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. 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). 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, 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). 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). 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. 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., 2026). 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).

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

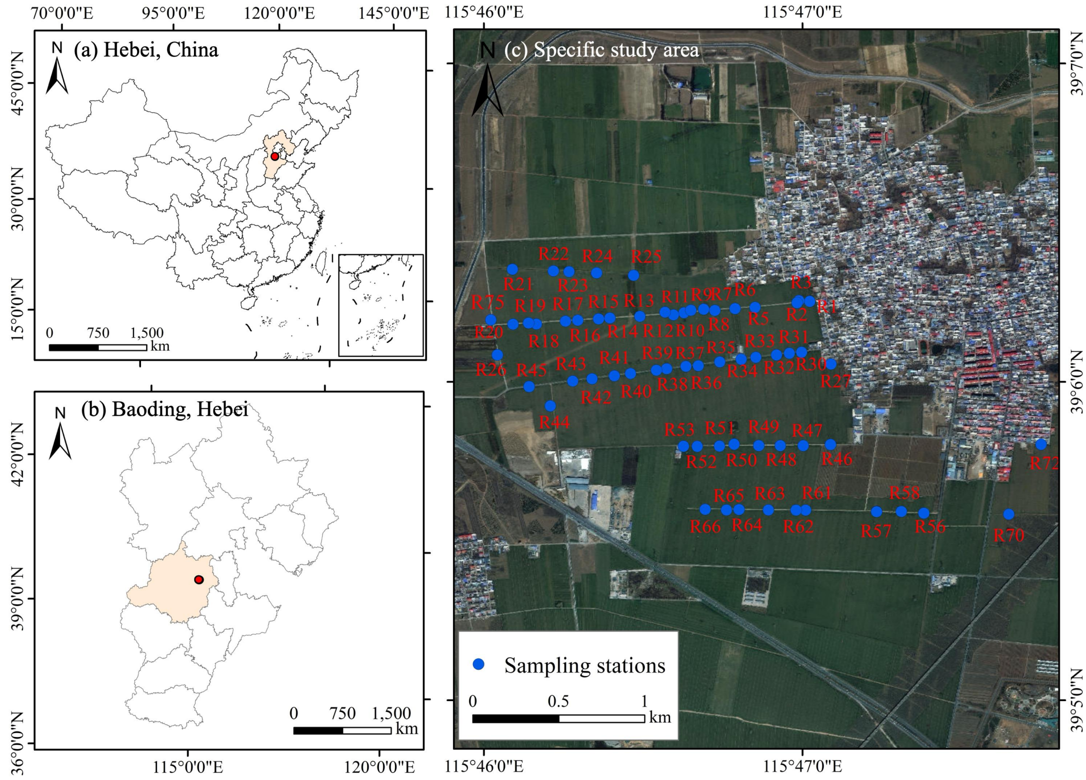

Fig.1. Study area map showing the sampling location and sampling points. The administrative boundaries are based on the standard map (approval No. (2023)2766) from the Standard Map Service of the Ministry of Natural Resources of China; the satellite basemap is from Google Earth (Imagery © 2026 Airbus, CNES/Airbus, MaxarTechnologies).

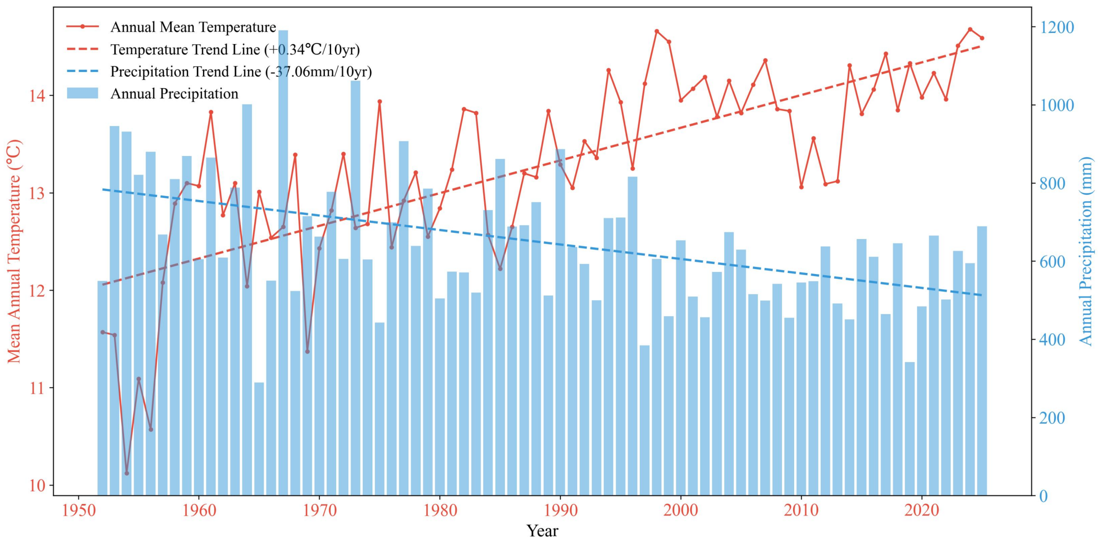

Fig.2. Temperature and precipitation trends in study area (1952-2025).

Cultivated land, predominantly dryland, occupies a large proportion of Xiong'an New Area, where traditional practices involve high application rates of nitrogen fertilizer and manure. The study site follows these practices, with an annual nitrogen application rate of 540-660 kg (N) ha^{-1} yr^{-1}, mainly as urea (46% N), posing a high risk of groundwater nitrogen pollution; the dense surrounding rural population adds further input from domestic sewage discharge. Irrigation is scheduled by crop phenology: wheat receives muddy water irrigation before sowing and at the overwintering, reviving, and jointing stages, while maize receives a single irrigation after sowing. The region has a temperate continental monsoon climate, with a mean annual temperature of 12.6°C (January average -5°C; July average 26.1°C). The mean annual precipitation is 480.8 mm, of which nearly 80% falls between June and September (Fig. 2). The multi-year average evaporation is 1,661.1 mm and the average water surface evaporation is 1,761.7 mm. The mean annual sunshine duration and wind speed are 2,335.2 h and 1.7 m s^{-1}, respectively, and the frost-free period averages 204 days. The soil texture is predominantly silt loam, with high-clay interlayers (clay and silty clay) at depths of 2~8.5 m. In the thick vadose zone, nitrogen exists predominantly as organic nitrogen (~97% of total nitrogen), and the shallow layer at 3-6 m depth stores about half of the total nitrate reserves in the North China Plain (Li et al., 2025). Groundwater occurs mainly in Quaternary unconsolidated porous aquifers (sampling well depths: 70~120 m); the shallow aquifer group consists of silty fine to medium sand with medium water abundance (single-well yields: ~1,000-3,000 m3/d). The depth to the shallow groundwater table ranges from 5.0 to 35.0 m, with fluctuations driven by precipitation infiltration and agricultural extraction. Precipitation is the primary recharge source for farmland groundwater, and artificial extraction for irrigation is the main discharge pathway. Detailed climate and hydrological data are given 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. In-situ physicochemical parameters (T, pH, TDS, DO, EC, and ORP) were measured using a Hach HQ400 multi-parameter meter (Li et al., 2022), and major ions, nitrogen species, and stable isotopes (δ^2H, $\delta^{17}O$, $\delta^{18}O$, $\delta^{15}N$-NO_3^-, $\delta^{18}O$-NO_3^-) were analyzed in the laboratory following standard protocols. Detailed sampling procedures, quality control measures, and analytical methods are provided in the Supplementary Material (Texts S1 and S2).

2.2.2 Google AlphaEarth Foundation

To compare with predictions based on in-situ field sampling data and to validate the accuracy of remote sensing-based nitrate prediction, this study incorporates the Google AlphaEarth Foundation (AEF) dataset. AEF provides a 64-dimensional surface semantic embedding vector (A00-A63) at 10 m spatial and annual temporal resolution, generated by pre-training on multi-source satellite imagery (e.g., Sentinel-2, Landsat). These embeddings implicitly encode environmental semantics such as land cover, vegetation dynamics, soil moisture, and human activity intensity (Alvarez et al., 2025; Tollefson et al., 2025). The 64-dimensional AEF vectors were extracted at each sampling point via the Google Earth Engine platform, and Principal Component Analysis (PCA) was applied to retain the minimum number of components explaining at least 95% of the cumulative variance. The PCA-reduced features served as model inputs. Details of the AEF data processing workflow are provided in the Supplementary Material (Text S3).

2.3 Statistical methods for source apportionment and driver analysis

2.3.1 MixSIAR model and isotopic composition of nitrate sources

MixSIAR incorporates prior information on end-member values and their errors and distributions, and estimates the proportional contribution of each source via Markov Chain Monte Carlo (MCMC) iteration (Stock et al., 2018). In this study, it was applied to apportion five potential NO_3^- sources: precipitation (NP), soil organic nitrogen (SON), synthetic NH_4^+ fertilizer (NHF), synthetic NO_3^- fertilizer (NOF), and domestic sewage & manure (DSM), with the source end-member values listed 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

2.3.2 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). 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), $0.001 \leq p < 0.01$ (very significant), $0.01 \leq p < 0.05$ (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).

2.3.3 Bayesian linear regression model

Bayesian linear regression was used to infer the posterior distributions of model parameters based on Bayes' theorem. This approach provides full uncertainty quantification of parameter effects and reduces the risk of overfitting in small-sample scenarios. The model is expressed as:

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

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; ϵ_i is the residual term, which follows a normal distribution with mean 0 and variance σ^2.

Weakly informative priors were assigned to all parameters to ensure stable fitting under small-sample conditions (Charizanos et al., 2023): a Student-t prior for the intercept, standard normal priors for the coefficients, and an exponential prior for the residual standard deviation. Posterior inference was performed via Markov Chain Monte Carlo (MCMC) sampling using the brms package in R (Addy et al., 2024), with 4 chains of 2000 iterations each (1000 warm-up). Model performance was assessed using the Bayesian coefficient of determination (Bayesian R^2), reported as the posterior mean with 95% credible interval. To quantify the directional effect of each predictor, the probability of direction (pd) was calculated as the proportion of the posterior distribution of β_j on the same side of 0 (Weller et al., 2022). pd ranges from 0.5 to 1, with higher values indicating stronger evidence for a directional effect.

2.4 t-SNE-GMM-KNN virtual sample generation and quality assessment

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

To mitigate overfitting and poor generalization caused by data sparsity and skewed distributions in small-sample modeling, we propose a three-stage virtual sample generation strategy: t-SNE-Gaussian Mixture Sampling with KNN inverse mapping, which preserves the non-linear manifold structure and multi-modal distribution of the original high-dimensional feature space while generating physically plausible and statistically consistent virtual samples. Virtual samples are generated independently for each hydrological season (normal, dry, wet) within the hydrochemical feature space, i.e., the space spanned by the in-situ water-quality parameters and nitrate concentration, without temporal extrapolation or cross-season fusion, thereby augmenting sample density within each season while preserving the inherent seasonal hydrochemical heterogeneity. The 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). 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). The perplexity was set to 10, which is suitable for small sample sizes (n<100) and balances local neighborhood preservation against 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).

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):

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

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

4. Virtual sample generation and inverse mapping

Virtual points were randomly sampled from the joint probability distribution of the optimal GMM in the 2D t-SNE latent space, thereby naturally inheriting the multi-modality and covariance structure of the original data. Generation factors of 1-10× the original sample size were tested, with 10× (n_virtual=n_original×10) yielding optimal model performance. The sampled points were then mapped back to the original 8-dimensional hydrochemical feature space (including NO_3^-) using a K-Nearest Neighbors (KNN) regressor fitted on the original t-SNE embeddings (inputs) and the original standardized hydrochemical data (outputs) (Niu et al., 2025). The regressor was configured with n_neighbors=min(5, n_original -1), balancing inverse-mapping smoothness against local fidelity and avoiding over-smoothing of extreme values. After inverse standardization, each virtual sample contains the synchronized values of all input variables and the target NO_3^- concentration, ensuring the one-to-one predictor-target correspondence required for subsequent RF modeling (hydrochemical parameters as X, NO_3^- as y).

5. Physical constraints and quality control

For the target variable NO_3^-, a non-negativity constraint was imposed post-generation to prevent non-physical values from the regression approximation, while other variables were allowed to fluctuate within reasonable ranges without hard clipping. The consistency between virtual and measured samples was validated by comparing their means, standard deviations, coefficients of variation, extreme-value ranges, and boxplot distributions, confirming that the generated data introduce no systematic bias or outliers. The virtual samples, each containing both predictor variables and nitrate concentration, were concatenated with the measured data to form an augmented training set, in which the 7 water-quality parameters (pH, T, EC, DO, ORP, Salt, TDS) served as inputs (X) and NO_3^- as the target (y). This increases the training sample size while preserving the geochemical integrity of feature-target relationships, thereby mitigating small-sample bias and improving the capture of heavy-tailed distributions.

