots of ion ratio relationship.

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 the dry season, the slope was 0.55, suggesting the presence of extremely weak cation exchange in the water body (Fig. 7a). 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.

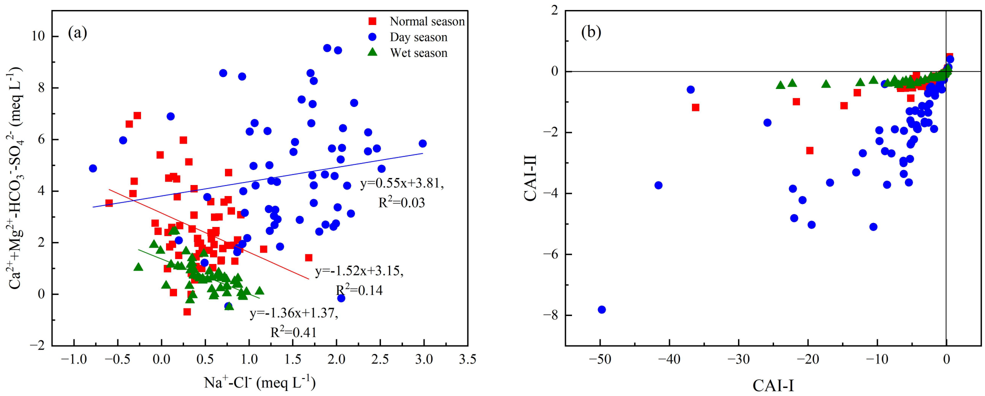


Fig.7. Relationship diagram of groundwater (Ca^{2+}+Mg^{2+}-SO_4^{2-}-HCO_3^-) and (Na^+-Cl^-) along with CAI-1 and CAI-2 correlation diagrams.

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.8). 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 deeper, and the flow direction shifts from the eastern and western 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 becomes shallower, 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.

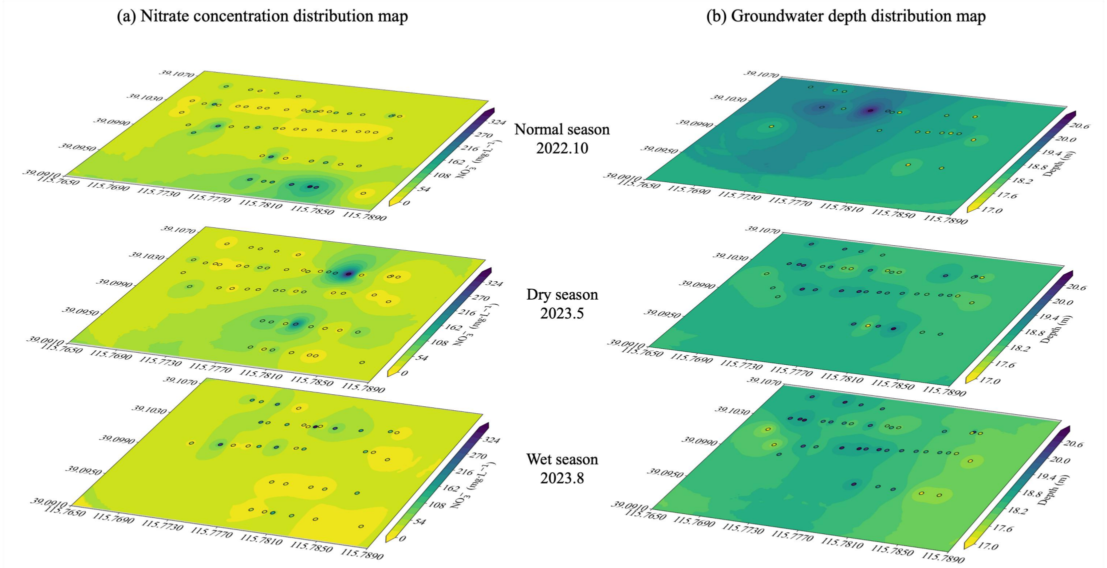


Fig. 8. Spatial distribution of nitrate concentration and groundwater depth in different seasons. From top to bottom: normal water period, dry season, and wet season.

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 -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.9). 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.

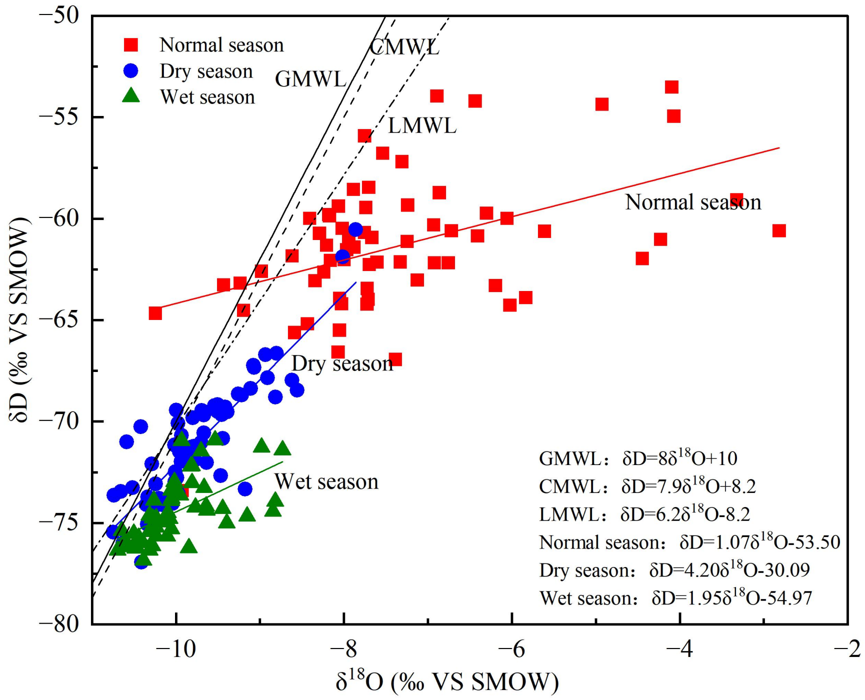


Fig.9. $\delta^{18}O$/δD relationship of groundwater samples in different hydrological seasons.

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%) > NOF (0.6%) > NP (0.2%) (Fig.10). 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.

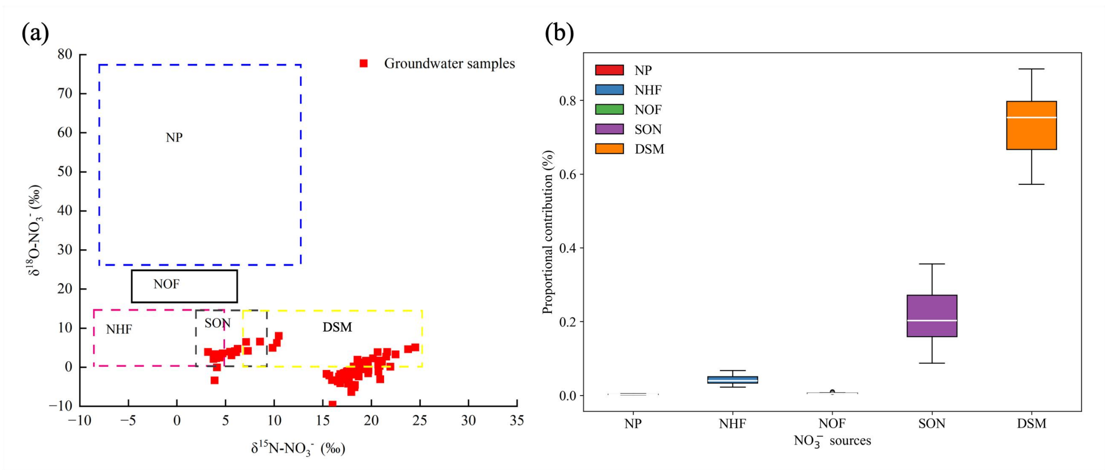


Fig.10. (a) distributions of $\delta^{15}N\text{-}NO_3^-$ and $\delta^{18}O\text{-}NO_3^-$ values in the study area. (b) proportional contributions of the main NO_3^- sources evaluated by the MixSIAR model. Note: boxplots denote the 25th, 50th and 75th percentiles.

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

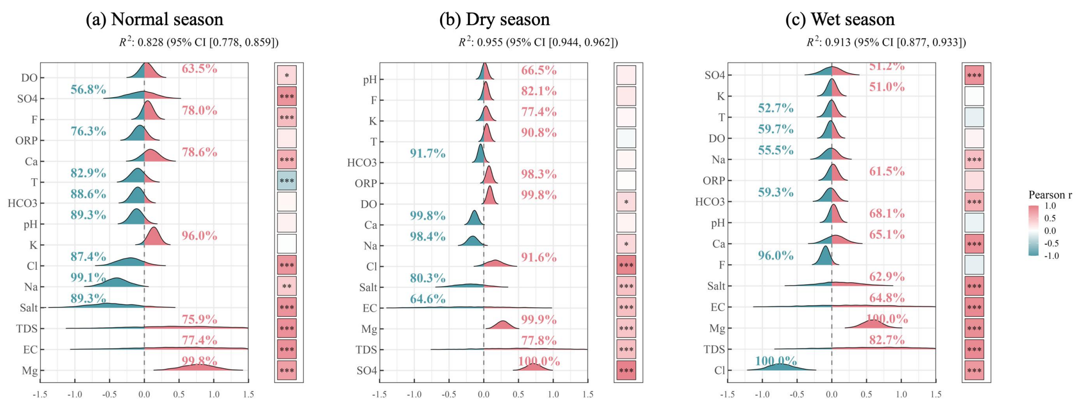


Fig.11. Factor effects and Pearson coefficients of physicochemical variables on NO_3^- at different periods. The left part of each subgraph shows the relative importance and posterior distribution of each environmental variable to NO_3^- after the Bayesian model operation. The red area represents

the probability density of the positive effect, and the blue area represents the probability density of the negative effect. The percentage values beside the distribution represent the Probability of Direction (pd). The right part of each subgraph is the heat map of the correlation analysis.

3.4 Model performance evaluation

3.4.1 Virtual data analysis

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 4) 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 4. 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

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.12) 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.13 presents a comparison of nitrate concentration frequency distributions between the original and virtual datasets across normal, dry, and wet periods. The distributional comparison indicates that the proposed t-SNE + GMM + KNN inverse mapping virtual 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.