This method offers four advantages: (1) t-SNE effectively captures local neighborhood structures and separates implicit subgroups under different hydrological conditions; (2) GMM provides a probabilistic generative framework that supports reasonable extrapolation of heavy-tailed distributions and extreme values; (3) the KNN-based inverse mapping requires no large training datasets, making it suitable for small-sample scenarios (Tang et al., 2022; Razavi-Termeh et al., 2024); and (4) joint modeling of predictors and target ensures that the generated nitrate concentrations remain chemically consistent with the accompanying hydrochemical indicators. The physical plausibility and distributional consistency of the virtual samples were validated using the Kolmogorov-Smirnov (KS) test, Jensen-Shannon (JS) divergence, and Maximum Mean Discrepancy (MMD), which assess distributional fidelity from local, overall, and high-dimensional perspectives, respectively.

2.4.2 Validation metrics for virtual sample quality assessment

1. Kolmogorov-Smirnov (KS) Test

The two-sample KS test was employed to quantify the maximum vertical distance between empirical cumulative distribution functions (ECDFs) of observed and virtual samples (Li et al., 2026). For each hydrochemical parameter, the test statistic D_{KS} is defined as:

$$D_{KS} = sup_x | F_n(x) - F_m(x) | \quad (3)$$

where F_n and F_m represent the ECDFs of observed (n samples) and virtual (m samples) datasets, respectively. The corresponding P-value tests the null hypothesis H_0 that both samples derive from the same underlying distribution.

Following established criteria for environmental data synthesis and geochemical validation (Farnia et al., 2023; Tung et al., 2023), we adopted D_{KS}<0.15 as the strict threshold for high fidelity and D_{KS}<0.20 as the acceptable limit for parameters with inherent geochemical heterogeneity. The KS test has been widely applied in hydrological studies to validate the consistency of generated synthetic hydrographs (Michalek et al., 2023; Li et al., 2023).

2. Jensen-Shannon (JS) Divergence

To measure the symmetric similarity between probability distributions in a bounded metric [0, 1], we computed the JS divergence between the observed distribution P and virtual distribution Q (Van et al., 2023):

$$JS(P \| Q) = \frac{1}{2} KL(P \| \mathcal{M}) + \frac{1}{2} KL(Q \| \mathcal{M}) \quad (4)$$

where $\mathcal{M}$=1/2(P+Q) represents the midpoint distribution, and KL denotes the Kullback-Leibler divergence. Unlike the asymmetric KL divergence, JS divergence is symmetric and mathematically bounded, making it particularly suitable for evaluating generative models in environmental applications (Balogun et al., 2024).

The JS divergence was calculated using histogram-based probability density estimation (20 bins) with Laplace smoothing to handle zero-frequency bins. A threshold of JS<0.05 was established to indicate high distributional consistency (Farnia et al., 2023). Values between 0.05-0.10 indicate moderate similarity, and JS>0.10 suggest significant distributional drift that may compromise model training integrity.

3. Maximum Mean Discrepancy (MMD)

Given the high-dimensional nature of hydrochemical feature space (d=8 parameters), we employed MMD in the Reproducing Kernel Hilbert Space (RKHS) to capture non-linear distributional differences without parametric assumptions (Kalinin et al., 2025). Using the Radial Basis Function (RBF) kernel $k(x,y)=\exp\left(-\|x-y\|^2/2\sigma^2\right)$ with bandwidth σ set via the median heuristic, the squared MMD between observed samples $X=\{x_i\}_{i=1}^{n}$ and virtual samples $Y=\{y_j\}_{j=1}^{m}$ was estimated as (Li et al., 2025):

$$MMD[X,Y]=\left[\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i,x_j)-\frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i,y_j)+\frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i,y_j)\right]^{\frac{1}{2}} \quad (5)$$

MMD has demonstrated superior performance in validating synthetic environmental data compared to parametric tests, particularly for small-sample geochemical datasets where underlying distribution assumptions are often violated (An et al., 2025).

2.5 Hybrid quantum-classical random forest

In this study, Random Forest (RF) was adopted as the baseline model. As an ensemble learning method, RF builds numerous decision trees through bootstrap sampling and random feature selection, and integrates their predictions (Abderzak et al., 2025). This approach effectively suppresses overfitting and improves generalization, making it well-suited for environmental data modeling with small samples, high dimensionality, non-linearity, and multicollinearity. 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. Feature importance was quantified by the mean decrease in Gini impurity to identify the critical hydrogeochemical drivers (Kaur et al., 2025). A hybrid quantum-enhanced Random Forest model was constructed by integrating quantum feature enhancement with classical RF. Standardized input features are encoded by a Parameterized Quantum Circuit (PQC) into quantum features with non-linear entanglement properties (Naresh et al., 2025), which are concatenated with the original features and fed into an RF regressor (Lamichhane et al., 2025). This hybrid framework retains the ensemble-learning strengths of classical RF while improving the capture of complex non-linear relationships in small-sample datasets, without relying on physical quantum hardware.

(1) Quantum Feature Encoding

Classical data were mapped into a quantum Hilbert space via single-qubit Z-gate and two-qubit CZ-gate entanglement operations, using the ZFeatureMap provided by Qiskit as the encoding circuit (Vedavyasa et al., 2025). The number of qubits was set equal to the number of input features (7 qubits for the in-situ hydrochemical parameters: pH, T, EC, DO, ORP, Salt, TDS), and the repetition layers (reps) were set to 1 to avoid circuit-depth-induced noise while maintaining expressibility. Its Hamiltonian form is given by (Khalil et al., 2025):

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

where $U_{ZMap}(x)$ is the unitary operator of the encoding circuit, x is the normalized input feature vector, R is the number of repetition layers, d is the number of qubits, Hi the Hadamard gate on the i-th qubit, $S\subseteq\{1,\ldots,d\}$ is the subset of qubit indices involved in entanglement, $\phi S(x)$ is the data-dependent rotation angle, and Zj is the Pauli-Z operator 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):

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

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}.

(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::

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

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 Evaluation methods and prediction process

2.6.1 SHAP analysis

This study adopts the SHapley Additive exPlanations (SHAP) method for local and global explainability analysis (Merabet et al., 2025). Through three typical visualization methods, namely summary plot, dependence plot, and waterfall plot, the following are revealed respectively: (1) The overall ranking and distribution of feature importance across all samples (global perspective); (2) The nonlinear relationship or interaction effects between a single predictor variable and the predicted nitrate concentration (conditional dependence); (3) The contribution decomposition of each feature in the prediction result of a representative sample (local attribution) (Alam et al., 2025; Li et al., 2024; Hollmannet al., 2025).

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

Given the limited sample size in each hydrological season, model performance was evaluated using Leave-One-Out Cross-Validation (LOOCV) (Ren et al., 2021). Prior to LOOCV, hyperparameters were tuned via 5-fold cross-validation grid search (n_estimators [100, 150], max_depth [5, 6, 7], min_samples_split [4, 6]) and then fixed for the LOOCV evaluation. To assess prediction uncertainty, 90% prediction intervals were estimated using quantile regression forests, with coverage probability calculated to validate their reliability; in addition, bootstrap resampling (1000 iterations) was used to derive 95% confidence intervals for all evaluation metrics, which are reported alongside the LOOCV metrics. Model accuracy was quantified by R^2, RMSE, and MAE (Gul et al., 2025).

2.6.3 Model robustness and uncertainty analysis

To quantitatively evaluate the response of the model's nitrate concentration predictions to perturbations in the input hydrochemical parameters, and to verify the robustness of the feature importance ranking, a multi-dimensional sensitivity analysis was conducted, including a SHAP-based global sensitivity index, one-at-a-time (OAT) perturbation tests, and elasticity coefficient calculations (Di et al., 2025). 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 for feature i is defined as the normalized mean absolute SHAP value:

$$S_i=\frac{\frac{1}{n}\sum_{j=1}^{n}|\phi_i^{(j)}|}{\sum_{k=1}^{p}\frac{1}{n}\sum_{j=1}^{n}|\phi_k^{(j)}|} \tag{9}$$

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:

Second, to assess the local stability of the model, an OAT perturbation test was conducted by varying each target feature within ±30% of its measured range while fixing the other inputs, and the perturbation sensitivity was quantified as the relative change rate of the model's R^2 score across perturbation levels (Anderson et al., 2018). Finally, the elasticity coefficient, defined as the ratio of the percentage change in the predicted output to the percentage change in the input, was used to quantify the directional response of the predictions to per-unit changes in each parameter. An absolute value greater than 1 indicates an elastic response, while a value less than 1 indicates an inelastic response. Monte Carlo simulation was specifically used to analyze the propagation of input measurement errors to the final predictions (Dega et al., 2023).

2.6.4 Standardized prediction workflow

To systematically evaluate how data inputs, virtual sample generation, and modeling strategy jointly affect seasonal nitrate prediction, we established a standardized, fully reproducible pipeline (Fig. 3), in which the output of each step serves as the direct input of the next: (1) Data inputs and preprocessing. Two complementary input feature sets were prepared from the seasonally partitioned observations (66, 65, and 50 samples for the normal, dry, and wet seasons, respectively): (i) in-situ field water-quality parameters and (ii) AlphaEarth Foundation (AEF) embeddings compressed by PCA. Both sets were Z-score standardized independently per season; the raw observations were used without interpolation or smoothing to preserve intrinsic variability. These standardized features, together with the measured NO_3^- concentrations, constitute the sole data basis for all subsequent steps. (2) Virtual sample generation and validation. Each standardized seasonal dataset was fed into the t-SNE–GMM–KNN generator to produce virtual samples at generation factors of 1-10×. Only virtual datasets whose distribution consistency and physical plausibility passed the validation tests were allowed to enter model training, thereby explicitly linking generation quality to downstream evaluation. (3) Model training. Classical Random Forest (RF) and quantum-enhanced RF were trained under unified hyperparameters on two tracks of training sets built from the validated samples: the original samples alone, and the original samples combined with 1-10× virtual samples. For the quantum-enhanced RF, input features were encoded via parameterized quantum circuits to generate ⟨Z⟩ quantum features, which were concatenated with the original features (2n inputs in total). (4) Model evaluation. Every trained model was evaluated using leave-one-out cross-validation on the measured samples. (5) Interpretability analysis. SHAP-based multi-scale interpretation (summary, dependence, and waterfall plots) was conducted and cross-verified against Bayesian modeling and Pearson correlation analysis.

All computations were performed on an Apple M2 Pro (10-core CPU, 16 GB RAM) using Python 3.9, with scikit-learn 1.7.2, qiskit 1.4.5, shap 0.49.1, pandas 2.3.3, numpy 2.2.6, scipy 1.15.3, joblib 1.5.2, and matplotlib 3.9.0. All random operations were seeded (random_state=42) to ensure deterministic results. The complete source code, covering quantum ⟨Z⟩ feature extraction, model training, LOOCV evaluation, bootstrap uncertainty quantification, and sensitivity analysis, is publicly available at github.

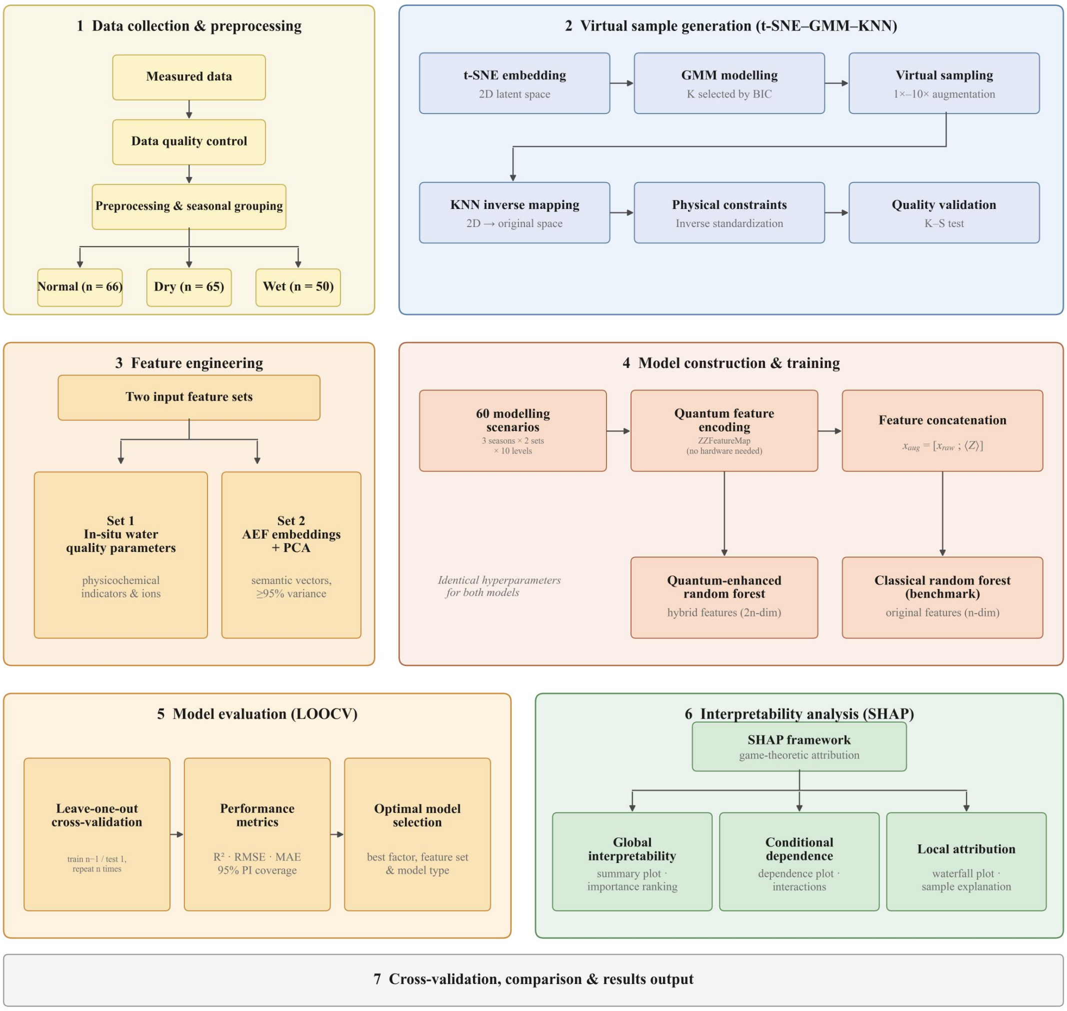

Fig.3. Process diagram for constructing prediction framework.

3. Results

3.1 Seasonal hydrochemical controls of nitrate distribution in farmland groundwater

3.1.1 Hydrochemical characteristics and water types

Groundwater pH was weakly alkaline in the normal water period with minimal variation and near-neutral in the dry and wet seasons, with a minimum of 5.77 in the wet season indicating occasional acidic water (Table S1). Temperature varied markedly among seasons but remained stable within each (CV ≈ 0.06). EC, salinity, and TDS followed consistent seasonal patterns, peaking in the dry season and reaching minima in the wet season. Redox indicators were highly variable: DO was slightly higher in the wet season, whereas ORP remained low across all seasons with large standard deviations. Among major ions, Ca^{2+}, Mg^{2+}, Na^{+}, Cl^{-}, and SO_4^{2-} all peaked in the dry season, while HCO_3^- was highest in the wet season and K^+ remained low. NO_3^- concentrations were highest in the dry season and lowest in the wet season, far exceeding those of NO_2^- and NH_4^+. Most variables were right-skewed, with notable extreme values for NO_3^- in the dry season (maximum=358.58, mean = 42.93 mg L^{-1}), Cl^- in the dry season (maximum=241.36, mean=24.90 mg L^{-1}), and F^- in the normal period (maximum=13.17, mean=3.70 mg L^{-1}). The Piper diagram (Fig. 4) shows that dry-season samples cluster tightly in the Ca-HCO_3^- facies, indicating dominant control by carbonate dissolution. In the wet season, the Ca-Mg-HCO_3^- type remains dominant but some samples shift toward sulfate and chloride types, reflecting leaching inputs of surface pollutants via rainfall infiltration. In the normal season, the hydrochemical types are most dispersed, showing mixed HCO_3^--Cl facies. Overall, the groundwater hydrochemistry in the study area is jointly controlled by precipitation-evaporation dynamics and carbonate weathering.

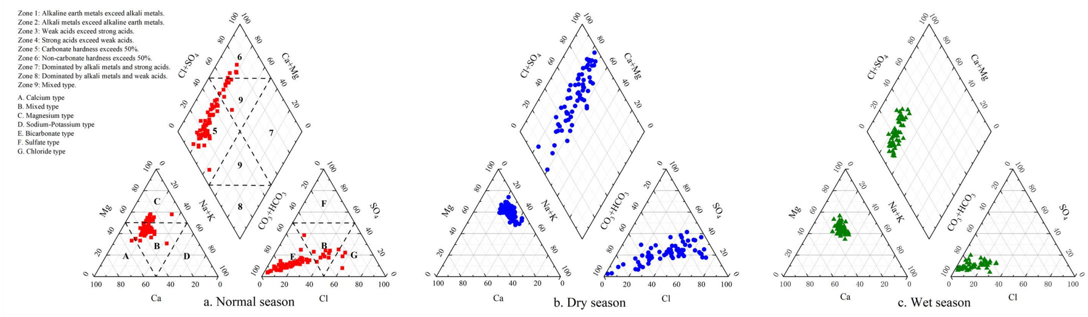

Fig.4. Piper diagram classifying the hydrochemical facies of the analyzed groundwater.

3.1.2 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 $\gamma(Na^+ + K^+)$ 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 $\gamma(Ca^{2+} + Mg^{2+})$ and γHCO_3^-, and between $\gamma(Ca^{2+} + Mg^{2+})$ and $\gamma(HCO_3^- + SO_4^{2-})$ (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 $\gamma(SO_4^{2-} + Cl^-)$ 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.

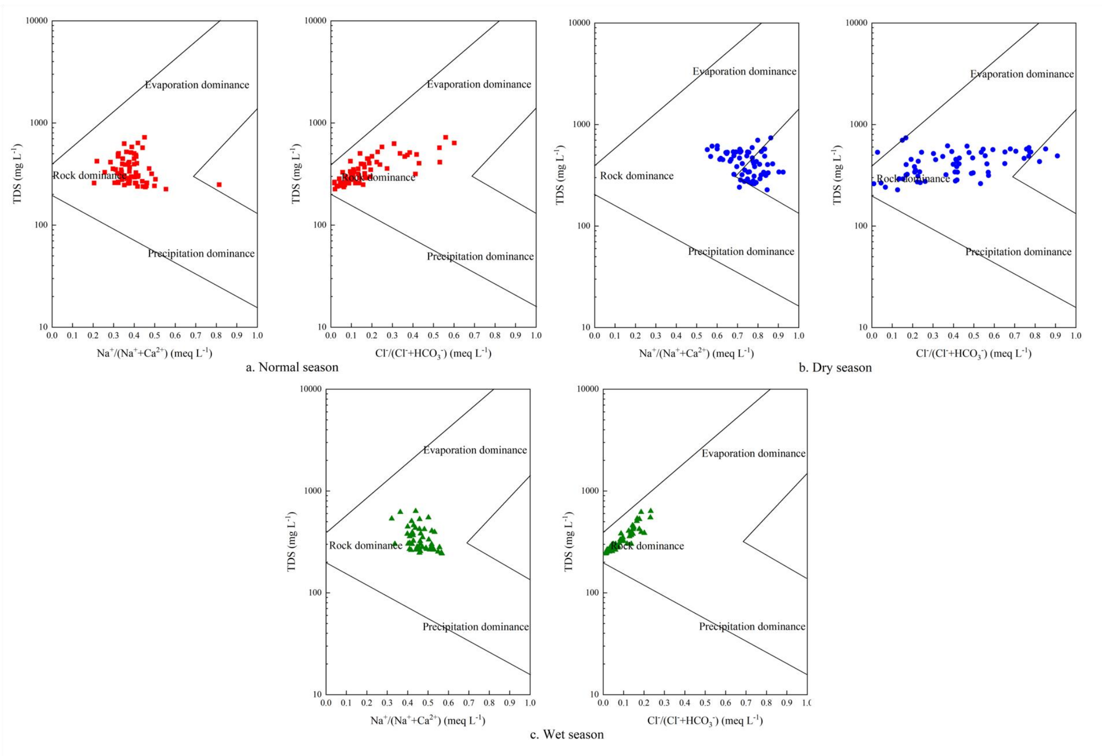

Fig.5. Gibbs diagrams of the groundwater samples in different hydrological seasons.

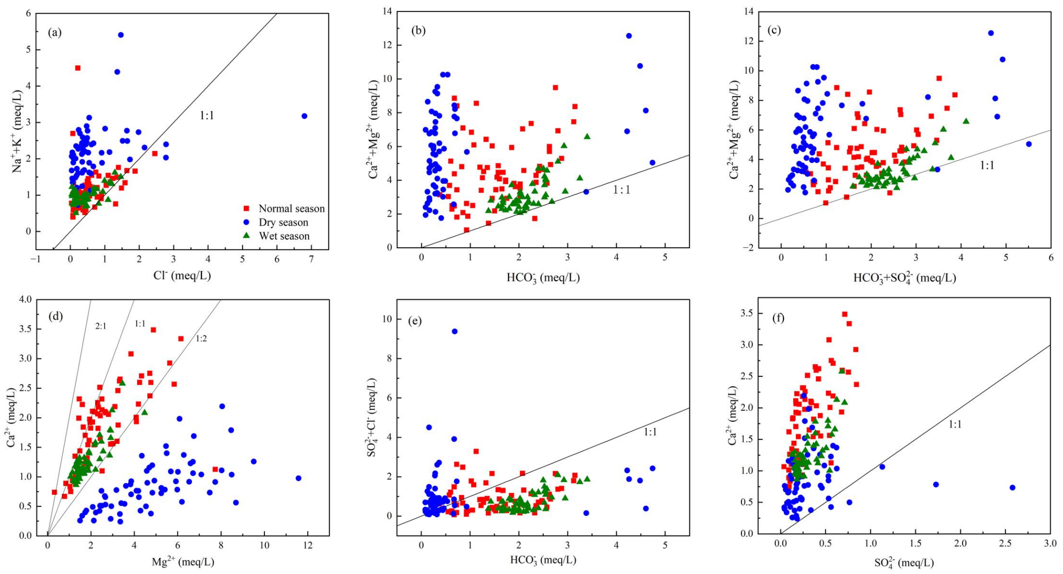

Fig.6. Plots of ion ratio relationship.

The Chloro-Alkaline Index (CAI) was used to assess cation exchange and adsorption between groundwater and sediments, where negative values indicate exchange and more negative values reflect stronger intensity. This was further examined through the relationship between [$\gamma(Ca^{2+}) + \gamma(Mg^{2+}) - \gamma(HCO_3^-) - \gamma(SO_4^{2-})$] and [$\gamma(Na^+) - \gamma(Cl^-)$] (Fig. 7). Samples from the normal and wet seasons plotted near the slope= -1 line (slopes of -1.52 and -1.36, respectively), confirming active cation exchange consistent with the CAI results; exchange was most pronounced in the rainy season (R^2=0.41), leading to Na^+ enrichment. In contrast, the dry-season slope of 0.55 indicates negligible cation exchange during this period.

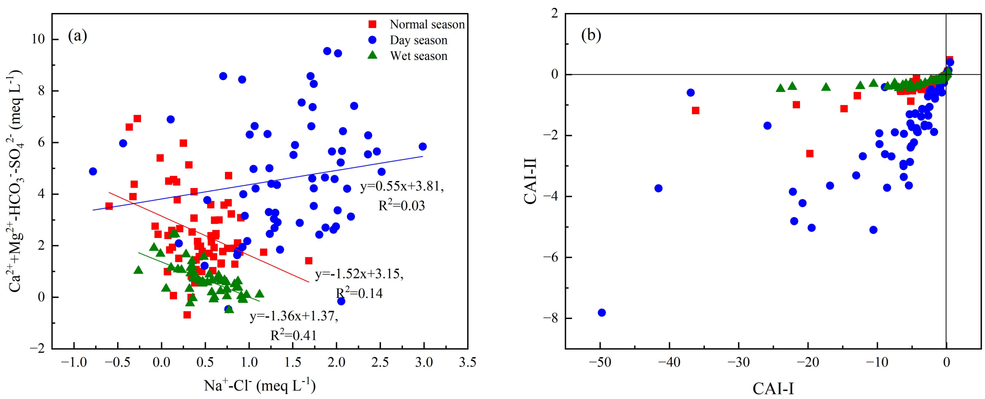

Fig.7. Relationship diagram of groundwater (Ca^{2+}+Mg^{2+}-SO_4^{2-}-HCO_3^-) and (Na^+-Cl^-) along with CAI-1 and CAI-2 correlation diagrams.

3.1.3 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.

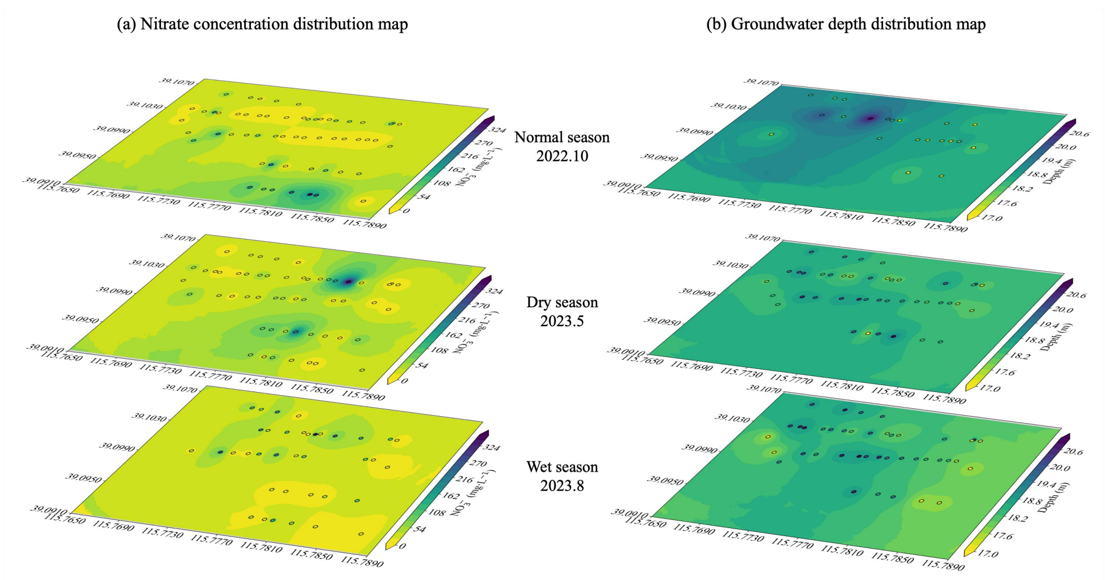

Fig. 8. Spatial distribution of nitrate concentration and groundwater depth in different seasons. From top to bottom: normal water period, dry season, and wet season.