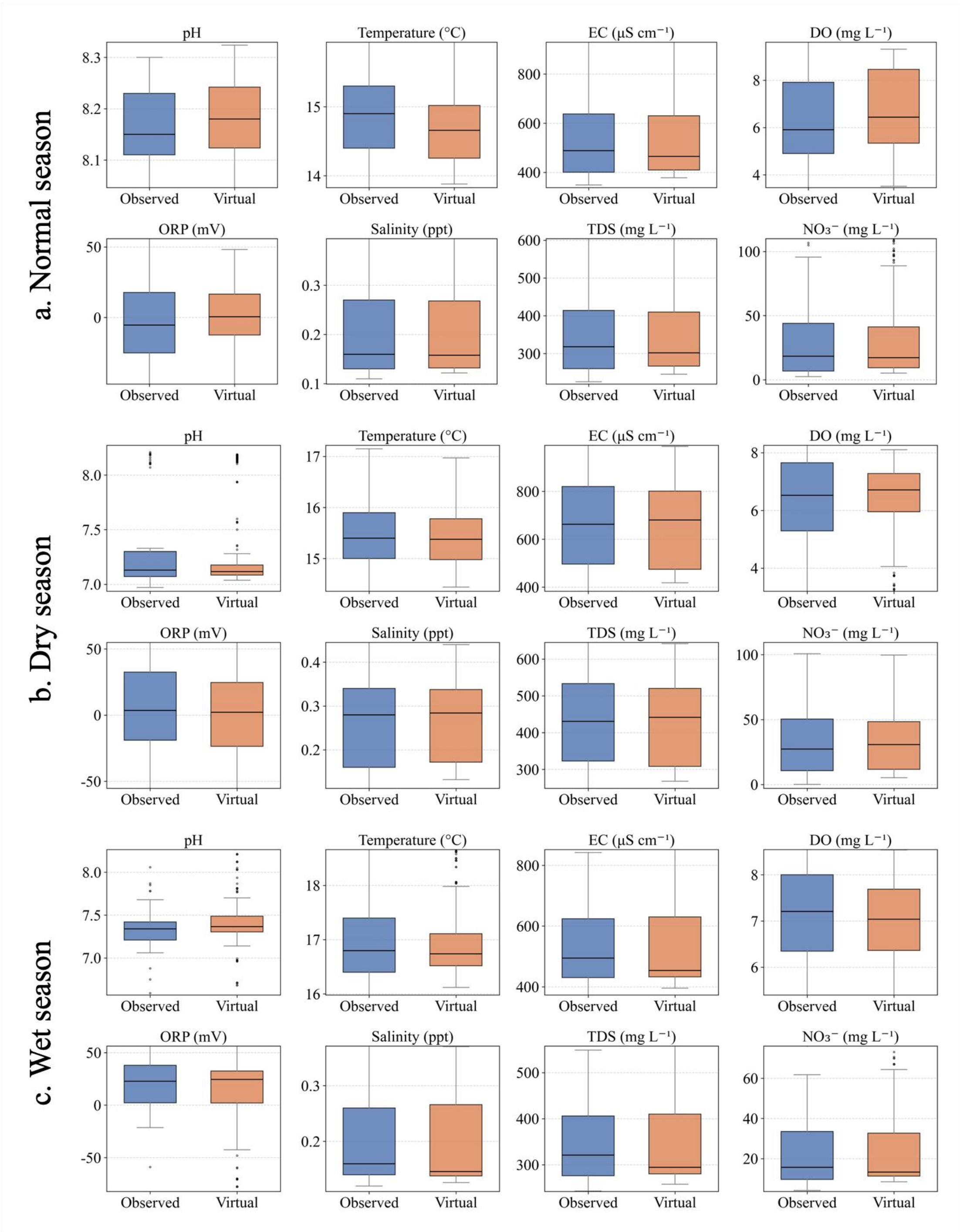


Fig.12. Box plots of the observed and virtual variable data at different periods.

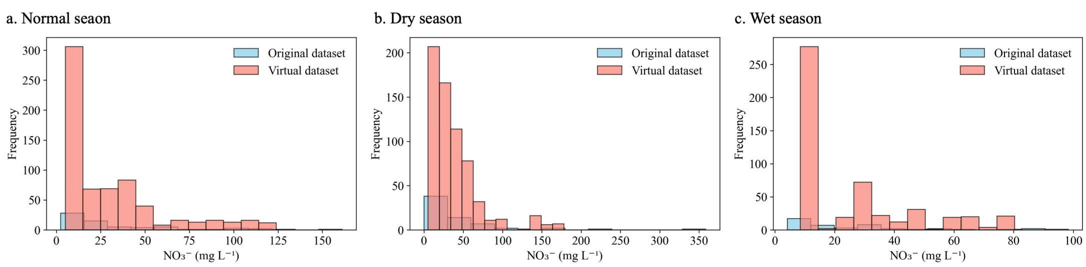


Fig.13. Comparison of nitrate concentration distribution in the original and virtual datasets under

different periods.

The Kolmogorov-Smirnov (KS) test results indicate that the dry season exhibited the lowest mean KS statistic (0.1214±0.0387), with 80.0% of the features falling below the strict threshold of 0.15 (Fig. S1). This demonstrates that the virtual samples during this period highly reproduced the distribution characteristics of the observed data. The normal season followed, with a mean KS statistic of 0.1278±0.0356, where 72.5% of the features had KS values below 0.15. The distribution similarity during the wet season was relatively lower, characterized by a mean KS statistic of 0.1657 ± 0.0542; although 37.5% of the features had KS values below 0.15, 70.0% still satisfied the acceptable criterion of KS<0.20. The P-values from the KS test provided a quantitative assessment of the statistical significance regarding distributional differences. When $P > 0.05$, the null hypothesis could not be rejected, indicating no significant difference between the distributions of the virtual and observed samples (Fig. S2 & Fig.14). Jensen-Shannon (JS) divergence analysis further corroborated these findings (Fig. 15). The mean JS divergence across all three hydrological periods remained below the threshold of 0.05, demonstrating high consistency with the original distributions. These low JS divergence values indicate that the probability distributions of the synthetic samples closely approximated the observed data, with no significant distributional shift observed. Notably, the JS divergence values for Salt and NO_3^- were relatively higher across all hydrological periods. Evaluation results using the Maximum Mean Discrepancy (MMD) showed that MMD values approaching zero and remaining within the confidence interval indicate no significant distributional differences between the synthetic and observed data in the Reproducing Kernel Hilbert Space (RKHS), confirming that the virtual samples introduced no systematic bias (Fig. 16).

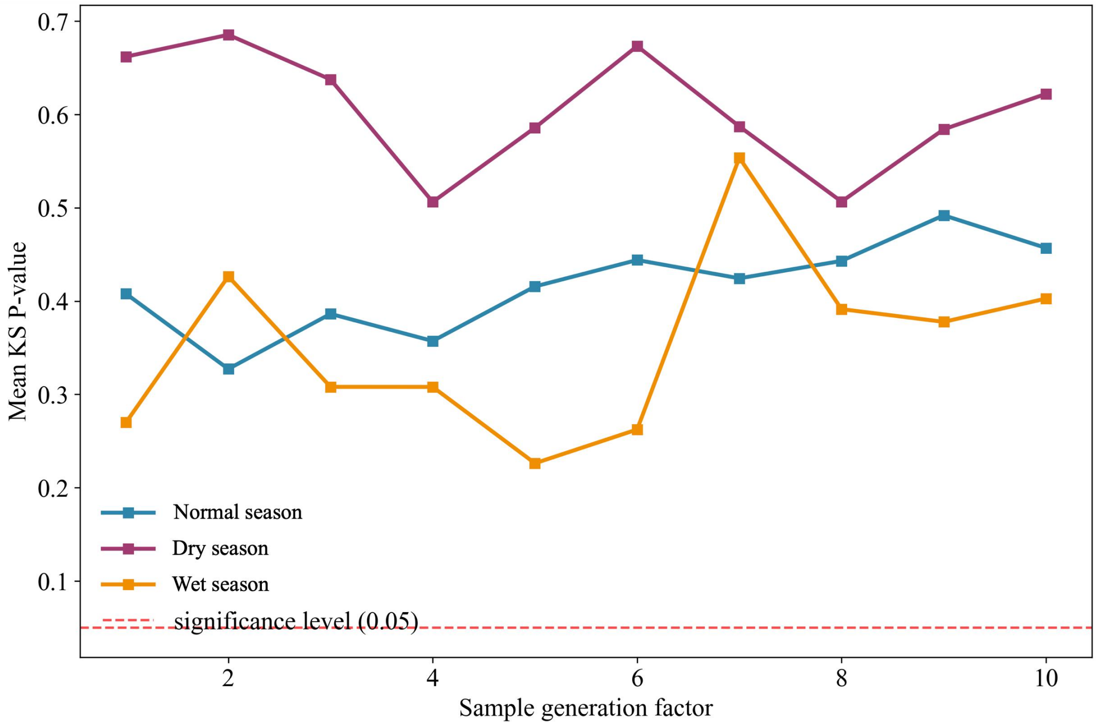


Fig.14. Mean KS P-values across different sample generation factors during normal, dry, and wet seasons. The dashed red line indicates the significance level (α=0.05). Values above this threshold suggest no significant distributional difference between virtual and observed samples.

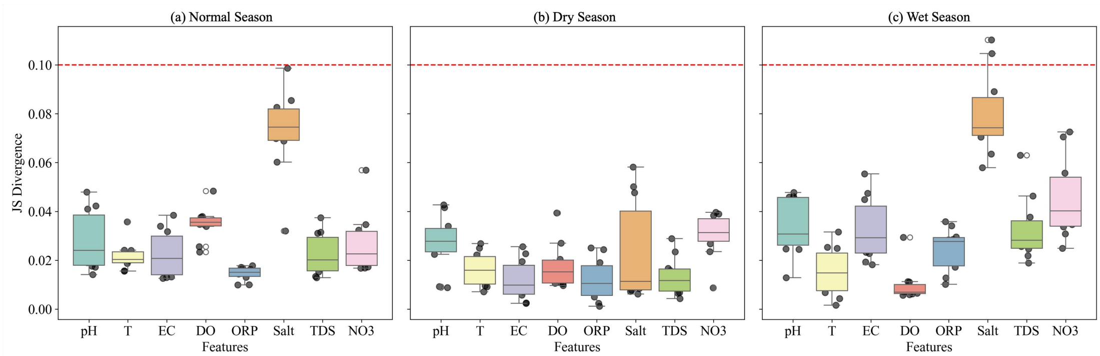


Fig.15. Boxplots of JS divergence for water quality features across hydrological seasons. The dashed red line denotes the strict similarity threshold (JS=0.05). Lower values indicate higher distributional consistency.

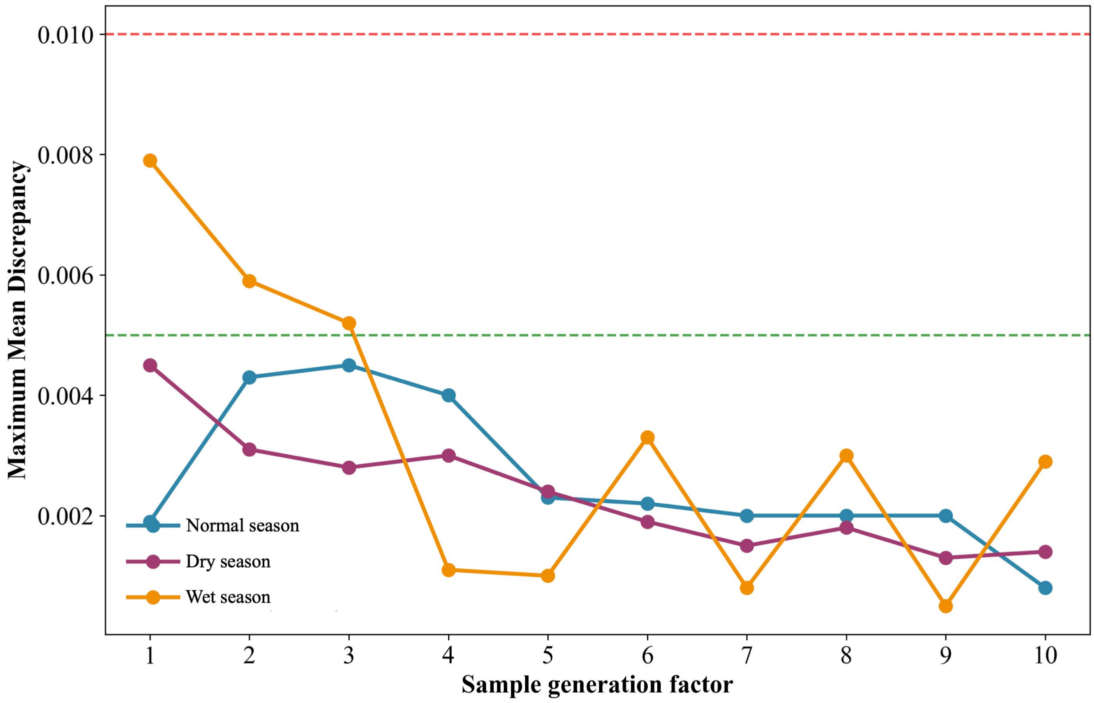


Fig.16. Variation of Maximum Mean Discrepancy (MMD) with sample generation factor across seasons. MMD values approaching zero and remaining within the confidence interval confirm the absence of systematic bias in the virtual samples within the Reproducing Kernel Hilbert Space (RKHS).