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.

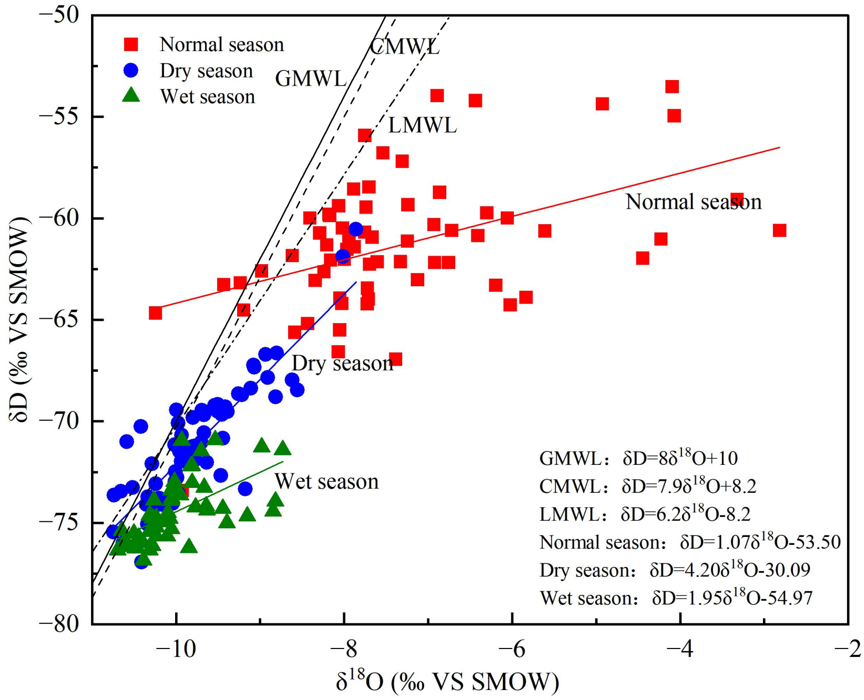

Fig.9. $\delta^{18}O/\delta D$ relationship of groundwater samples in different hydrological seasons.

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.

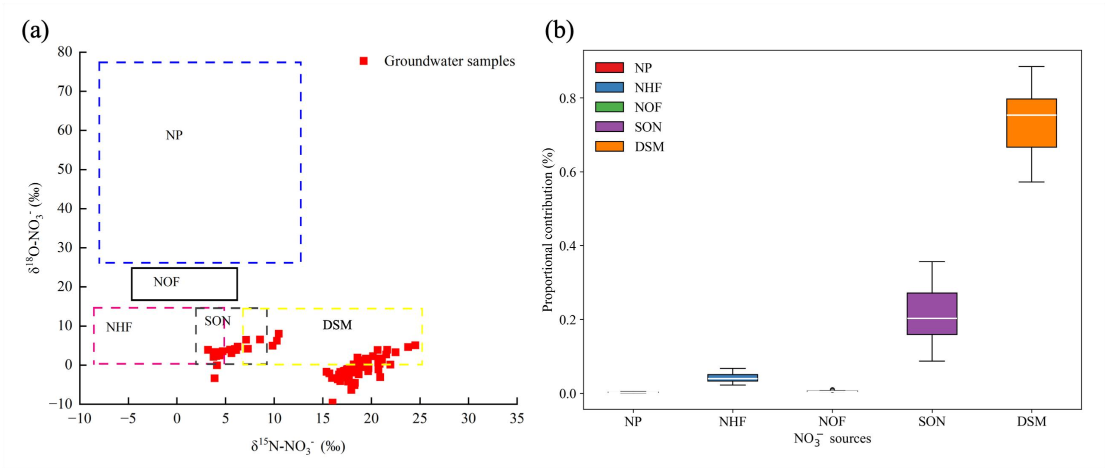

Fig.10. (a) distributions of $\delta^{15}N$-NO_3^- and $\delta^{18}O$-NO_3^- values in the study area. (b) proportional contributions of the main NO_3^- sources evaluated by the MixSIAR model. Note: boxplots denote the 25th, 50th and 75th percentiles.

3.3 Bayesian model analysis and correlation analysis

In the normal season, the Bayesian model identified Mg^{2+} as the central driver of NO_3^-, consistent with its strong correlation (Fig. 11). Na^+ showed a significant negative effect despite only a weak correlation (r=0.39), suggesting its variation reflects hydrological processes rather than direct involvement in NO_3^- transformation. Although TDS and EC correlated strongly with NO_3^- (r>0.8), their probabilities of direction (pd<80%) indicate indirect effects through collinearity with other ions. In the dry season, SO_4^{2-} emerged as the primary positive driver, agreeing with its high correlation with NO_3^- (r=0.96), whereas Na^+ and Ca^{2+} exerted significant negative effects despite weak positive correlations, likely because their enrichment reflects evaporation while NO_3^- derives from anthropogenic inputs, with no direct causal link. In the wet season, Cl^- and Mg^{2+} were the dominant factors with clear directional effects, matching their correlations with NO_3^- (r=-0.78 and 0.84, respectively) and confirming their direct influence. TDS and EC again showed wide posterior distributions and low pd values, indicating effects attributable to collinearity with Mg^{2+} and Cl^- rather than independent contributions.

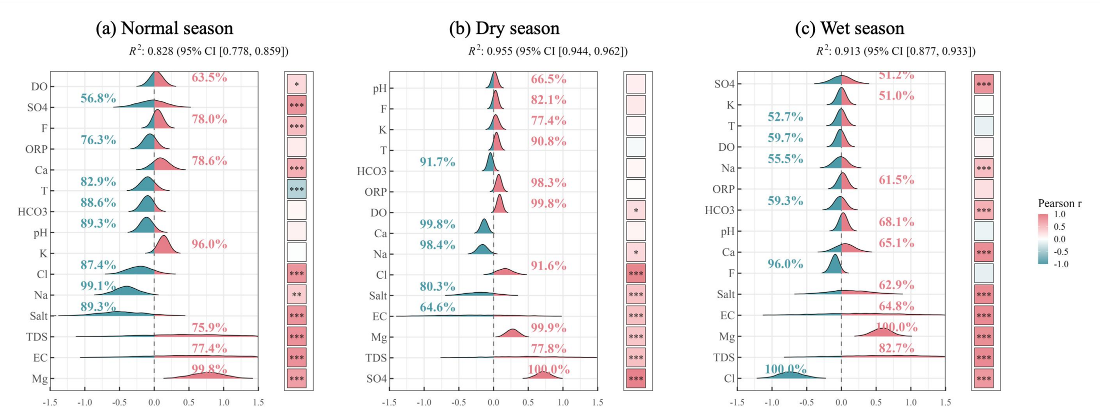

Fig.11. Factor effects and Pearson coefficients of physicochemical variables on NO_3^- at different. periods. The left part of each subgraph shows the relative importance and posterior distribution of each environmental variable to NO_3^- after the Bayesian model operation. The red area represents the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

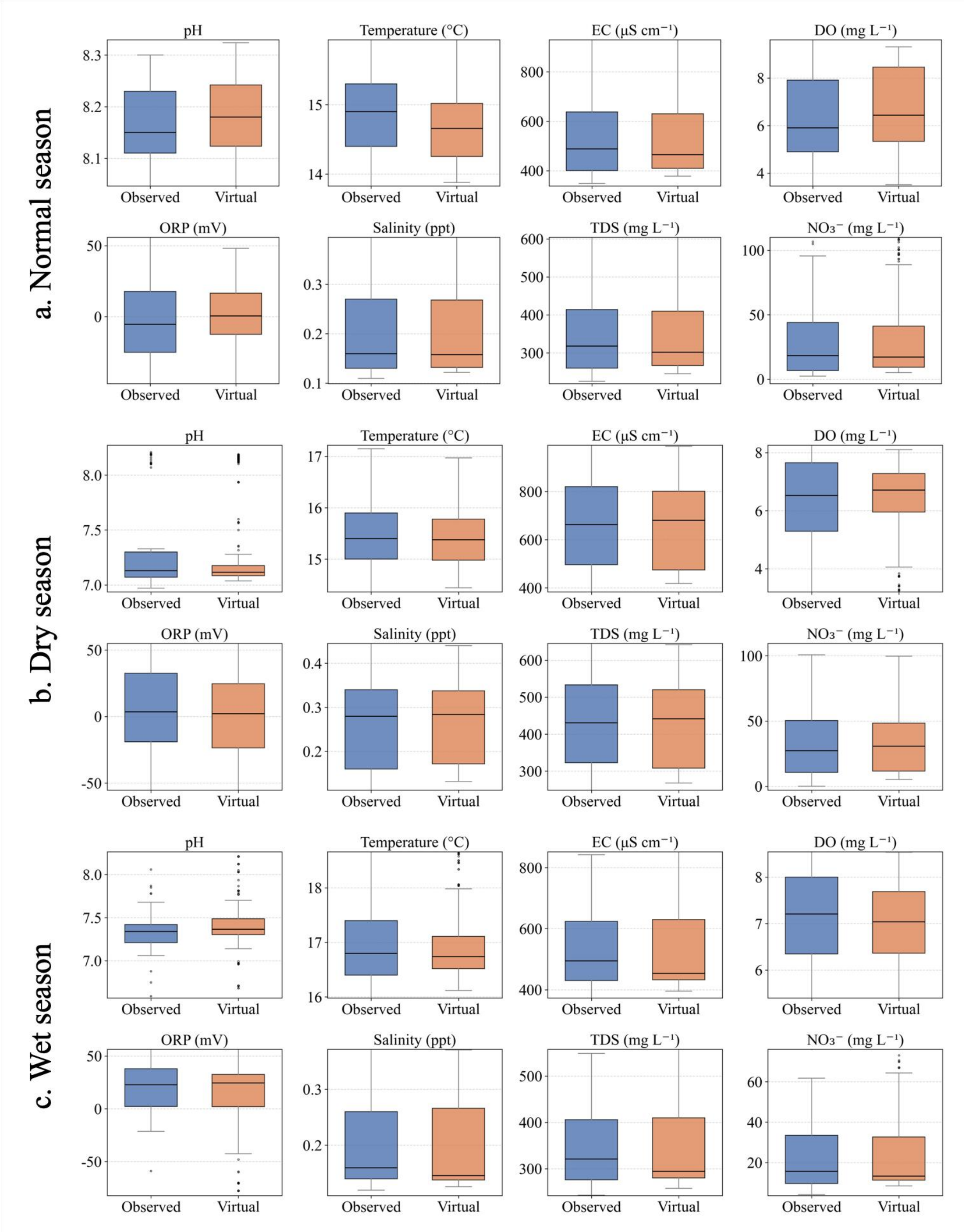

Fig.12. Box plots of the observed and virtual variable data at different periods.

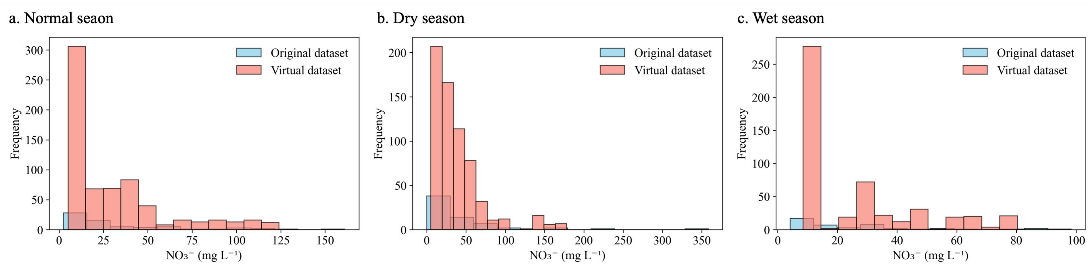

Fig.13. Comparison of nitrate concentration distribution in the original and virtual datasets under.

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. S2). 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).

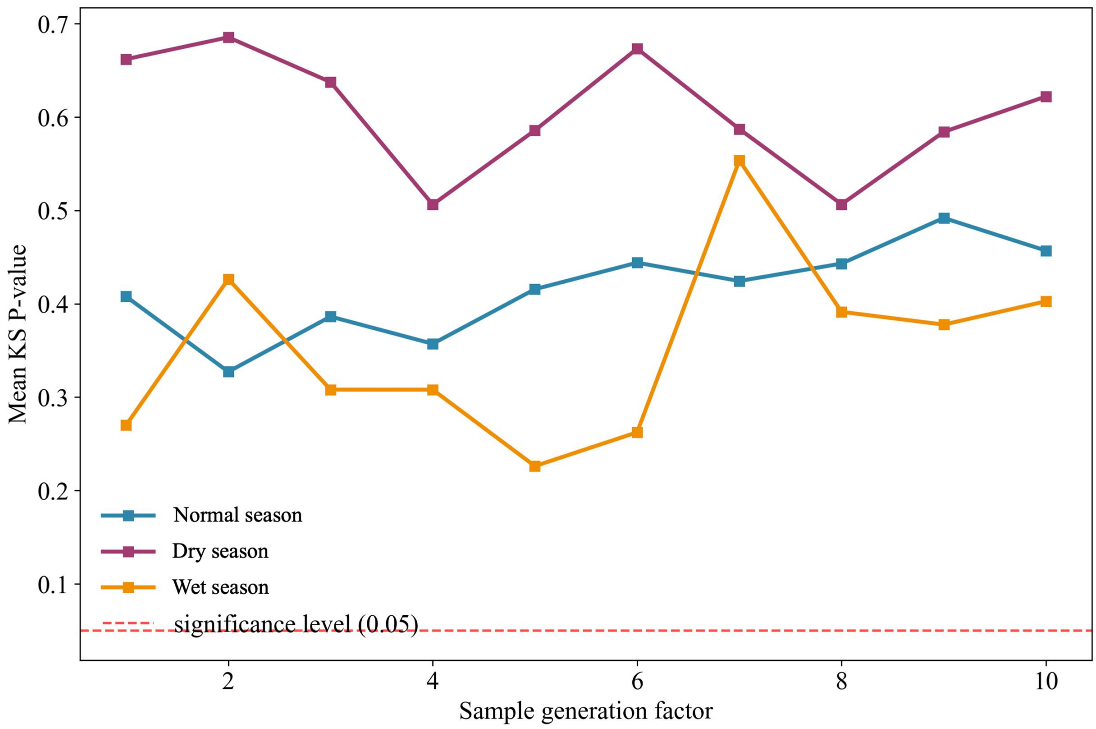

Fig.14. Mean KS P-values across different sample generation factors during normal, dry, and wet seasons. The dashed red line indicates the significance level (α=0.05). Values above this threshold suggest no significant distributional difference between virtual and observed samples.

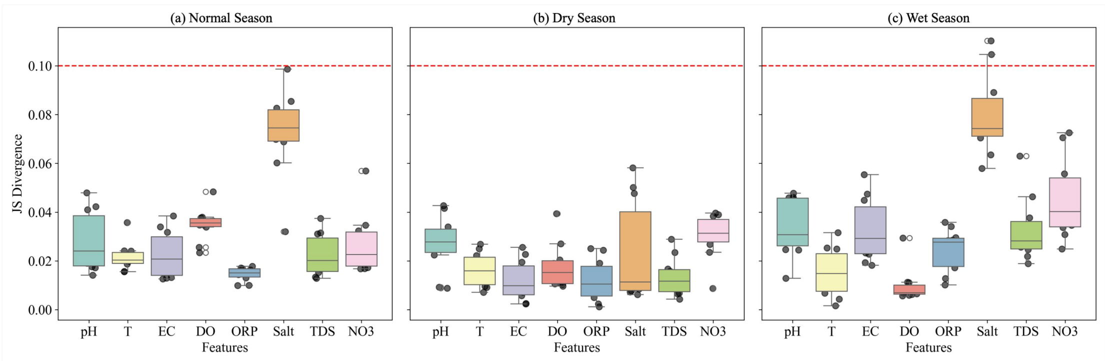

Fig.15. Boxplots of JS divergence for water quality features across hydrological seasons. The dashed red line denotes the strict similarity threshold (JS=0.05). Lower values indicate higher distributional consistency.

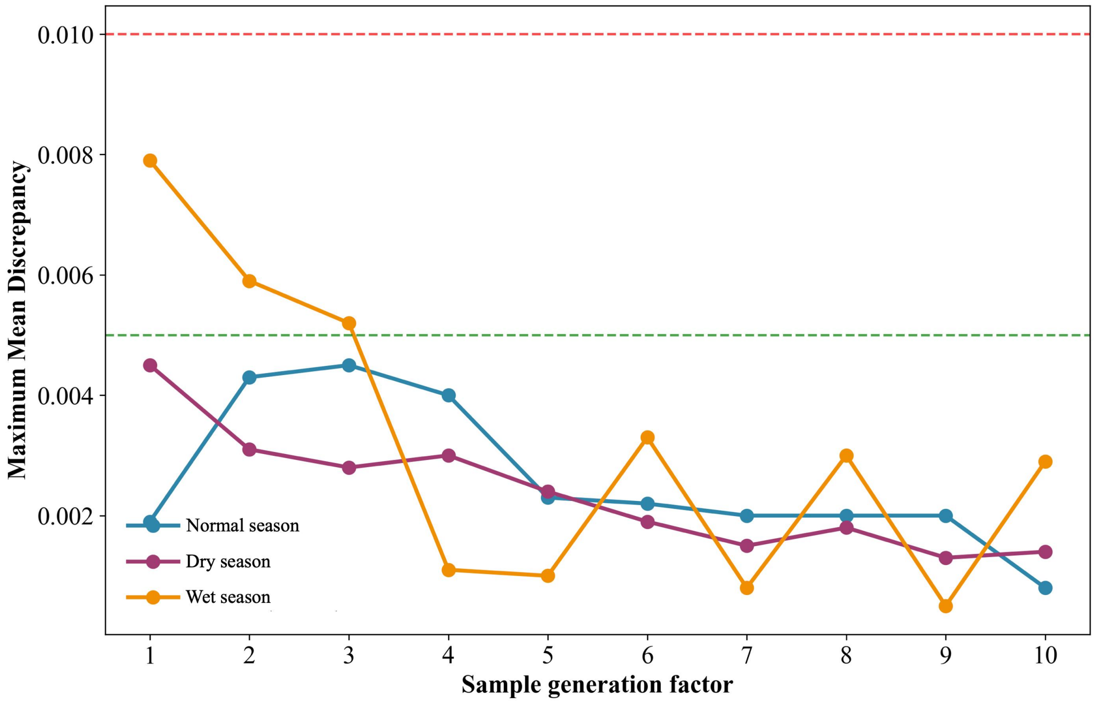

Fig.16. Variation of Maximum Mean Discrepancy (MMD) with sample generation factor across seasons. MMD values approaching zero and remaining within the confidence interval confirm the absence of systematic bias in the virtual samples within the Reproducing Kernel Hilbert Space (RKHS).

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

In the normal season, baseline and quantum enhanced RF models achieved R^2 of 0.6727 and 0.6602, respectively, with wide bootstrap 95% CIs reflecting the limited sample size (Table S2). Augmenting virtual samples from 1× to 10× the original size steadily raised R^2 above 0.958 (quantum enhanced RF reaching 0.9622, 95% CI: [0.9744, 0.9881]), with gains plateauing beyond approximately 500 virtual samples. The dry season showed the poorest baseline performance (R^2=0.2841, 95% CI: [0.2884, 0.7276]; Table S2-S3), attributable to high NO_3^- variability and outlier sensitivity under data sparsity. Virtual sampling markedly improved accuracy: R^2 reached 0.527 at 1× generation and 0.854 (95% CI: [0.8431, 0.9221]) at 8×. Although the quantum enhanced model slightly lagged at 2× generation or less, both models converged to high accuracy, confirming that virtual samples effectively mitigated data sparsity and skewed distributions. The wet season yielded the best baseline performance (R^2=0.7331, 95% CI: [0.6715, 0.8926]; Table S4), owing to lower NO_3^- concentrations and smaller spatial variability. Augmentation further raised accuracy to R^2=0.962 at 4× generation and a stable 0.977 at 10× (RMSE as low as 3.03 mg L^{-1}; 95% CI: [0.9787, 0.9875]). Quantum enhanced and classical RF performed nearly identically, indicating limited marginal benefit from quantum feature encoding when data quality is high and relationships are more linear.

In the normal season (Fig. 17A1-A2), both models trained on the 66 original samples showed marked deviations between predicted and observed values, with highly dispersed predictions and medians far from the observed median, consistent with the low R^2 of small sample modeling. As virtual samples increased, predicted distributions gradually converged toward the observed values and the boxplot IQR and whiskers narrowed, indicating substantially improved stability and accuracy. At 10× generation (660 virtual samples), predicted and observed boxplots nearly overlapped, with RMSE reduced to 6.02 mg L^{-1}; the quantum enhanced model slightly outperformed classical RF (R^2=0.9622, bootstrap 95% CI: [0.9744, 0.9881]), confirming robust accuracy despite the small original sample. In the dry season (Fig. 17B1-B2), both models systematically overestimated NO_3^- on the original data owing to its extremely high and skewed concentrations, placing predicted boxplots entirely above the observed values (R^2=0.28, 95% CI: [0.2884, 0.7276]). Virtual sampling brought substantial improvement: predicted medians and ranges began converging from 1× generation, and at 8× or higher the models captured the high concentration intervals, with RF reaching R^2=0.8542 (95% CI: [0.8431, 0.9221]) at 9×. Classical RF slightly outperformed the quantum enhanced model at low generation levels, but their performance converged as sample size grew, demonstrating that virtual sample generation effectively alleviates challenges from data sparsity and extreme values. In the wet season (Fig. 17C1-C2), predictions already agreed well with observations on the original data, and virtual samples further narrowed the IQR and concentrated predictions within the observed distribution.

At 10× generation, agreement was exceptionally high (R^2=0.977, RMSE=3.03 mg L^{-1}), demonstrating excellent predictive performance.

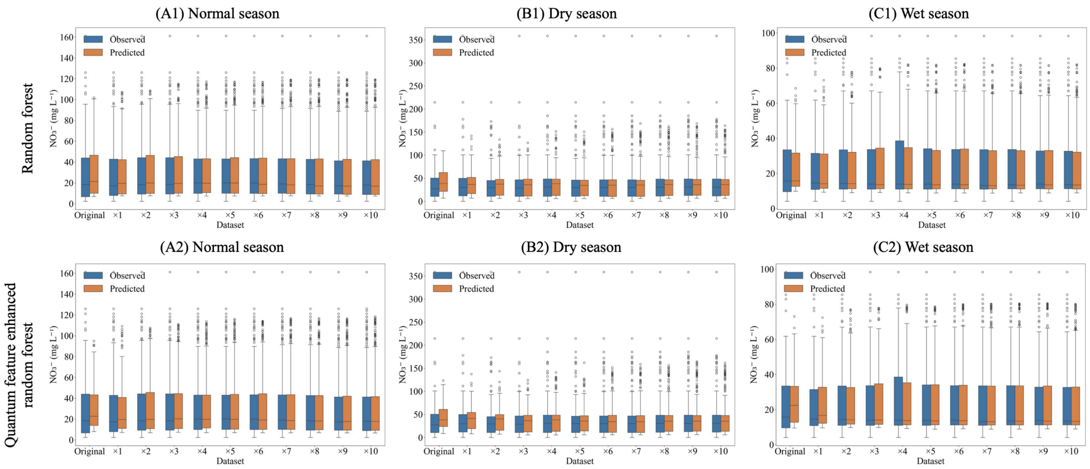

Fig.17. Comparison of observed and predicted NO_3^- concentrations across data generation levels for random forest and quantum feature-enhanced random forest models in normal, dry, and wet Seasons.

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.

In the normal season, models trained on the original samples performed poorly, with R^2 of 0.167 (RF) and 0.119 (quantum enhanced RF) and RMSE as high as 32.89 and 33.82 mg L^{-1}, respectively (Table S5), indicating that AEF embedding features alone cannot adequately capture the hydrological processes of this period under limited sample size. Virtual sampling markedly improved performance: at 10× generation, RF reached R^2=0.860 with RMSE reduced to 10.73 mg L^{-1} and a narrowed bootstrap 95% CI of [0.8792, 0.9326], while the quantum enhanced RF achieved R^2=0.844 with a consistent trend, confirming statistically stable performance even with indirect remote sensing inputs. Boxplot comparison (Fig. 18A) shows that initial predictions severely overestimated low to medium concentrations while underestimating high value tails; as sample size expanded, predicted distributions progressively converged toward the observed values, with much improved agreement in median and IQR. This confirms that virtual samples effectively enhanced the capability of AEF features to represent nonlinear patterns.