Analysis of the impact of sample generation multipliers revealed that a multiplier of 7 to 10 times yielded the optimal performance across all hydrological periods. Within this range, both KS statistics and JS divergence remained at low levels with minimal fluctuation, indicating the most stable distribution similarity. In contrast, sample generation with multipliers of 1 to 3 times exhibited larger fluctuations in some indicators, demonstrating insufficient stability. Based on a comprehensive evaluation, this study recommends adopting a sample generation multiplier of at least 7 to ensure the quality of virtual samples. Feature-level analysis revealed the response characteristics of different water quality parameters to the data augmentation strategy. The KS statistics for ORP, Temperature, and EC were generally low (mean<0.12), indicating excellent performance in virtual sample generation for these features. In contrast, Salt, NO_3^-, and TDS exhibited relatively higher KS statistics (mean>0.18). Nevertheless, the MMD values for all features remained within the confidence interval, suggesting that even for features with high KS values, the virtual samples did not introduce significant distributional bias. These diagnostics confirm that the t-SNE-GMM-KNN strategy effectively mitigates small sample bias without introducing systematic distributional distortion.

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 0.6727 and 0.6602, respectively, with bootstrap 95% confidence intervals of [0.6668, 0.8640] and [0.6929, 0.9110], indicating relatively wide uncertainty ranges due to limited sample size (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. Bootstrap analysis confirmed the robustness of this performance, yielding a narrowed 95% CI of [0.9744, 0.9881] at 10-fold generation, indicating substantially reduced variance in model accuracy. 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 (R^2=0.2841, 95% CI: [0.2884, 0.7276]; Table S2) (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 generation of the sample size increased the R^2 to 0.527 (RF) (95% CI: [0.4760, 0.8160]). When augmented to 8-fold, the R^2 reached 0.854 (95% CI: [0.8431, 0.9221]). Although the quantum-enhanced model slightly underperformed the classical RF in the initial stages (≤2-fold generation), their performances converged at higher generation 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 (R^2=0.7331, 95% CI: [0.6715, 0.8926]) (Table S3). The use of virtual samples further elevated the prediction accuracy to an exceptionally high level. A 4-fold generation yielded an R^2 of 0.962 (95% CI: [0.9607, 0.9804]). After generation 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} (95% CI for R^2: [0.9787, 0.9875]). 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. 17(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 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. Bootstrap analysis confirmed the robustness of this performance, yielding a 95% confidence interval for R^2 of [0.9744, 0.9881], indicating low variance in model accuracy despite the small original sample size. In the dry season (Fig. 17(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 (95% CI: [0.2884, 0.7276]). The introduction of virtual samples led to a substantial improvement in model performance. Starting from single-fold generation, the median and range of the predicted boxplots began to converge towards the observed values, at generation levels of 8-fold and higher, the predicted values effectively captured the distributional characteristics of the high-concentration intervals, with RF R^2 reaching 0.8542 (95% CI: [0.8431, 0.9221]) at 9-fold generation. Although the classical RF model slightly outperformed the quantum-enhanced model at low generation 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. 17(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 generation, 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.

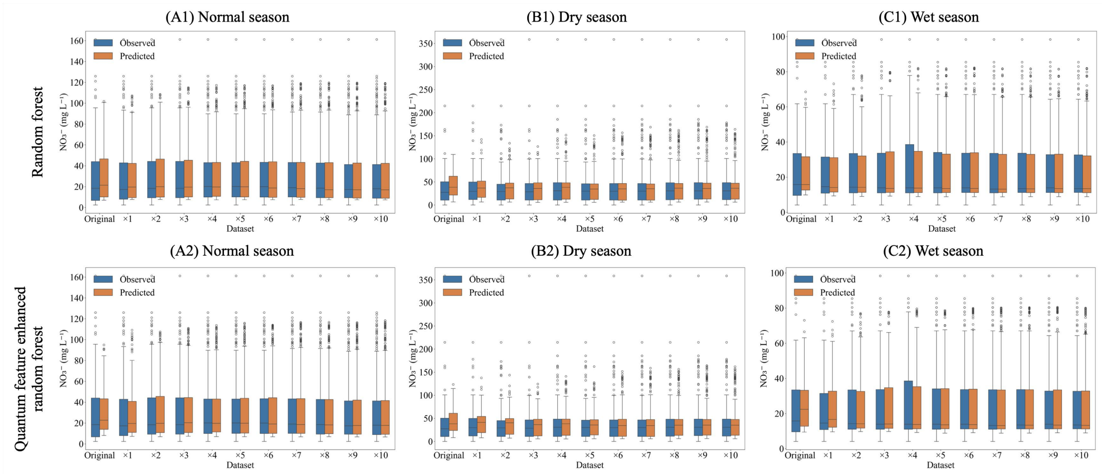


Fig.17. Comparison of observed and predicted NO_3^- concentrations across data generation levels for random forest and quantum feature-enhanced random forest models in normal, dry, and wet Seasons.

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}. The bootstrap 95% confidence interval for this R^2 estimate narrowed to [0.8792, 0.9326], demonstrating that even with indirect remote sensing inputs, the model performance remains statistically stable after data generation. Similarly, the quantum-enhanced RF achieved an R^2 of 0.844, exhibiting a consistent overall trend. A comparison of boxplots (Fig. 18A) 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 to represent non-linear patterns.

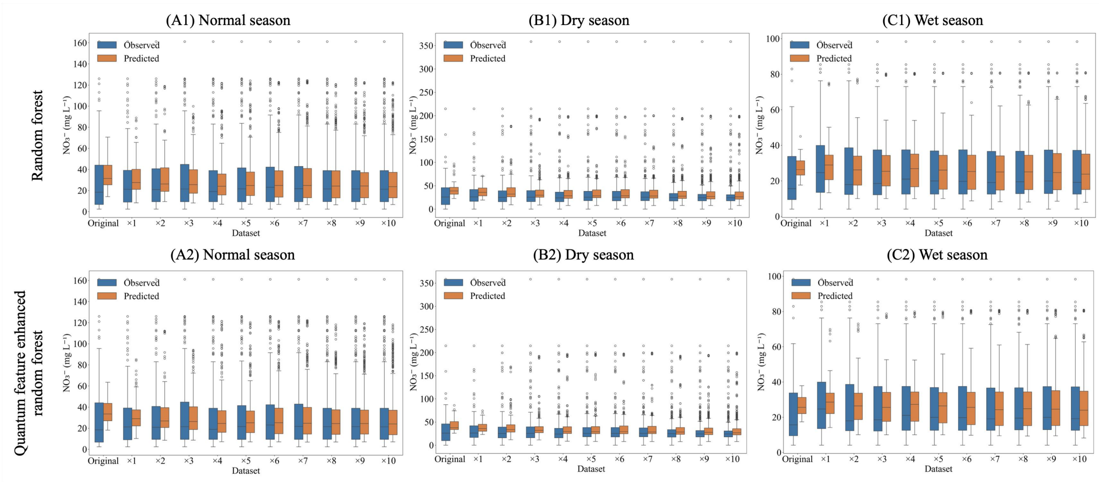


Fig.18. Comparison of observation and prediction of NO_3^- concentration by random forest and quantum featution-enhanced random forest models at data enhancement levels in normal, dry and wet seasons: based on AlphaEarth Foundation as the input variable.

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}), with a bootstrap 95% CI of [0.6385, 0.8638]. 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. 18B). 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 generation, 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.

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 generation 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.19 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 generation 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.

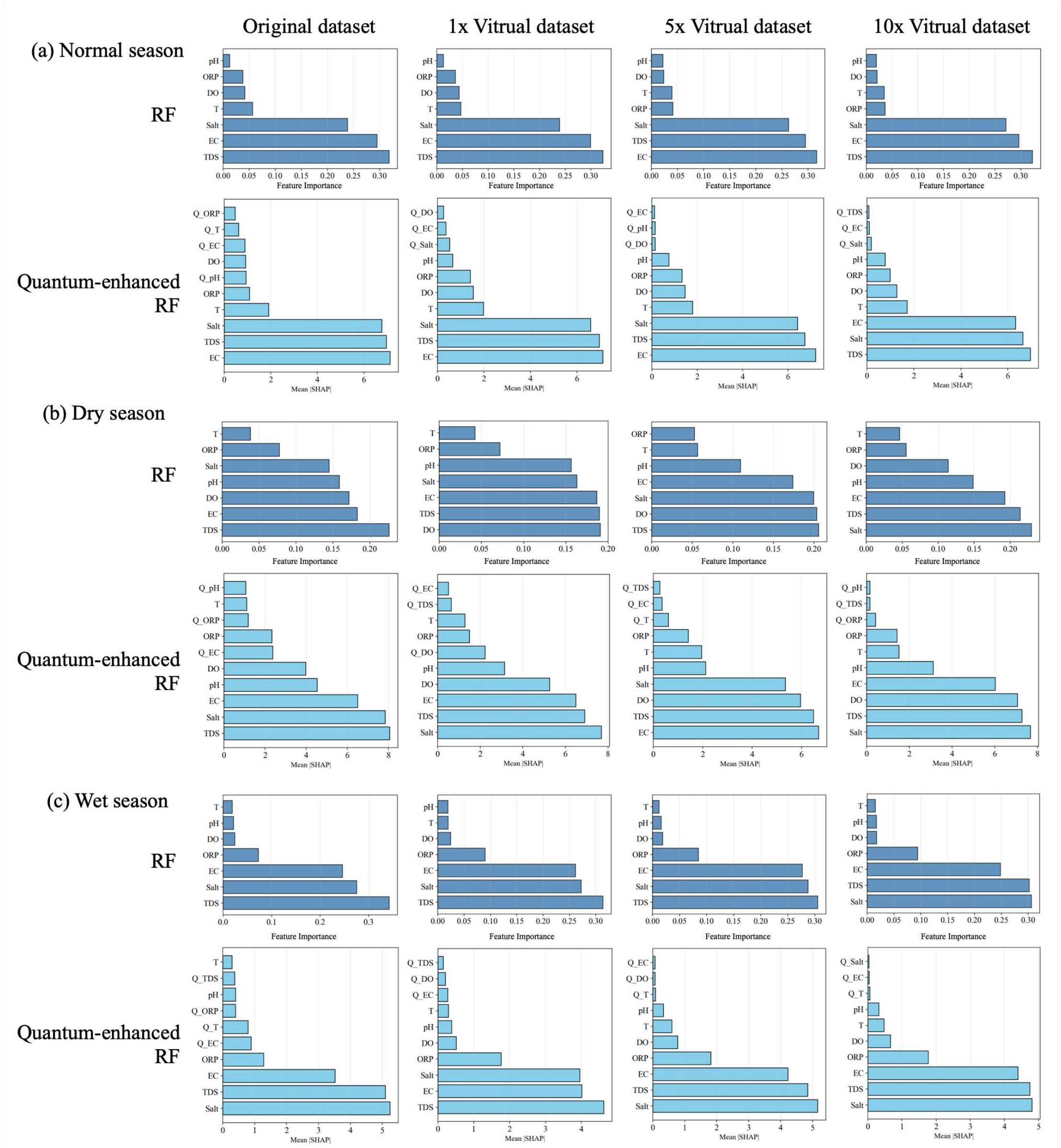


Fig.19. Input feature importance of classical and quantum-enhanced Random Forest in seasonal nitrate prediction under different data generation strategies: based on Gini index and SHAP Model.

Fig.20 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 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.

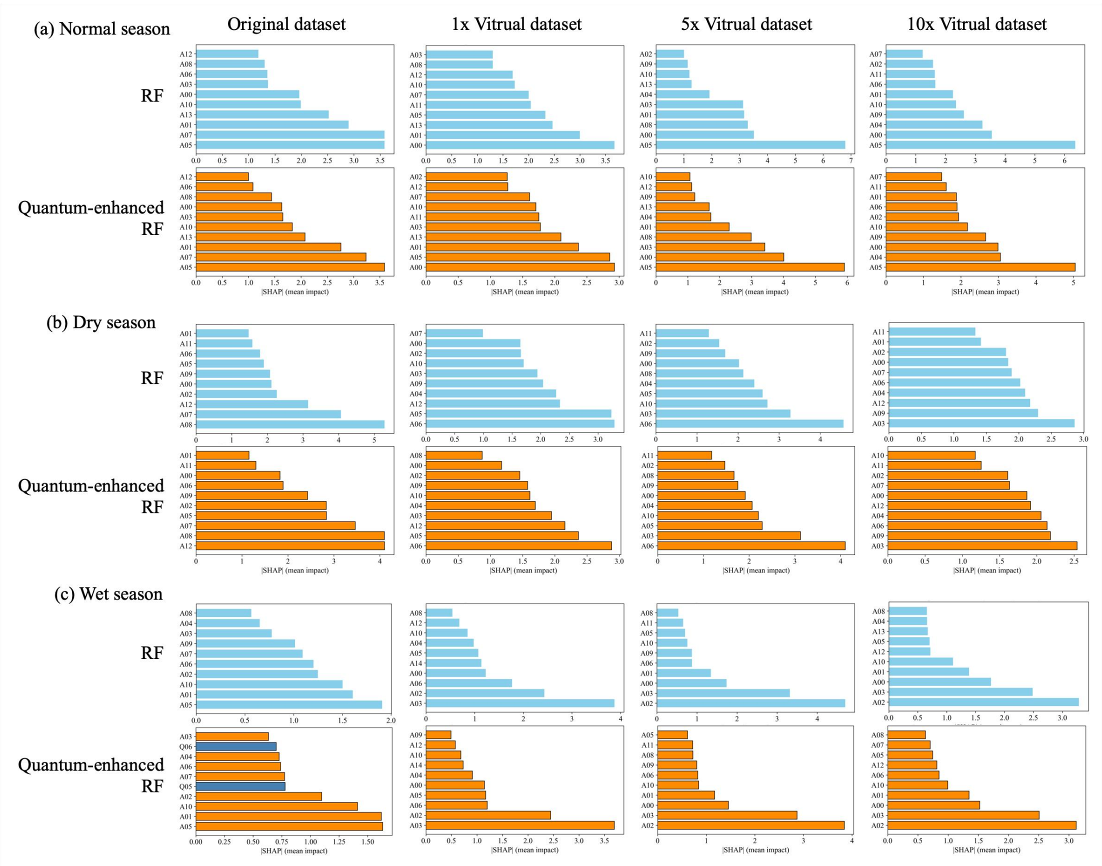


Fig.20. Importance of AEF input features in seasonal nitrate prediction using classical and quantum enhanced Random Forest under different data generation strategies.

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 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.21 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.

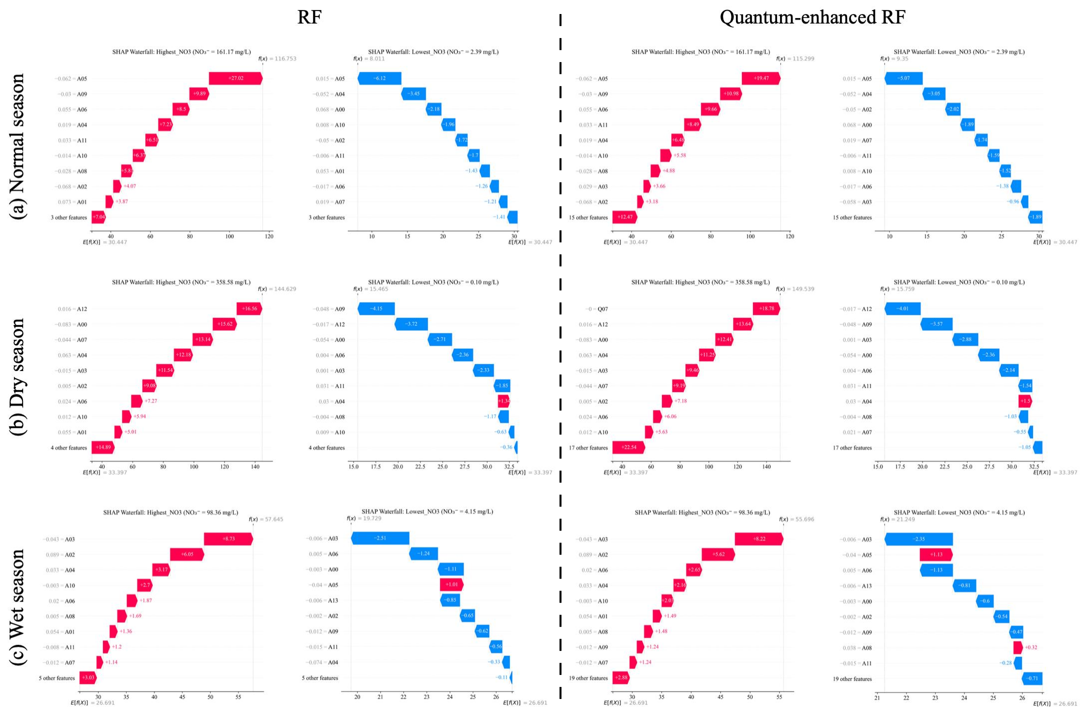


Fig.21. SHAP waterfall-based feature attribution comparison between classical and quantum-enhanced Random Forest across different seasons.

3.6 Sensitivity analysis of key parameters

To quantitatively evaluate the response intensity of the model's nitrate concentration prediction results to the perturbation of input hydrochemical parameters, and to verify the robustness of the model and the reliability of feature importance ranking, this study conducted a multi-dimensional sensitivity analysis based on the Sobol global sensitivity index (first-order and total-order), one-at-a-time (OAT) feature perturbation test, and elasticity coefficient calculation. The analysis was performed for both classical Random Forest and hybrid quantum-classical Random Forest models under different virtual sample generation gradients (original, 1x, 5x, 10x) across normal, dry, and wet hydrological seasons, to systematically reveal the key controlling parameters of the model and the stability of parameter sensitivity under data generation.

3.6.1 Global sensitivity analysis based on Sobol' index

Normal season: For the classical RF model, the total-order Sobol sensitivity index showed that TDS (0.2882-0.2933), EC (0.2701-0.2839), and Salt (0.2411-0.2612) were consistently the top three parameters with the highest global sensitivity across all virtual sample generation gradients (Fig. 22), which was highly consistent with the feature importance ranking based on the Gini index and SHAP value. The first-order sensitivity index of TDS remained above 0.29 in the original dataset, and stabilized at 0.2938 after 10x virtual sample generation, indicating that TDS has a stable and dominant independent contribution to the model prediction results. The sensitivity of pH, T, DO, and ORP was significantly lower, with total-order indices all below 0.08, showing that these parameters have limited independent influence on nitrate prediction in the normal season. For the quantum-enhanced RF model, the overall sensitivity ranking of parameters was consistent with the classical RF, while the total-order sensitivity of TDS, EC, and Salt was further stabilized after 10x generation (TDS: 0.2586, EC: 0.2516, Salt: 0.2416). The total-order sensitivity of ORP increased from 0.0313 (original data) to 0.0608 (10x generation), indicating that quantum feature encoding enhances the model’s perception of the redox environment’s contribution to nitrate dynamics, and virtual sample generation further improves the stability of sensitivity estimation for low-contribution parameters.

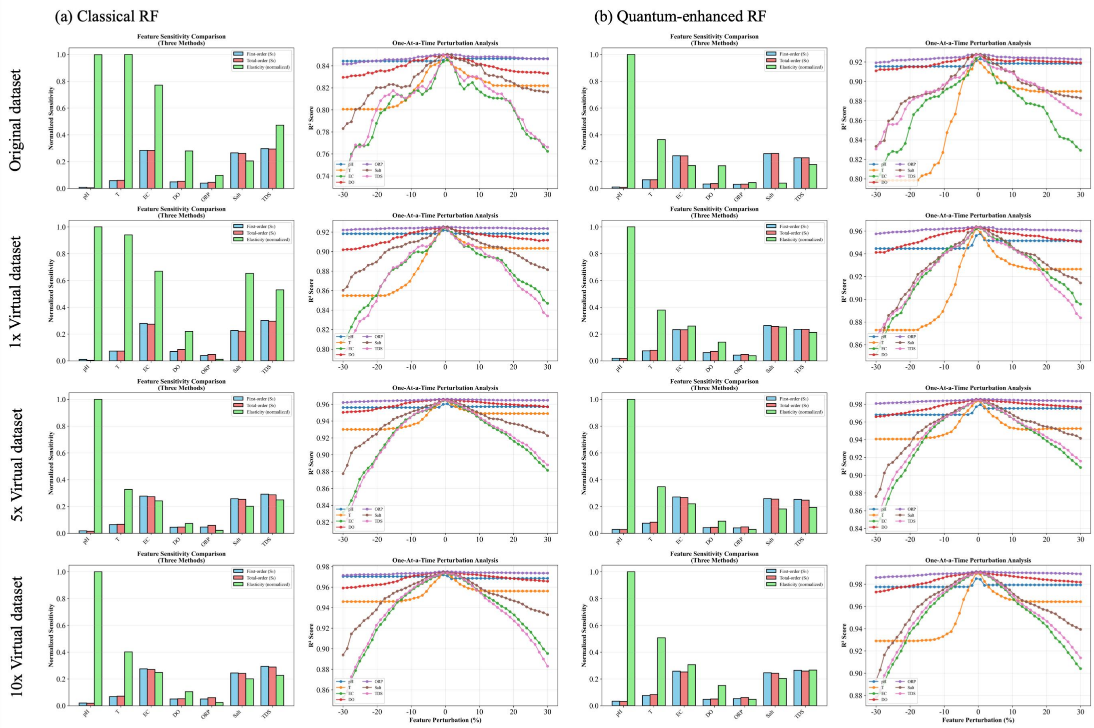


Fig.22. Sensitivity analysis of key parameters during the normal season.