In the dry season, the original dataset yielded a negative R^2, reflecting the weak generalization of AEF features under high variability and heavy tailed distributions (Table S6). Performance improved steadily with virtual sampling: R^2 rose to 0.039 at 1×, 0.641 at 8×

(RMSE=21.05 mg L^{-1}, bootstrap 95% CI: [0.6385, 0.8638]), and 0.674 at 10× (RMSE=19.86 mg L^{-1}). The quantum enhanced RF slightly outperformed standard RF at high expansion levels, suggesting that quantum encoding helps mitigate the influence of extreme values and enhances robustness (Fig. 18B). Prediction distributions show that the initial model entirely failed to capture the high concentration clustering of dry season NO_3^-, whereas after data generation the predicted boxplots progressively covered the true high value intervals, with tail behavior aligning with observations. In the wet season, although baseline R^2 was also negative, improvement was the most rapid (Table S7): R^2 reached 0.5 at only 2× expansion, 0.685 at 5×, and stabilized at 0.784 (RF, RMSE=8.27 mg L^{-1}) and 0.781 (quantum enhanced RF) at 10×. Boxplots confirm that the initially dispersed and systematically biased predictions rapidly converged to the dense intervals of observed values, with final IQR and whisker ranges showing high overlap.

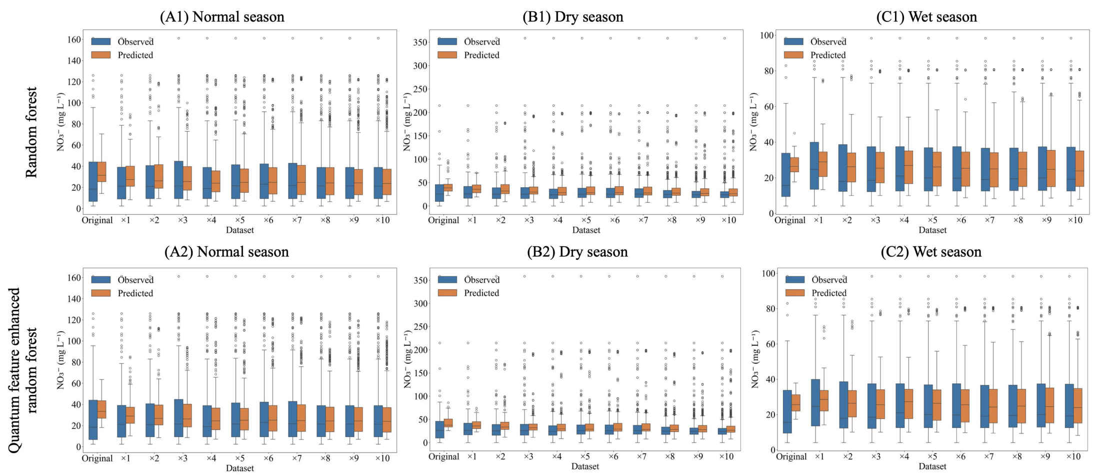

Fig.18. Comparison of observation and prediction of NO_3^- concentration by random forest and quantum featution-enhanced random forest models at data enhancement levels in normal, dry and wet seasons: based on AlphaEarth Foundation as the input variable.

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

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 (Fig. 19). 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.

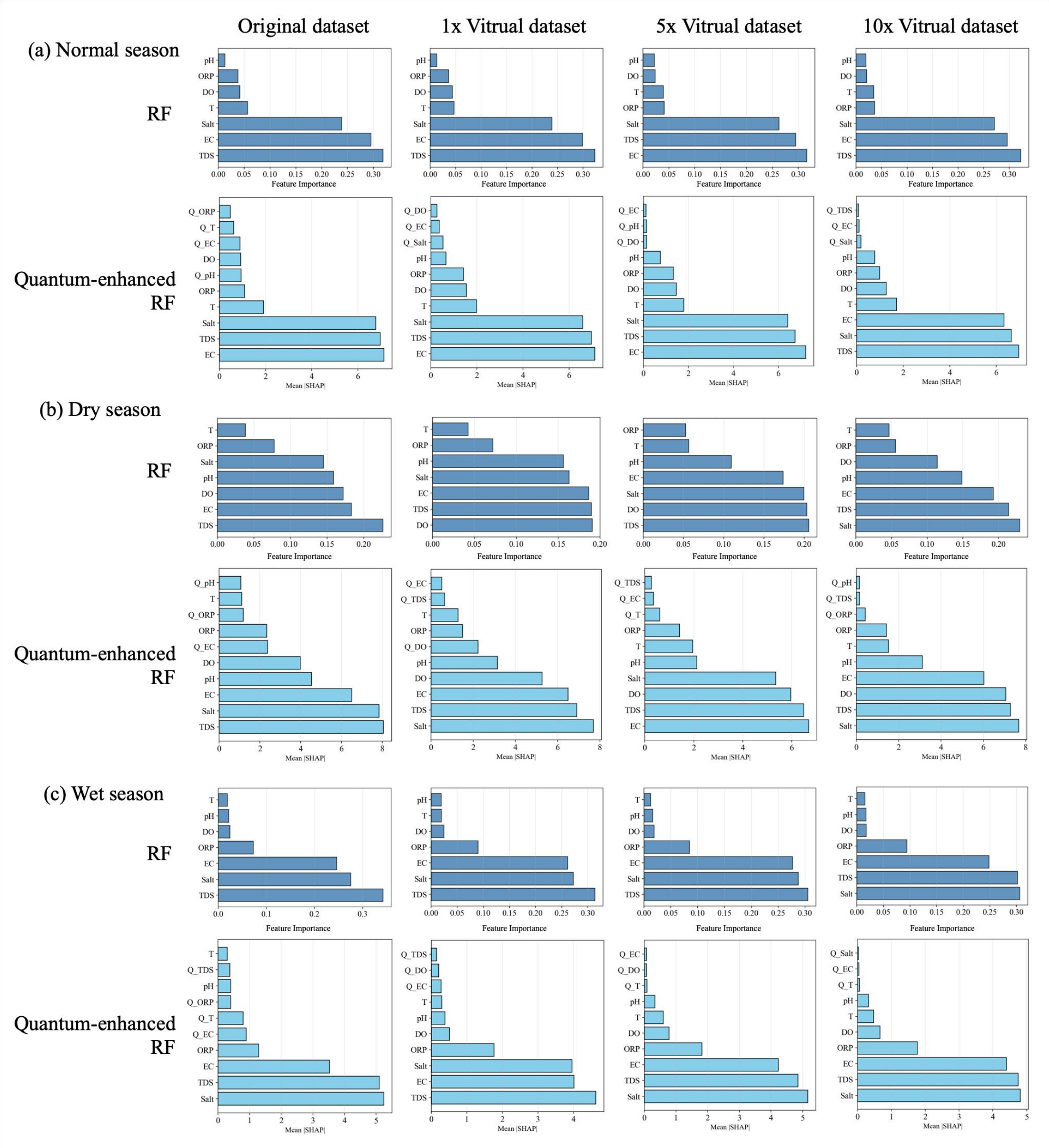

Fig.19. Input feature importance of the classical Random Forest (RF) and quantum-enhanced RF models for seasonal nitrate prediction using in-situ measured water quality parameters. Panels (a)-(c) correspond to the normal, dry, and wet seasons, respectively. Columns represent the original, 1x, 5x, and 10x virtual-sample datasets.

Fig. 20 presents the feature importance of the RF and quantum enhanced RF models using only the 64 dimensional AEF embedding vectors as inputs. Although these abstract features cannot be assigned explicit physical meanings, their importance rankings reveal which remote sensing semantics are critical for nitrate prediction. Compared with in situ parameters, AEF feature importance fluctuates considerably across seasons and data volumes, with no consistent core feature set, indicating that AEF embeddings, despite their rich environmental information, correlate only weakly with nitrate and require large data volumes to establish a robust mapping. Seasonally, A05, A07, and A00 ranked highest in the normal season (likely encoding land use, soil moisture, or vegetation cover); A08, A06, and A05 in the dry season (associated with surface dryness, bare land, or human activity intensity, matching the spatial distribution of high concentration pollution sources); and A02, A03, and A05 in the wet season (related to runoff, vegetation growth, or soil water content, reflecting rainfall driven pollutant migration). Virtual sample generation proved crucial for stabilizing these rankings: the initially chaotic importance ordering on the original data gradually clarified as virtual samples increased, highlighting core features and further demonstrating the effectiveness of the generation strategy for small sample modeling. The quantum enhanced model showed a similar importance distribution to classical RF but occasionally assigned slightly higher weights to certain features, suggesting that quantum feature encoding helps extract more discriminative information from the high dimensional remote sensing semantic space and marginally refines feature selection.

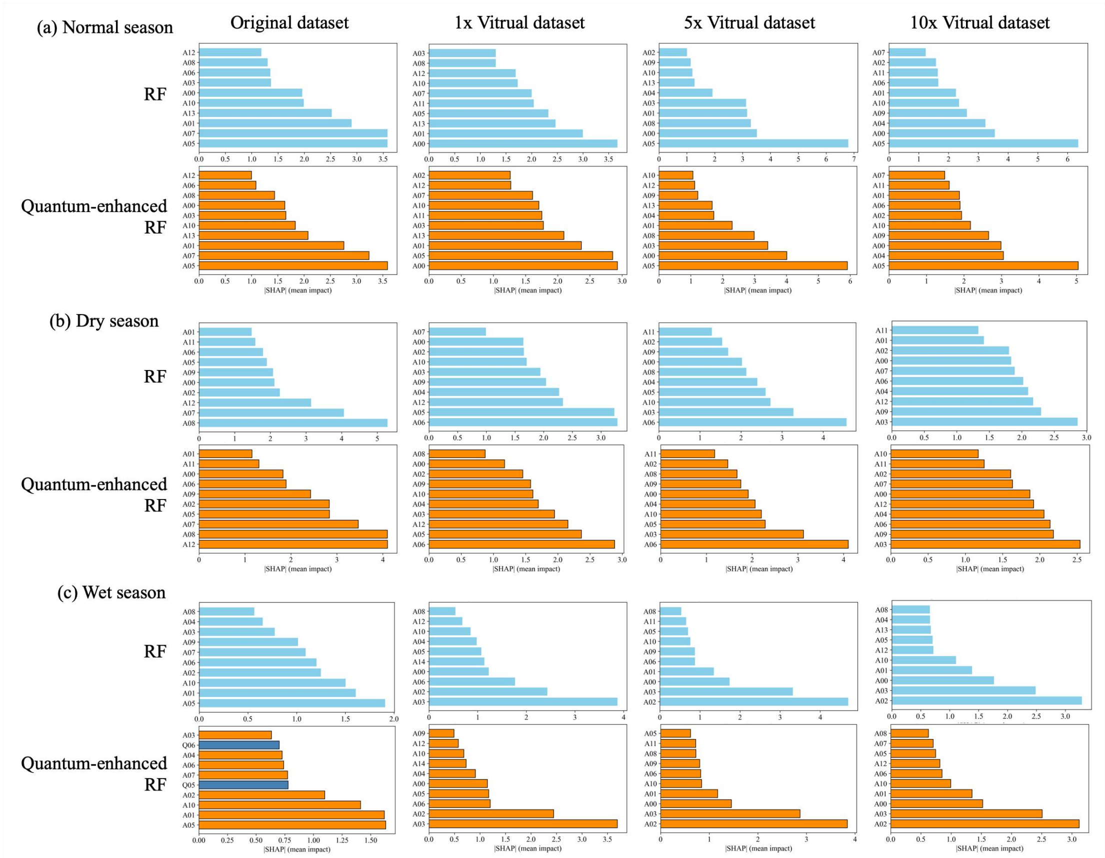

Fig.20. Importance of AlphaEarth Foundation (AEF) embedding features for seasonal nitrate prediction using the classical Random Forest (RF) and quantum-enhanced RF models. Panels (a)-(c) correspond to the normal, dry, and wet seasons, respectively. Columns represent the original, 1x, 5x, and 10x virtual-sample datasets.

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.