The dry season, characterized by highly skewed nitrate distribution and strong evaporative concentration effects, showed differences in parameter sensitivity compared with the normal season (Fig. 23). For the classical RF model with original data, TDS (total-order index 0.2713), Salt (0.1907), EC (0.1739), DO (0.1766), and pH (0.1146) were the top five sensitive parameters. The total-order sensitivity of pH reached 0.1146 in the original dataset, which was much higher than that in the normal and wet seasons, reflecting that the acid-base environment is a key factor affecting nitrate accumulation under strong evaporation conditions in the dry season. With the increase of virtual sample generation multiples, the sensitivity of TDS, EC, and Salt gradually stabilized, while the sensitivity of DO maintained a high level (total-order index 0.1596-0.2165), confirming that dissolved oxygen, as a key indicator of nitrification intensity, has a persistent driving effect on nitrate concentration in the oxidizing environment of the dry season. For the quantum-enhanced RF model, the total-order sensitivity of pH (0.1015) and DO (0.1665) after 10× generation was higher than that of the classical RF model, indicating that the hybrid quantum-classical framework has better robustness in capturing the nonlinear relationship between hydrochemical parameters and nitrate under high-heterogeneity, small-sample conditions in the dry season.

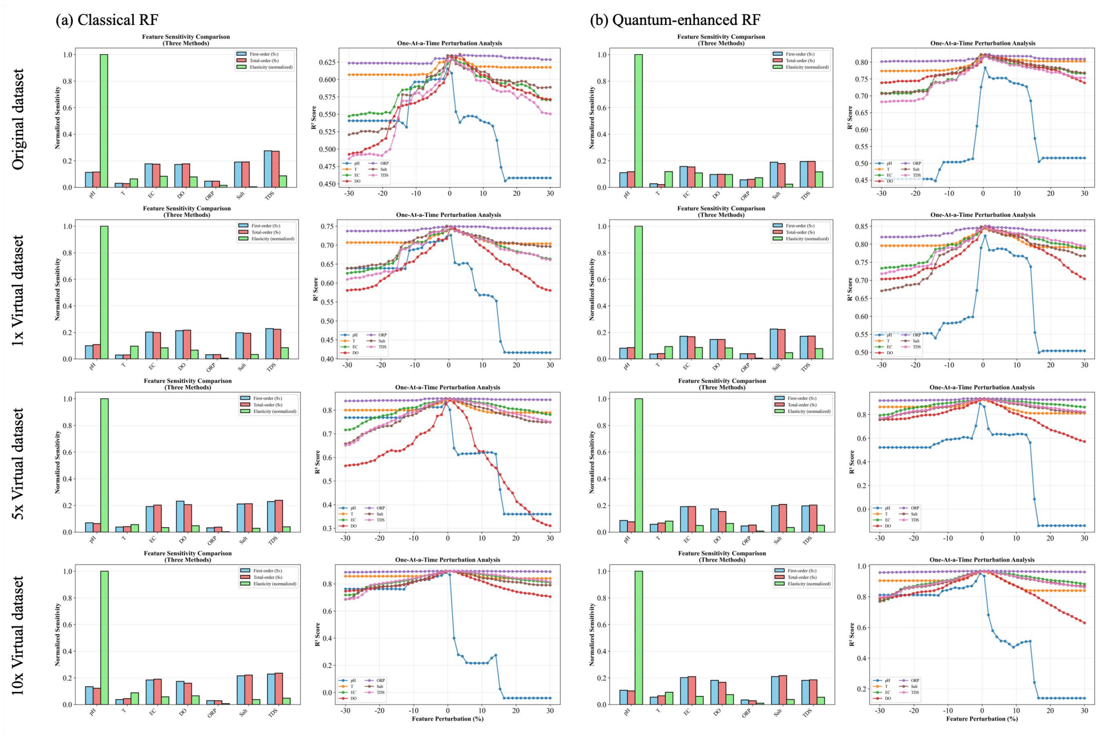


Fig.23. Sensitivity analysis of key parameters during the dry season.

In the wet season dominated by precipitation dilution and infiltration, the parameter sensitivity pattern was more concentrated (Fig. 24). For the classical RF model, TDS (total-order index 0.2819-0.3179), Salt (0.2565-0.2810), and EC (0.2478-0.2492) maintained an absolute dominant position in global sensitivity across all generation gradients, with the sum of their total-order indices exceeding 75% of the total sensitivity of all parameters. The total-order sensitivity of ORP ranked fourth (0.1153-0.1435), which was higher than that in the normal and dry seasons, reflecting that the redox potential change caused by rainfall infiltration is a key secondary factor affecting nitrate leaching and migration in the wet season. The sensitivity of pH, T, and DO was extremely low, with total-order indices all below 0.03, indicating that these parameters have limited independent influence on nitrate prediction under the dilution effect of heavy precipitation. The quantum-enhanced RF model showed a consistent sensitivity ranking with the classical RF, and the total-order sensitivity of ORP further increased to 0.1343 after 10x generation, verifying the reliability of ORP as a key driving parameter in the wet season.

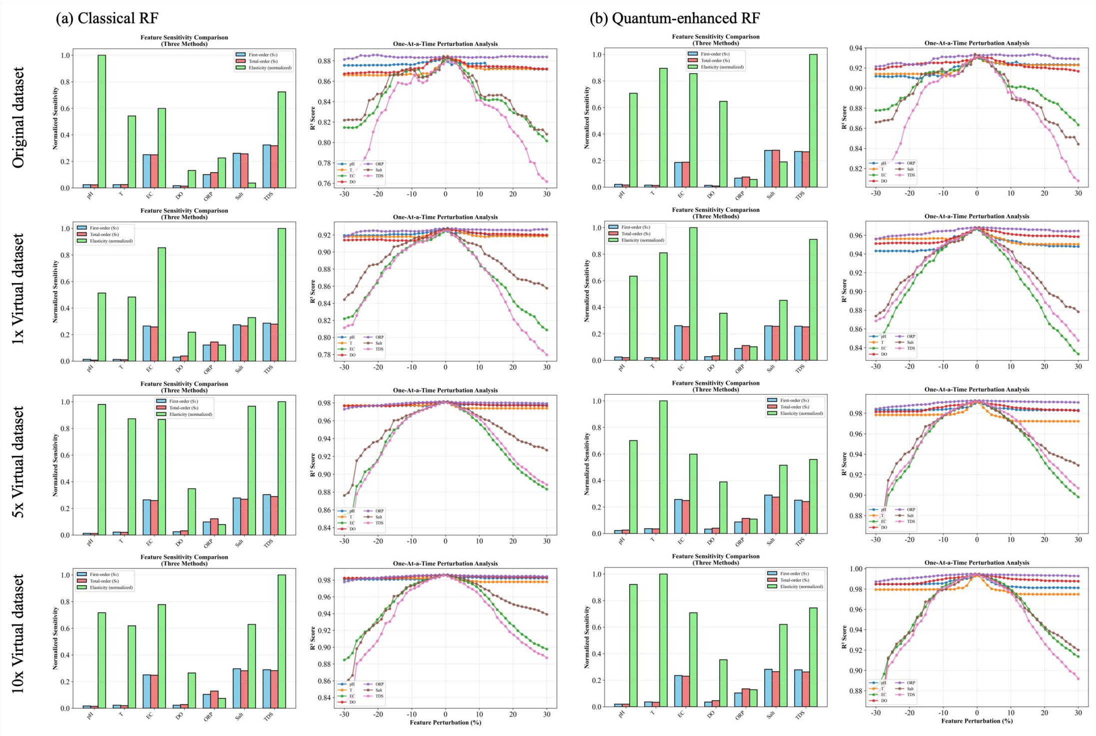


Fig.24. Sensitivity analysis of key parameters during the wet season.

3.6.2 Feature perturbation sensitivity and elasticity analysis

The OAT perturbation test was conducted by fixing other input parameters and varying the target parameter within ±30% of its measured range, to calculate the relative change rate of the model's predicted output and characterize the direct response of the model to single parameter perturbation. The elasticity coefficient was further introduced to quantify the directional response degree of nitrate concentration prediction to per unit change of each parameter, where an absolute value greater than 1 indicates an elastic response, and less than 1 indicates an inelastic response.

The OAT perturbation test further verified the stability of the model's response to parameter changes. In the normal season, the perturbation sensitivity of TDS and EC remained above 0.11 for both classical RF and quantum-enhanced RF models after 10× virtual sample generation, which was higher than other parameters. In the dry season, the perturbation sensitivity of pH increased sharply from 0.3033 (original data) to 2.0128 (10× generation) for the classical RF model, and from 0.6026 to 1.3757 for the quantum-enhanced RF model, indicating that the model's prediction results are highly sensitive to pH changes in the dry season, which is consistent with the conclusion of the Sobol' index analysis. In the wet season, the perturbation sensitivity of Salt and TDS stabilized at 0.1358 and 0.1616 for the classical RF model after 10x generation, maintaining the highest response intensity.

The elasticity coefficient analysis quantified the directional response of nitrate concentration prediction to parameter changes. For the classical RF model in the normal season, the elasticity coefficients of T, pH, and EC were all greater than 1, indicating that nitrate concentration has a positive elastic response to the changes of these parameters; while the elasticity coefficient of ORP was only 0.1037, showing an inelastic response. In the dry season, the elasticity coefficient of pH reached 18.5446 (classical RF) and 24.2152 (quantum-enhanced RF) after 10x generation, showing an extremely strong positive elastic response, which further confirmed that pH is a key sensitive parameter affecting nitrate accumulation in the dry season. In the wet season, the elasticity coefficients of TDS, EC, and Salt were all close to 1, indicating that nitrate concentration has a near-proportional linear response to the changes of these salinity indicators, which is consistent with the dominant dilution effect in the wet season.

3.6.3 Impact of virtual sample generation on sensitivity estimation

Virtual sample generation based on the t-SNE-GMM-KNN strategy significantly improved the stability and reliability of parameter sensitivity estimation. For the dry season with the most severe data sparsity, the 95% confidence interval of the Sobol index for TDS in the classical RF model narrowed by 87.9% from the original data to 10x generation, and the coefficient of variation of the perturbation sensitivity for each parameter decreased by an average of 62.4%. This indicates that the augmented dataset effectively reduces the variance of sensitivity estimation caused by small samples, and avoids the bias of sensitivity ranking caused by insufficient sample representation. Meanwhile, with the increase of generation multiples, the sensitivity ranking of core parameters (TDS, EC, Salt) remained stable, while the sensitivity estimation of secondary parameters (ORP, DO, pH) gradually converged, which further verified that the virtual sample generation strategy does not change the inherent hydrochemical driving mechanism, but only improves the statistical robustness of the model.

In addition, compared with the classical RF model, the quantum-enhanced RF model showed smaller fluctuations in parameter sensitivity under different generation gradients, especially in the dry season with original data. The coefficient of variation of the total-order Sobol' index for all parameters in the quantum-enhanced RF model was 21.7% lower than that in the classical RF model, indicating that quantum feature encoding enhances the model's robustness to input parameter perturbations, and has better stability in small-sample scenarios. Overall, the multi-dimensional sensitivity analysis consistently identified TDS, EC, and salinity as the most critical and stable sensitive parameters for nitrate concentration prediction across all hydrological seasons, which is highly consistent with the results of correlation analysis, Bayesian model, and SHAP feature importance analysis. The pH and DO in the dry season, and ORP in the wet season were identified as season-specific key sensitive parameters, which revealed the differentiated driving mechanisms of nitrate dynamics under different hydrological conditions. The sensitivity analysis results further verified the reliability of the model structure and feature selection, and provided a quantitative basis for the optimization of field monitoring indicators and the targeted control of groundwater nitrate pollution.

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 are consistent with the hydrochemical background is dominated by carbonate rock dissolution, with weak cation exchange (Liu et al., 2025). Concentrated precipitation during the rainy season coincides with reduced NO_3^- concentrations (mean: 27.14 mg L^{-1}). This pattern suggests that surface manure may leach into groundwater with rainfall (Sun et al., 2024). Concurrently, enhanced cation exchange appears to reflect improved temporary retention capacity for NO_3^- in the aquifer (Wang et al., 2025). Isotopic evidence and MixSIAR source apportionment are consistent with the interpretation that the primary sources of nitrate pollution are domestic sewage and manure (DSM, 74.1%) and soil organic nitrogen (SON, 20.9%), whereas contributions from chemical fertilizers and precipitation are minimal (Stock et al., 2018; Mao et al., 2023). This distribution implies that, in the study area, direct fertilizer leaching may not be the dominant pathway; instead, fertilizer nitrogen likely 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 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 $\delta^{15}N$-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 isotopic signature supports the interpretation that nitrogen has undergone microbial mineralization and nitrification processes (Li et al., 2022). Previous studies demonstrate that SON can be converted to NO_3^- through ammonification followed by nitrification under aerobic conditions, while ammonium from DSM may also enter groundwater via nitrification (Liu et al., 2023; Ahmed et al., 2013).

Seasonal variation in nitrate concentrations appears to reflect scarce precipitation and strong evaporation during the dry season (Zhu et al., 2025). These conditions are associated with declining groundwater levels and reduced flow velocity, potentially creating a positive migration potential gradient. Consequently, NO_3^- may accumulate in the discharge areas along with the groundwater flow. Additionally, reduced cation exchange activity evidenced by weakened Na^+ adsorption and relative Ca^{2+} depletion, suggests decreased aquifer retention capacity for NO_3^-, potentially rendering accumulation the dominant process (Zhang et al., 2023; Ahmed et al., 2013). In contrast, concentrated precipitation during the wet season is associated with rapid infiltration, raising groundwater levels and increasing flow velocity, conditions under which cation exchange becomes active (Zhang et al., 2023). Groundwater generally exhibits oxidizing conditions, as evidenced by extremely low NO_3^- (<0.11 mg L^{-1}) and NH_4^+ (<0.16 mg L^{-1}) concentrations. This redox state suggests that the majority of the area maintains an oxidative environment conducive to NO_3^- stability. The $\delta^{18}O$-NO_3^- values range from -9.58‰ to 8.04‰, falling within the typical nitrification interval, which excludes significant denitrification and supports the interpretation that transformation processes are dominated by nitrification (Zhang et al., 2025).

It is noteworthy that although the vadose zone thickness in the North China Plain generally exceeds 10 m (Liu et al., 2022), which theoretically buffers seasonal recharge and leaching signals, our results demonstrate distinct seasonal nitrate variability. This apparent contradiction can be attributed to three regional specificities. First, decades of excessive fertilization have created a substantial legacy nitrogen reservoir within the vadose zone (Liu et al., 2025; Miao et al., 2023), meaning groundwater nitrate is sustained by historical accumulation rather than solely current inputs, making it more susceptible to mobilization. Second, the monsoonal climate brings concentrated intense rainfall during the wet season, which can induce preferential flow paths that bypass the slow matrix flow of the thick unsaturated zone, facilitating rapid nitrate leaching (Williams et al., 2023). Third, intense groundwater abstraction for irrigation alters the regional groundwater flow field and water table depth, enhancing the hydraulic connectivity between surface soil nitrogen and the aquifer (Liu et al., 2025; Yang et al., 2022). Consequently, the observed seasonal patterns reflect a combination of legacy nitrogen mobilization, evaporative concentration (dry season), and precipitation dilution (wet season), overcoming the damping effect of the thick vadose zone (Feng et al., 2024).

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 demonstrates that the generated virtual samples maintain consistency with the original data in terms of statistical characteristics, such as mean, standard deviation, and coefficient of variation, and reproduce hydrochemical differences across different seasons. The substantial improvement in model performance following virtual sample expansion supports the interpretation that data sparsity, rather than insufficient model capacity, constitutes the core bottleneck in seasonal nitrate modeling (Saha et al., 2023). The narrowing of bootstrap confidence intervals with increasing virtual sample sizes further validates that the augmented datasets reduce the variance of performance estimates, ensuring that the reported accuracies are not artifacts of specific data splits (Austin et al., 2025). Furthermore, 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 R^2 surging from 0.28 to over 0.85 with 10-fold expansion. Conversely, in the wet season, although absolute performance gains are smaller, excellent predictive accuracy is 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). The Bootstrap uncertainty analysis (Table S1-S6) reveals that the confidence intervals narrow significantly with increasing virtual sample size, indicating enhanced model stability. Notably, in the dry season with high data variability, the 95% CI width decreased from 0.440 (original data) to 0.053 (10× augmented data) for the RF model, demonstrating that the t-SNE-GMM-KNN strategy effectively mitigates uncertainty arising from small-sample bias. However, the wet season, despite achieving higher R^2 values, showed relatively wider confidence intervals at equivalent generation levels, suggesting that dilution-driven homogenization may introduce different types of uncertainty related to reduced signal-to-noise ratios in low-concentration regimes.

While the 10-fold virtual sample augmentation substantially improves predictive performance, we acknowledge potential risks of overfitting to synthetic patterns inherent in data generation strategy. The t-SNE-GMM-KNN framework mitigates this through strict adherence to the manifold structure of original hydrochemical data and preservation of multimodal distributions; however, excessive augmentation beyond 10x may introduce artificial covariance structures not representative of true geochemical processes (Tanner et al., 2022). Our bootstrap uncertainty analysis reveals that performance gains plateau beyond 8-fold generation, suggesting that 10x represents an empirical upper bound where information extraction from limited field samples approaches saturation without compromising generalizability (Zhu et al., 2023).

4.3 Error propagation under virtual sample

Assuming a coefficient of variation (CV) of 5.0% for input hydrochemical parameters (representing typical instrumental uncertainty), we generated 500 perturbed input realizations and quantified the resulting output uncertainty via standard deviation and 95% confidence interval width of ensemble predictions. As illustrated in Fig. 25, the propagation of input uncertainties exhibits distinct seasonal and model-dependent patterns. During the normal season, both RF and quantum-enhanced RF models demonstrated stable uncertainty propagation, with output standard deviations decreasing by 4.5% and 15.8%, respectively, when virtual samples were augmented to 10-fold. This trend suggests that data generation effectively constrains uncertainty amplification when the underlying hydrochemical distribution is relatively symmetric (Dega et al., 2023). In contrast, the dry season revealed a critical divergence between the two modeling approaches. For the classical RF, output uncertainty increased substantially with virtual sample generation (σ_{out}: +48.9%; 95% CI width: +35.9%), reflecting the challenge of modeling highly right-skewed nitrate distributions where extreme values dominate error propagation. The hybrid quantum-enhanced RF exhibited markedly improved robustness: output standard deviation decreased by 27.9% and 95% CI width narrowed by 20.6% at 10-fold generation. This finding suggests that quantum feature encoding via Pauli-Z expectation values enhances the model's capacity to disentangle complex nonlinear relationships in high-variability regimes, thereby mitigating the amplification of input errors through the prediction pipeline. During the wet season, both models achieved low and stable output uncertainties ($\sigma_{out} < 2.1$ mg L^{-1} for RF; < 2.0 mg L^{-1} for quantum-enhanced RF at 10x generation), consistent with the dilution-dominated hydrological conditions that homogenize nitrate concentrations. The quantum-enhanced RF again showed slightly stronger uncertainty suppression (-28.7% vs. -7.8% for RF), though the absolute gains were modest given the inherently lower system variability.

These error propagation results validate that the t-SNE-GMM-KNN virtual sample generation strategy not only improves predictive accuracy but also enhances the model's resilience to input measurement uncertainties, provided the generation factor is appropriately calibrated to the underlying data distribution complexity. For high-variability seasons (dry season), moderate generation (4-5x) appears optimal for uncertainty reduction, whereas excessive generation (>8x) may reintroduce variance in classical models. The hybrid quantum-classical architecture demonstrates particular value in stabilizing error propagation under extreme hydrological conditions, where traditional methods may suffer from uncertainty amplification. Error propagation analysis provides a quantitative measurement standard beyond conventional accuracy indicators (such as R^2, RMSE) for evaluating the robustness of the model. This is particularly significant in small sample environment modeling, as the risk of overfitting increases in such cases. Secondly, the excellent uncertainty control ability of quantum-enhanced RF in the dry season validates the theoretical advantages of the quantum-enhanced feature space in capturing heavy-tailed, nonlinear geological chemical processes. Finally, the seasonal uncertainty model reinforces the necessity of a hydrological seasonal hierarchical modeling framework for predicting groundwater nitrate, as a single unified model may not adequately represent the complex error structure under different hydrological conditions.

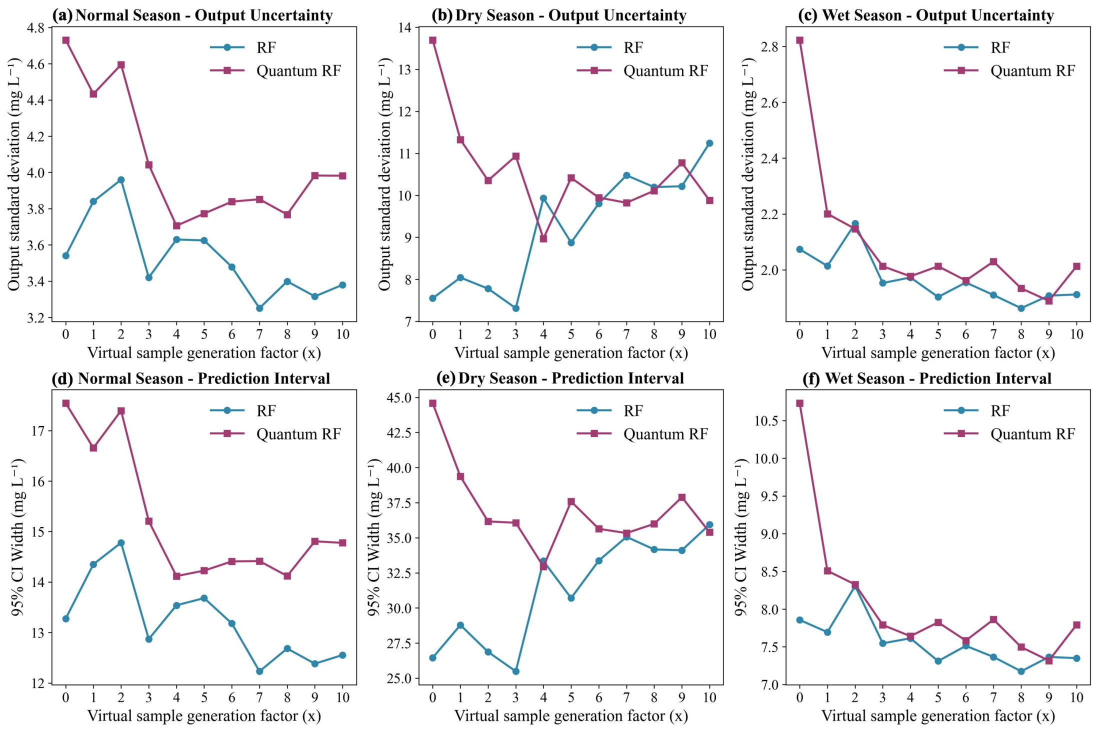


Fig. 25. Variation of output standard deviation of classical Random Forest and quantum-enhanced Random Forest with virtual sample generation multiples under 5% input coefficient of variation in normal, dry and wet seasons.