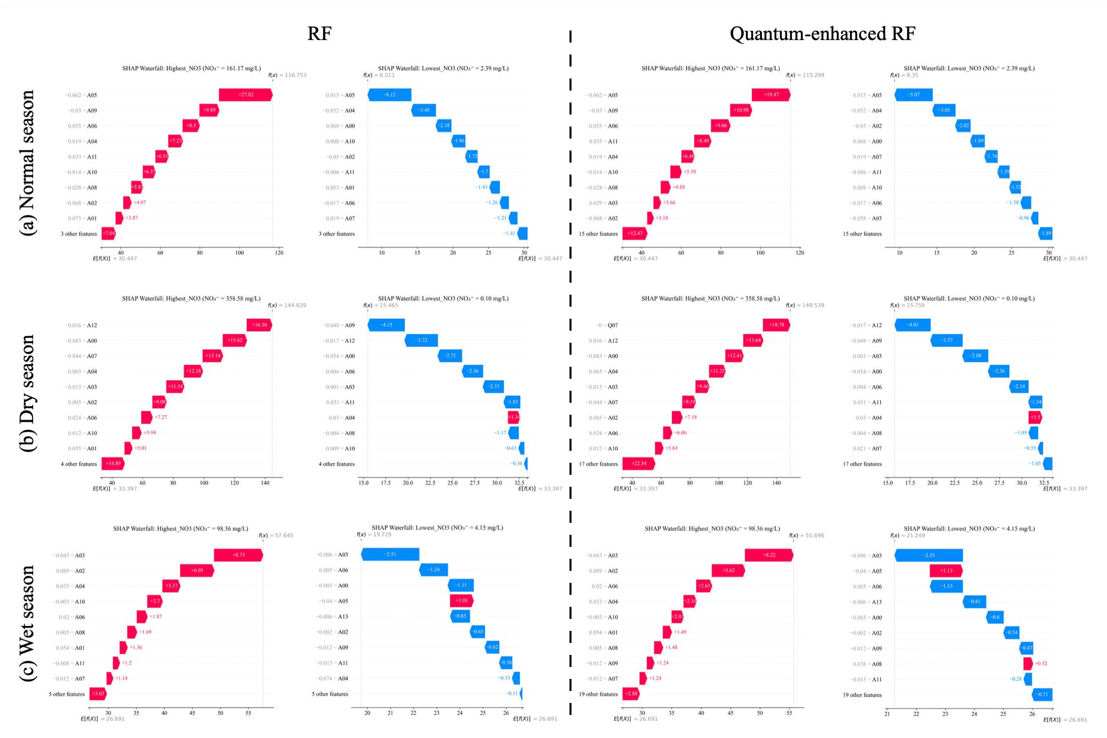

Fig.21. SHAP waterfall-based local feature attribution for representative samples with the highest and lowest observed NO_3^- concentrations. Panels (a)-(c) correspond to the normal, dry, and wet seasons, respectively. In each panel, the left group shows the classical RF model and the right group shows the quantum-enhanced RF model; within each group, the left and right plots correspond to the highest- and lowest-concentration samples, respectively.

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

In the normal season, the total order Sobol indices of the classical RF consistently ranked TDS (0.2882 to 0.2933), EC (0.2701 to 0.2839), and Salt (0.2411 to 0.2612) as the three most sensitive parameters across all generation gradients (Fig. 22), in close agreement with the Gini and SHAP importance rankings. The first order index of TDS remained above 0.29 in the original data and stabilized at 0.2938 after 10× generation, indicating a stable and dominant independent contribution to nitrate prediction. In contrast, pH, T, DO, and ORP showed total order indices below 0.08, reflecting limited influence. The quantum enhanced RF yielded the same overall ranking, with the sensitivity of TDS, EC, and Salt further stabilized at 10× generation (0.2586, 0.2516, and 0.2416, respectively). Notably, the total order sensitivity of ORP rose from 0.0313 (original) to 0.0608 (10×), suggesting that quantum feature encoding strengthens the model's perception of redox contributions to nitrate dynamics and that virtual sampling improves the stability of sensitivity estimates for low contribution parameters.

In the dry season, characterized by highly skewed nitrate distributions and strong evaporative concentration, parameter sensitivity differed from the normal season (Fig. 23). For classical RF on the original data, the five most sensitive parameters were TDS (0.2713), Salt (0.1907), DO (0.1766), EC (0.1739), and pH (0.1146). The notably higher pH sensitivity than in other seasons indicates that the acid-base environment strongly controls nitrate accumulation under intense evaporation. With increasing virtual sample generation, the sensitivity of TDS, EC, and Salt gradually stabilized, while DO remained high (0.1596 to 0.2165), confirming its persistent control over nitrate as an indicator of nitrification intensity in the oxidizing dry season environment. After 10× generation, the quantum enhanced RF showed higher sensitivity for pH (0.1015) and DO (0.1665) than classical RF, suggesting the hybrid quantum-classical framework is more robust in capturing nonlinear hydrochemistry-nitrate relationships under the highly heterogeneous, small sample conditions of the dry season.

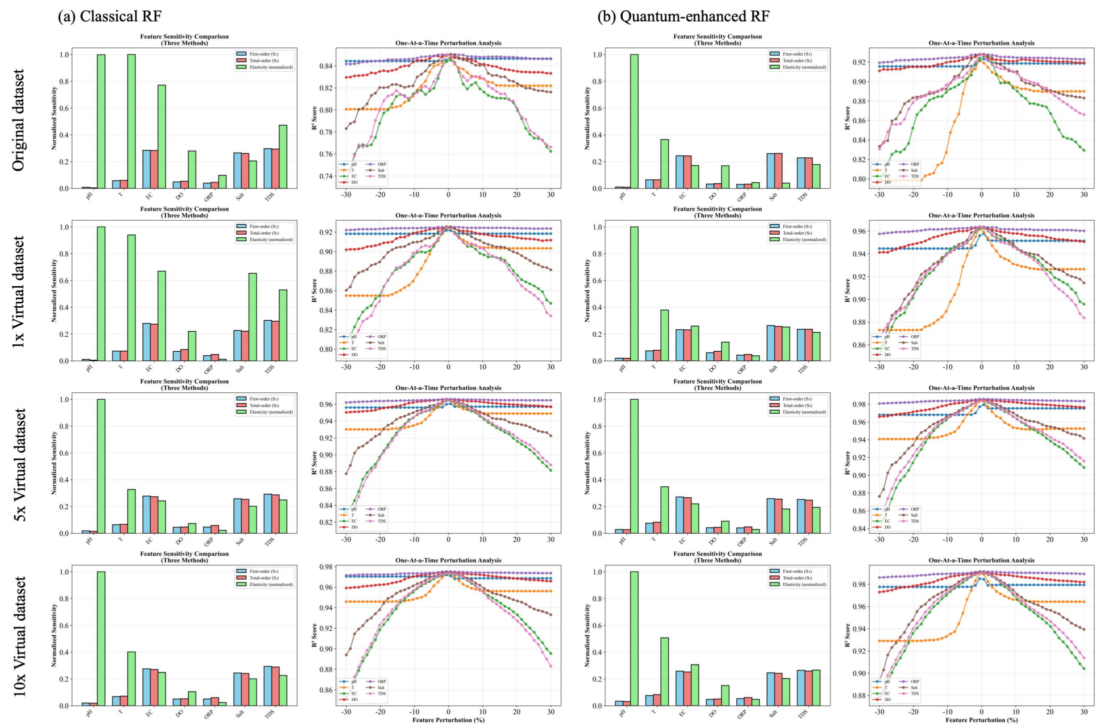

Fig.22. Sensitivity analysis of key parameters during the normal season.

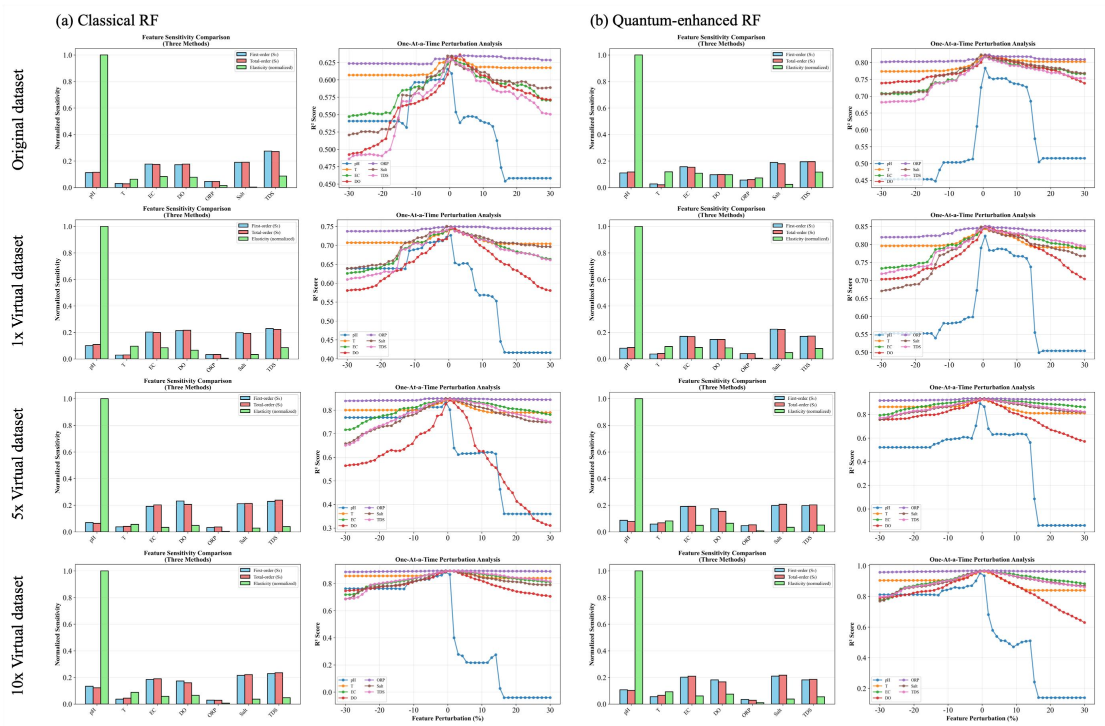

Fig.23. Sensitivity analysis of key parameters during the dry season.

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.

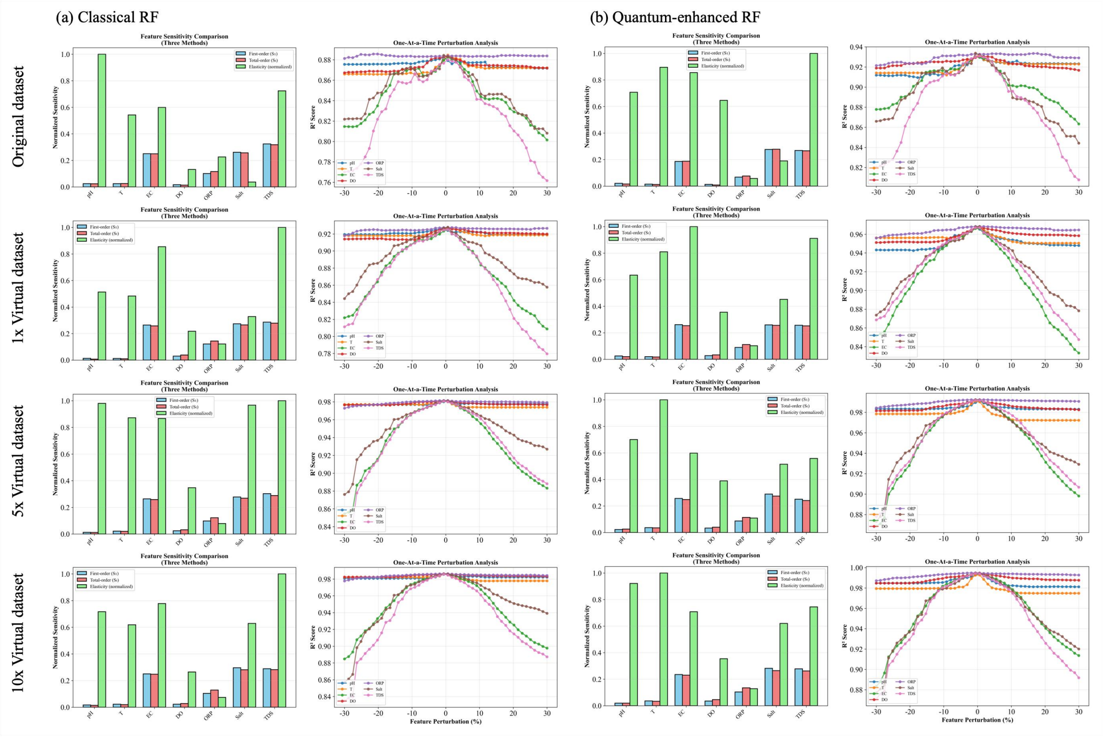

Fig.24. Sensitivity analysis of key parameters during the wet season.

3.6.2 Feature perturbation sensitivity and elasticity analysis

The OAT perturbation test was performed by varying each target parameter within ±30% of its measured range while fixing others, and elasticity coefficients were calculated to quantify the directional response of predicted nitrate per unit parameter change (|value| > 1 indicates an elastic response). The results corroborated the Sobol analysis. In the normal season, the perturbation sensitivity of TDS and EC remained highest (>0.11) for both models after 10× generation. In the dry season, pH sensitivity rose sharply from 0.3033 to 2.0128 (classical RF) and from 0.6026 to 1.3757 (quantum enhanced RF), confirming the model's strong responsiveness to pH; correspondingly, pH elasticity coefficients reached 18.5446 and 24.2152 after 10× generation, an extremely strong positive elastic response identifying pH as the key control on dry season nitrate accumulation. In the wet season, the perturbation sensitivity of Salt and TDS stabilized at 0.1358 and 0.1616 (classical RF, 10×) as the highest responses. Elasticity analysis further showed that in the normal season T, pH, and EC had coefficients above 1 (positive elastic response), whereas ORP was inelastic (0.1037); in the wet season, the coefficients of TDS, EC, and Salt were all close to 1, indicating a near proportional linear response of nitrate to salinity indicators, consistent with the dominant dilution effect.