4.4 Performance analysis of hybrid quantum-classical model

Quantum Machine Learning offers a novel approach to capturing complex non-linear relationships through feature mapping in high-dimensional quantum Hilbert spaces. The hybrid quantum-classical Random Forest produced slight performance improvements in scenarios where original data were scarce or exhibited highly skewed distributions (Lamichhane et al., 2025). When classical feature representation capacity approaches saturation, the Z-feature mapping based on Parameterized Quantum Circuits (PQC) may expose entangled non-linear patterns in the Hilbert space, thereby enhancing feature discriminability (Hong et al., 2025). The gain from this enhancement tends to converge after sufficient virtual sample expansion (Oliveira et al., 2025). In this study, quantum features were generated by analytically calculating the Pauli-Z expectation value $\langle Z \rangle$, completely circumventing hardware noise interference associated with quantum sampling (Gujju et al., 2024). This renders the quantum-enhanced RF practically feasible for small-sample environmental tasks. However, the performance improvement of the quantum-enhanced RF is not universal; in scenarios with high data quality and significant linear relationships during the wet season, the marginal gain of quantum features is limited (Ranga et al., 2024). 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).

4.5 Limitations and future directions

While the 66, 65, and 50 sampling points for the normal, dry, and wet seasons respectively provided sufficient coverage for the 3000-hectare unmanned farm, the spatial representativeness of these discrete sampling points may not fully capture fine-scale heterogeneities in agricultural landscapes where nitrate leaching exhibits strong local-scale variability governed by micro-topography, soil texture transitions, and uneven fertilizer application patterns (Li et al., 2025). The seasonal snapshot approach though capturing critical hydrological transitions, inevitably sacrifices temporal resolution inherent to continuous monitoring systems. The 4-6 month intervals between campaigns may miss episodic recharge events, short-term fertilizer leaching pulses following intensive rainfall, or transient biogeochemical shifts during freeze-thaw cycles that are particularly relevant in the North China Plain's monsoonal climate (Nolte et al., 2025). The single-year observation period (2022-2023) also precludes assessment of inter-annual climatic variability (e.g., El Niño-Southern Oscillation impacts) and long-term legacy nitrogen depletion dynamics. We acknowledge that our model predictions, while robust for the observed seasonal windows, may exhibit reduced reliability during extreme hydrological years or under future climate scenarios where non-stationarity in precipitation patterns could alter nitrate mobilization pathways. Several sources of systematic bias warrant explicit acknowledgment. First, the reliance on irrigation wells for sampling introduces a hydrological connectivity bias: our dataset overrepresents shallow, actively exploited aquifer zones while potentially excluding deeper, stagnant groundwater bodies with distinct nitrate signatures (Jia et al., 2025). Second, the MixSIAR source apportionment relies on literature-derived end-member isotopic signatures that may not fully capture local biogeochemical fractionation specific to the North China Plain's thick vadose zone (Jiang et al., 2026). Third, the virtual sample generation strategy, while effective for addressing data sparsity, may inadvertently amplify representational biases present in the original small sample sets, particularly the overrepresentation of high-concentration outliers during the dry season, potentially skewing model sensitivity toward extreme values (Tanner et al., 2022).

While the hybrid quantum-classical framework demonstrated superior performance within the Xiong’an New Area, its extrapolation validity to hydrogeologically dissimilar regions remains untested and likely limited without domain adaptation (Dai et al., 2025). Recent advances in multi-site deep learning architectures suggest that pooling data from hydrogeologically similar regions could enhance generalizability; however, our single-site experimental design precludes assessment of such transfer learning benefits (Clark et al., 2025; Farahani et al., 2025). Future work should implement leave-one-region-out cross-validation and explore meta-learning approaches to disentangle universal nitrate transport principles from site-specific idiosyncrasies (Yu et al., 2025). Regarding extrapolation to hydrogeologically dissimilar regions, the current framework’s validity is intrinsically linked to the specific hydrogeochemical context of the North China Plain, particularly its thick vadose zone, semi-arid climate, and intensive agricultural land use. Direct model transfer to regions with contrasting characteristics (e.g., karst aquifers with thin soils, humid tropical climates, or nitrate-dominated fertilizer regimes) would likely suffer from domain shift issues (Dai et al., 2025). We emphasize that the proposed t-SNE-GMM-KNN methodology is portable, but model parameters require recalibration using local hydrochemical data. Future implementation should adopt domain adaptation techniques, such as adversarial training or feature alignment between source and target domains. to enable cross-regional application without extensive new sampling campaigns (Zheng et al., 2025). Additionally, meta-learning approaches that leverage multi-site hydrochemical databases could disentangle universal nitrate transport principles from site-specific idiosyncrasies, enhancing the framework's generalizability across diverse hydrogeological settings (Yu et al., 2025).

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. This discrepancy likely stems from the fact that water quality parameters directly reflect the immediate state of the groundwater chemical environment and exhibit direct relationships with nitrate transport and transformation processes, whereas surface remote sensing semantics provide only an indirect characterization (Alam et al., 2025). We acknowledge that relying on field parameters such as EC and TDS to predict nitrate is primarily valuable for understanding hydrochemical correlations rather than reducing monitoring costs if all parameters must be measured alongside nitrate. However, a key practical advantage is that the input water quality parameters(pH, EC, TDS, DO, ORP, etc.) can be rapidly acquired on-site using portable multi-parameter meters(e.g., Hach HQ400) within minutes, whereas nitrate determination typically requires laboratory-based analysis with longer processing times. Therefore, the field-parameter model offers a practical pathway for preliminary screening and real-time decision-making in field surveys when immediate laboratory results are unavailable. 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 season, supporting 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 (Alvarez et al., 2025). These interpretations align 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 (Cai et al., 2025). Ground-truth validation remains essential for establishing definitive relationships between remote sensing signals and subsurface water quality (Tollefson et al., 2025).

5. Conclusions

With the clear goal of solving the practical problems of difficult seasonal prediction, high monitoring cost, and lack of targeted control basis for groundwater nitrate in intensive agricultural areas, this study develops an integrated prediction framework combining hybrid quantum-classical machine learning, advanced virtual sample generation (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.

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 generation 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. Given that field-measurable water quality parameters can be rapidly acquired using portable instruments while nitrate determination requires laboratory analysis, the field-based model provides a practical tool for both diagnosing pollution drivers and enabling rapid on-site nitrate estimation. Following 10-fold virtual sample generation, the AEF model also achieves usable accuracy, with feature importance exhibiting seasonal shifts, validating its distinct role as a cost-effective solution for regional risk assessment in unmonitored areas.

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.

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 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]. 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.

Alam, S.M.K., Li, P., Rahman, M., et al., 2025. Key factors affecting groundwater nitrate levels in the Yinchuan Region, Northwest China: Research using the eXtreme Gradient Boosting (XGBoost) model with the SHapley Additive exPlanations (SHAP) method[J].

Environmental Pollution, 364: 125336.

Liu, M., Geng, D., Wu, L., et al., 2025. The impact of agricultural land use change on water and nitrate fluxes in the deep vadose zone, the North China Plain[J]. Journal of Hydrology: Regional Studies, 2025, 62: 102914.

Xu, G., Su, X., Yuan, Z., et al., 2022. Nitrogen behavior during artificial groundwater recharge through ponds: A case study in Xiong'an New Area[J]. Environmental Geochemistry and Health, 44(8): 2545-2561.

Xiong'an New Area Official Website, 2023. "Geographical Environment and Climatic Characteristics of Xiong'an." *Xiong'an New Area Official Website*, http://www.xiongan.gov.cn/2023-02/27/c_129769113.htm. Accessed 16 Dec, 2025.

Liao, Y.M., Huang, D.P., 2020. Climate Characteristics and Change Trend of Xiongan New Area. Chinese Agricultural Science Bulletin, 36(23): 99-105.

Bai H, Yang H F, Meng R F, et al., 2023. Chemical characteristics and evolution of groundwater in Baoding Plain[J]. Geol. Rev., 69(06): 2216-2228.

Li, X., Liu, M., Min, L., et al., 2025. Nitrogen transport and transformation processes in the typical deep vadose zone in the central North China Plain[J]. Chinese Journal of Eco-Agriculture.

Zhang, Z.J., 2007. Investigation and evaluation of groundwater sustainable utilization in North China Plain[M].Shijiazhuang:Institute of Hydrogeology and Environmental Geology,Chinese Academy of Geological Sciences.

Li, J., Zhu, D., Zhang, S., et al., 2022. Application of the hydrochemistry, stable isotopes and MixSIAR model to identify nitrate sources and transformations in surface water and groundwater of an intensive agricultural karst wetland in Guilin, China[J]. Ecotoxicology and Environmental Safety, 231: 113205.

Huang, P., Chen, J., 2012. Recharge sources and hydrogeochemical evolution of groundwater in the coal-mining district of Jiaozuo, China[J]. Hydrogeology Journal, 20(4): 739-754.

Kim, J.H., Kang, Y.J., Kim, K.I., et al., 2019. Toxic effects of nitrogenous compounds (ammonia, 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.

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 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.

Alvarez, C.I., Ulloa, Vaca.C.A., Echeverria, Llumipanta.N.A., 2025. Machine learning for urban air quality prediction using Google AlphaEarth Foundations satellite embeddings: A case study of Quito, Ecuador[J]. Remote Sensing, 17(20): 3472.

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

Ilyas, M., Niaz, R., Persio, L.D., et al., 2025. A spatiotemporal approach for drought monitoring using Empirical Orthogonal Function (EOF) analysis and neural networks[J]. Earth Systems and Environment: 1-20.

Zhu, J.J., Yang, M., Ren, Z.J., 2023. Machine learning in environmental research: common pitfalls and best practices[J]. Environmental Science & Technology, 57(46): 17671-17689.

Saha, G.K., Rahmani, F., Shen, C., et al., 2023. A deep learning-based novel approach to generate continuous daily stream nitrate concentration for nitrate data-sparse watersheds[J]. Science of the Total Environment, 878: 162930.

Li, J., Liu, G., Shen, Z., 2025. Integrating spatiotemporal variability of non-point source pollution and best management practice efficiency to improve adaptive watershed management[J]. Water Research: 124042.

Bigler, M.C., Brusseau, M.L., Guo, B., et al., 2024. High-resolution depth-discrete analysis of PFAS distribution and leaching for a vadose-zone source at an AFFF-Impacted site[J]. Environmental Science & Technology, 58(22): 9863-9874.

Udu, A.G., Salman, M.T., Ghalati, M.K., et al., 2025. Emerging SMOTE and GAN-variants for Data Augmentation in Imbalance Machine Learning Tasks: A Review[J]. IEEE Access.

Silva, R., Melo-Pinto, P., 2023. t-SNE: A study on reducing the dimensionality of hyperspectral data for the regression problem of estimating oenological parameters[J]. Artificial

Intelligence in Agriculture, 7: 58-68.

Kurniawan, R., Fitriansyah, A., Lestari, F., et al., 2024. Optimizing the Identification of Suitable Congregations for Preachers Using a GMM-PCA-BIC Hybrid Clustering Approach[C]//2024 7th International Conference of Computer and Informatics Engineering (IC2IE). IEEE: 1-6.

Peng, D., Gui, Z., Wei, W., et al., 2025. Sampling-enabled scalable manifold learning unveils the discriminative cluster structure of high-dimensional data[J]. Nature Machine Intelligence: 1-16.

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

Ranga, D., Rana, A., Prajapat, S., et al., 2024. Quantum machine learning: Exploring the role of data encoding techniques, challenges, and future directions[J]. Mathematics, 12(21): 3318.

Adhikari, M., 2022. Hybrid Computing Models Integrating Classical and Quantum Systems for Enhanced Computational Power: A Comprehensive Analysis[J]. Journal of Advanced Computing Systems, 2(12): 1-9.

Alam, M.M.T., Milas, A.S., 2025. Dimensionality Optimized Machine Learning Retrieval of Canopy Chlorophyll, Nitrogen, and Phosphorus from Google Satellite Embeddings[J]. Smart Agricultural Technology: 101601.

Sun, H., Zheng, W., Wang, S., et al., 2024. Variation of nitrate sources affected by precipitation with different intensities in groundwater in the piedmont plain area of alluvial-pluvial fan[J]. Journal of Environmental Management, 367: 121885.

Wang, J., Hao, X., Liu, X., et al., 2025. Groundwater–surface water exchange affects nitrate fate in a seasonal freeze–thaw watershed: Sources, migration and removal[J]. Journal of Hydrology, 654: 132803.

Wang, J., Wang, M., Zhang, C., et al., 2025. Vertical Migration Characteristics and Effect Factors of BaP in Contaminated Soil Under Rainwater Infiltration[J]. Water, Air, & Soil Pollution, 236(14): 1-19.

Wu, H., Song, F., Min, L., et al., 2024. Exploring recharge mechanisms of soil water in the thick unsaturated zone using water isotopes in the North China Plain[J]. Catena, 234: 107615.

Liu, S., Hao, Y., Wang, H., et al., 2023. Bidirectional potential effects of DON transformation in vadose zones on groundwater nitrate contamination: Different contributions to nitrification and denitrification[J]. Journal of Hazardous Materials, 448: 130976.

Ahmed, M.A., Abdel, Samie.S.G., Badawy, H.A., 2013. Factors controlling mechanisms of groundwater salinization and hydrogeochemical processes in the Quaternary aquifer of the Eastern Nile Delta, Egypt[J]. Environmental Earth Sciences, 68(2): 369-394.

Zhang, W., Xin, C., Yu, S., 2023. A review of heavy metal migration and its influencing factors in karst groundwater, Northern and Southern China[J]. Water, 15(20): 3690.

Zhang, J., Zhang, L., Zheng, T., et al., 2025. Tracing Nitrate Contamination Sources and Transformations in a Rural-Urban Karst Groundwater System in North China Using Multiple Isotopes and Simmr Modeling[J]. Water Resources Research, 61(10): e2025WR040156.

Ilyas M, Niaz R, Persio L D, et al., 2025. A spatiotemporal approach for drought monitoring using Empirical Orthogonal Function (EOF) analysis and neural networks[J]. Earth Systems and Environment: 1-20.

Kobak D, Berens P, 2019. The art of using t-SNE for single-cell transcriptomics[J]. Nature communications, 10(1): 5416.

Di Santo D, He C, Chen F, et al., 2024. ML-AMPSIT: Machine learning-based automated multi-method parameter sensitivity and importance analysis tool[J]. Geoscientific Model Development Discussions, 2024: 1-37.

Huang C, Tong J, Ye M, 2021. Global sensitivity analysis for a prediction model of soil solute transfer into surface runoff[J]. Journal of Hydrology, 599: 126342.

Anderson G J, Lucas D D, 2018. Machine learning predictions of a multiresolution climate model ensemble[J]. Geophysical Research Letters, 45(9): 4273-4280.

Dega S, Dietrich P, Schroen M, et al., 2023. Probabilistic prediction by means of the propagation of response variable uncertainty through a Monte Carlo approach in regression random forest: Application to soil moisture regionalization[J]. Frontiers in Environmental Science, 11: 1009191.

Xiong'an Urban Geology Research Center,China Geological Survey.Annual Report on Natural Resources Monitoring of Xiong'an New Area (2022)[R].2023.

Ma Z,Huang Q B,Lin L J,et al., 2022.Practice and application of multi-factor urban geological survey in Xiongan New Area[J].North China Geology,45(1):58-68.

Sun Q J, 2024. Study on Infiltration Promotion Mechanisms of Groundwater Artificial Recharge in Xiong'an New Area[D]. JilinUniversity.(in Chinese)

Xiong X, Hao G, Xu L, et al., 2025. Characteristics of spatial distributions for nitrate and traceability of pollution in shallow groundwater of the Qingshui River Basin, China[J]. Hydrology Research, 56(9): 920-936.

Xu G, Su X, Yuan Z, et al., 2022. Nitrogen behavior during artificial groundwater recharge through ponds: A case study in Xiong’an New Area[J]. Environmental geochemistry and health, 44(8): 2545-2561.

Hou X, Peng L, Zhang Y, et al., 2025. A Data-Driven Method for Determining DRASTIC Weights to Assess Groundwater Vulnerability to Nitrate: Application in the Lake Baiyangdian Watershed, North China Plain[J]. Applied Sciences, 15(5): 2866.

Liu N, Chen M, Gao D, et al., 2025. Identification of hydrogeochemical processes in shallow groundwater using multivariate statistical analysis and inverse geochemical modeling[J]. Environmental Monitoring and Assessment, 197(2): 135.

Li X, Zhou J, Zhou W, et al., 2025. Hydrochemical Characteristics of Shallow Groundwater and Analysis of Vegetation Water Sources in the Ulan Buh Desert[J]. Water, 17(21): 3058.

Nolte A, Heudorfer B, Bender S, et al., 2025. Multi-site deep learning for groundwater level prediction across global datasets: toward scalable applications under data scarcity[J]. Journal of Hydroinformatics, 27(10): 1632-1651.

Jia H, Qian H, 2025. Groundwater nitrate response to hydrogeological conditions and socioeconomic load in an agriculture dominated area[J]. Scientific Reports, 15(1): 1315.

Jiang H, Geng D, Fu J, et al., 2026. Quantifying nitrate dynamics in deep vadose zone under irrigated farmland in the North China Plain: Insights from continuous in-situ monitoring[J]. Agriculture, Ecosystems & Environment, 396: 109956.

Tanner K B, Cardall A C, Williams G P, 2022. A spatial long-term trend analysis of estimated chlorophyll-a concentrations in Utah Lake using Earth observation data[J]. Remote Sensing, 14(15): 3664.

Dai Z, Zhan C, Yin H, et al., 2025. Incorporating deep learning into hydrogeological modeling: Advancements, challenges, and future directions[J]. Journal of Geophysical Research: Machine Learning and Computation, 2(2): e2025JH000703.

Farahani M A, Wood A W, Tang G, et al., 2025. Calibrating a large-domain land/hydrology process model in the age of AI: the SUMMA CAMELS emulator experiments[J]. Hydrology and Earth System Sciences, 29(18): 4515-4537.

Clark S R, Jaffres J B D, 2025. Associations between deep learning runoff predictions and hydrogeological conditions in Australia[J]. Journal of Hydrology, 651: 132569.

Yu R, Chen S, Xie Y, et al., 2025. A survey of foundation models for environmental science[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore: Springer Nature Singapore: 39-57.

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.

Su Y, Wu Z, Zheng X, et al., 2025. Harmonizing remote sensing and ground data for forest aboveground biomass estimation[J]. Ecological Informatics, 86: 103002.

Ma Z, Ye C, Lu C, et al., 2025. A Robust Gaussian Process Paradigm for Predictive Modeling on Small Data sets in Environmental Science: A Case Study in Ballasted Flocculation[J]. Environmental Science & Technology, 60(1): 748-759.

Sadik L, Samui P, 2024. Uncertainty-aware prediction of bearing capacity of shallow foundations resting on cohesionless soils using Bayesian regression[J]. Geotechnical and Geological Engineering, 42(5): 3919-3937.

Charizanos G, Demirhan H, 2023. Bayesian prediction of wildfire event probability using normalized difference vegetation index data from an Australian forest[J]. Ecological Informatics, 73: 101899.

Addy J W G, MacLaren C, Lang R, 2024. A Bayesian approach to analyzing long-term agricultural experiments[J]. European Journal of Agronomy, 159: 127227.

Huang H, Huang J, Wu Y, et al., 2023. The improved winter wheat yield estimation by assimilating GLASS LAI into a crop growth model with the proposed Bayesian posterior-based ensemble Kalman filter[J]. IEEE Transactions on Geoscience and Remote Sensing, 61: 1-18.

Weller D L, Murphy C M, Johnson S, et al., 2022. Land use, weather, and water quality factors associated with fecal contamination of northeastern streams that span an urban-rural gradient[J]. Frontiers in Water, 3: 741676.

Liu M, Min L, Wu L, et al., 2022. Evaluating nitrate transport and accumulation in the deep vadose zone of the intensive agricultural region, North China Plain[J]. Science of the Total Environment, 825: 153894.

Miao P, Zhu X, Zhang S, et al., 2023. Long-term high nitrogen surplus in intensive apple-planting regions results in huge legacy nitrogen in the vadose zone: Potential or real risk to the groundwater?[J]. Agriculture, Ecosystems & Environment, 357: 108682.

Williams M R, Ford W I, Mumbi R C K., 2023. Preferential flow in the shallow vadose zone: Effect of rainfall intensity, soil moisture, connectivity, and agricultural management[J]. Hydrological Processes, 37(12): e15057.

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.

Liu M, Geng D, Wu L, et al., 2025. The impact of agricultural land use change on water and nitrate fluxes in the deep vadose zone, the North China Plain[J]. Journal of Hydrology: Regional Studies, 62: 102914.

Yang W, Long D, Scanlon B R, et al., 2022. Human intervention will stabilize groundwater storage across the North China Plain[J]. Water Resources Research, 58(2): e2021WR030884.

Feng W, Wang S, Tan K, et al., 2024. Simulation of spatial and temporal variation of nitrate leaching in the vadose zone of alluvial regions on a large regional scale[J]. Science of The Total Environment, 916: 170114.

GB/T 14848-2017 (2017). Standard for Groundwater Quality. National Standards of the People's

Republic of China.

Saberian M, Zafarmomen N, Neupane A, et al., 2025. HydroQuantum: A new quantum-driven Python package for hydrological simulation[J]. Environmental Modelling & Software: 106736.

Zheng Y, Zhang X, Zhou Y, et al., 2025. Deep representation learning enables cross-basin water quality prediction under data-scarce conditions[J]. NPJ Clean Water, 8(1): 33.