3.6.3 Impact of virtual sample generation on sensitivity estimation

The t-SNE-GMM-KNN virtual sample generation markedly improved the stability and reliability of sensitivity estimation. For the dry season with the most severe data sparsity, the 95% CI of the Sobol index for TDS narrowed by 87.9% from original to 10× generation, and the coefficient of variation of perturbation sensitivity decreased by an average of 62.4%, indicating that augmentation effectively reduces small sample variance and avoids ranking bias from insufficient representation. As generation multiples increased, the ranking of core parameters (TDS, EC, Salt) remained stable while estimates for secondary parameters (ORP, DO, pH) gradually converged, confirming that virtual sampling improves statistical robustness without altering the inherent hydrochemical driving mechanism. The quantum enhanced RF further showed smaller sensitivity fluctuations across generation gradients than classical RF, particularly on the original dry season data (21.7% lower coefficient of variation in total order indices), demonstrating that quantum feature encoding enhances robustness under small sample conditions. Overall, the multi-dimensional sensitivity analysis consistently identified TDS, EC, and salinity as the most critical and stable parameters across all seasons, in close agreement with the correlation, Bayesian, and SHAP analyses, while dry season pH and DO and wet season ORP emerged as season specific drivers, revealing differentiated controls on nitrate dynamics under varying hydrological conditions. These results validate the model structure and feature selection and provide a quantitative basis for optimizing monitoring indicators and targeted control of groundwater nitrate pollution.

4. Discussion

4.1 Nitrogen sources, migration, and transformation

The Piper and Gibbs diagrams, together with ion ratios, indicate a Ca-Mg-HCO_3^- hydrochemical type dominated by carbonate dissolution with weak cation exchange (Liu et al., 2025). The lower NO_3^- concentrations during the rainy season, combined with enhanced cation exchange, suggest rainfall-driven manure leaching accompanied by temporary aquifer retention of NO_3^- (Sun et al., 2024; Wang et al., 2025). Isotopic evidence and MixSIAR apportionment identify domestic sewage and manure (DSM, 74.1%) and soil organic nitrogen (SON, 20.9%) as the dominant nitrate sources, with negligible fertilizer and precipitation inputs (Mao et al., 2023). This pattern implies that direct fertilizer leaching is not the dominant pathway; rather, given the thick (>10 m) vadose zone of the North China Plain, elevated NO_3^- likely derives from historical fertilizer residues and long-term manure infiltration, especially where farmlands adjoin rural residences (Wang et al., 2025; Wu et al., 2024). The $\delta^{15}N$-NO_3^- values (12.2‰~18.2‰) fall within the manure/SON range rather than that of chemical fertilizers, indicating microbial mineralization and nitrification, consistent with ammonification and nitrification of SON and DSM-derived ammonium under aerobic conditions (Li et al., 2022; Liu et al., 2023; Ahmed et al., 2013).

Seasonal nitrate variation reflects contrasting recharge regimes. During the dry season, scarce precipitation, strong evaporation, and declining groundwater levels create a migration potential gradient that drives nitrate accumulation in discharge areas; concurrently weakened cation exchange (reduced Na^+ adsorption and relative Ca^{2+} depletion) lowers aquifer retention capacity, making accumulation the dominant process (Zhang et al., 2023). In the wet season, rapid infiltration raises groundwater levels and velocities and reactivates cation exchange. The oxidizing aquifer conditions, indicated by extremely low nitrite and ammonium concentrations, together with $\delta^{18}O$ values of nitrate (-9.58‰ to 8.04‰) within the typical nitrification range, exclude significant denitrification and confirm nitrification as the dominant transformation process (Zhang et al., 2025).

Although the vadose zone of the North China Plain generally exceeds 10 m and should theoretically damp seasonal recharge signals (Liu et al., 2022), the distinct seasonal nitrate variability observed here can be attributed to three regional factors. First, decades of excessive fertilization have built a legacy nitrogen reservoir in the vadose zone, so groundwater nitrate is sustained by readily mobilized historical accumulation rather than current inputs alone (Liu et al., 2025; Miao et al., 2023). Second, intense monsoonal rainfall can generate preferential flow paths that bypass slow matrix flow, enabling rapid nitrate leaching despite the thick unsaturated zone (Williams et al., 2023). Third, intensive irrigation abstraction alters the flow field and water table depth, enhancing hydraulic connectivity between soil nitrogen and the aquifer (Yang et al., 2022). Together, these factors explain why legacy nitrogen mobilization, dry season evaporative concentration, and wet season dilution override the damping effect of the thick vadose zone (Feng et al., 2024).

4.2 Virtual sample generation mitigates small sample bias and reveals seasonal sensitivity

Model overfitting and poor generalization due to small samples are prevalent challenges in environmental forecasting (Zhu et al., 2023). The virtual samples generated by the t-SNE/GMM/KNN strategy reproduce the statistical characteristics (mean, standard deviation, coefficient of variation) and seasonal hydrochemical differences of the original data, and the substantial performance gains after augmentation indicate that data sparsity, rather than model capacity, is the core bottleneck in seasonal nitrate modeling (Saha et al., 2023). Narrowing bootstrap confidence intervals with increasing sample size further confirms that augmentation reduces variance in performance estimates rather than exploiting specific data splits. The gains diverge seasonally: under the highly right skewed nitrate distribution of the dry season, R^2 rises from 0.28 to over 0.85 with tenfold expansion, because evaporation driven concentration intensifies spatial heterogeneity and process nonlinearity, requiring richer samples to capture tail behaviors (Li et al., 2025); in the wet season, rainfall dilution homogenizes the system, so excellent accuracy is achieved with smaller marginal gains (Bigler et al., 2024).

The proposed t-SNE/GMM/KNN strategy outperforms traditional oversampling and deep generative models in preserving multimodal structures and heavy tailed covariance, as the latter either neglect the manifold geometry of high dimensional geochemical spaces or require large training datasets unavailable in this study (Udu et al., 2025). Its advantages over GMM and GAN based methods are threefold: t-SNE captures clustering structures driven by distinct hydrological processes; BIC based cluster optimization avoids subjective parameterization; and KNN inverse mapping reconstructs high dimensional samples without large scale training, suiting small sample scenarios (Silva et al., 2023; Peng et al., 2025). Bootstrap analysis confirms that confidence intervals narrow with increasing virtual sample size; for the dry season RF model, the 95% CI width drops from 0.440 to 0.053 under tenfold augmentation, indicating effective mitigation of small sample bias. Although tenfold virtual sample augmentation substantially improves predictive performance, overfitting to synthetic patterns remains a potential risk. The t-SNE/GMM/KNN framework mitigates this risk by adhering to the manifold structure and multimodal distributions of the original data; nevertheless, augmentation beyond tenfold may introduce artificial covariance unrepresentative of true geochemical processes (Tanner et al., 2022). Bootstrap analysis shows that performance gains plateau beyond eightfold generation, indicating that tenfold augmentation represents an empirical upper bound at which information extraction from limited field samples approaches saturation without compromising generalizability (Zhu et al., 2023).

4.3 Error propagation under virtual sample

Assuming a 5.0% coefficient of variation for input hydrochemical parameters, we propagated uncertainty through 500 perturbed realizations and quantified output variability via standard deviation and 95% confidence interval width (Fig. 25). Uncertainty propagation shows clear seasonal and model dependent patterns. In the normal season, both RF and quantum enhanced RF models remained stable, with output standard deviations decreasing by 4.5% and 15.8% at tenfold augmentation, indicating that data generation constrains uncertainty amplification under relatively symmetric distributions (Dega et al., 2023). The dry season reveals a critical divergence: classical RF uncertainty increased markedly (σ_{out} +48.9%; CI width +35.9%), reflecting the dominance of extreme values in highly right skewed nitrate distributions, whereas the quantum enhanced RF reduced output standard deviation by 27.9% and CI width by 20.6%. This robustness suggests that quantum feature encoding via Pauli Z expectation values better disentangles nonlinear relationships in high variability regimes, limiting error amplification. In the wet season, dilution driven homogenization yields low, stable uncertainties for both models ($\sigma_{out}<2.1$ mg L^{-1}), with the quantum enhanced RF again suppressing uncertainty more strongly (-28.7% versus -7.8%), though absolute gains remain modest.

These results confirm that virtual sample generation enhances not only accuracy but also resilience to measurement uncertainty, provided the generation factor matches the complexity of the underlying distribution: for the high variability dry season, moderate augmentation is optimal, whereas excessive generation (beyond eightfold) may reintroduce variance in classical models. Three broader implications follow. First, error propagation analysis offers a robustness criterion beyond conventional accuracy metrics (R^2, RMSE), which is especially valuable in small sample modeling where overfitting risk is elevated. Second, the superior uncertainty control of the quantum enhanced RF in the dry season supports the theoretical advantage of quantum feature spaces in capturing heavy tailed, nonlinear geochemical processes, highlighting the value of hybrid quantum classical architectures under extreme hydrological conditions. Third, the distinct seasonal uncertainty structures reinforce the necessity of seasonally stratified modeling frameworks, as a single unified model cannot adequately represent error behavior across contrasting hydrological regimes.

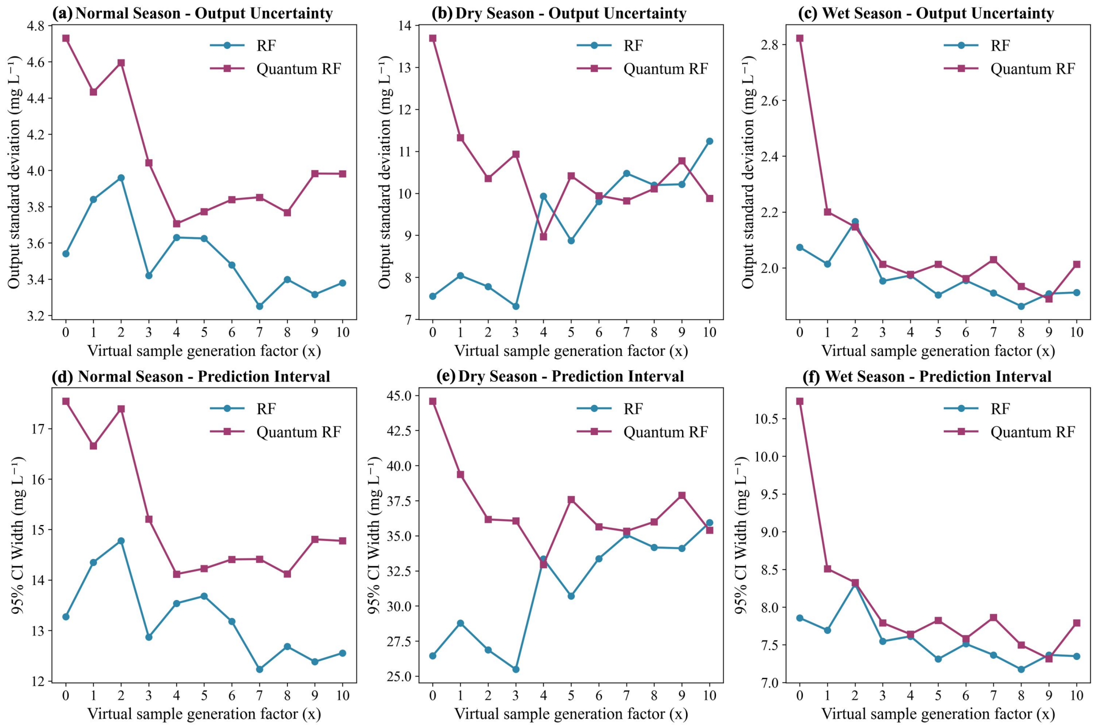

Fig. 25. Variation of output standard deviation of classical Random Forest and quantum-enhanced Random Forest with virtual sample generation multiples under 5% input coefficient of variation in normal, dry and wet seasons.

4.4 Performance analysis of hybrid quantum-classical model

Quantum machine learning captures complex nonlinear relationships through feature mapping in high dimensional Hilbert spaces, offering gains when data are scarce or highly skewed (Lamichhane et al., 2025). When classical feature representation saturates, the Z feature mapping of parameterized quantum circuits may expose entangled nonlinear patterns that enhance discriminability, though such gains converge once virtual sample expansion is sufficient (Hong et al., 2025; Oliveira et al., 2025). Here, quantum features were generated by analytically computing Pauli Z expectation values, circumventing hardware noise and making the quantum enhanced RF feasible for small sample environmental tasks (Gujju et al., 2024). The advantage is conditional rather than universal: gains are marginal in the wet season, where relationships are largely linear, but substantial in the dry season, where sparse data and extreme values make quantum encoding more robust to measurement noise (Ranga et al., 2024). Hybrid quantum classical modeling therefore adds value mainly in complex, information limited scenarios by expanding representational capacity rather than replacing classical logic, supporting the potential of quantum machine learning for small sample problems in earth sciences even when its absolute advantage remains modest.

In a comparable intensive agricultural setting on the North China Plain, Zhao et al. (2026)

evaluated RF, XGBoost, and LightGBM on 157 groundwater samples from Handan City, with the best-performing LightGBM reaching a test R^2 of 0.753 (RMSE=3.67 mg L^{-1}). Similarly, Alam et al. (2025) reported XGBoost performance below R^2=0.9 for nitrate prediction in the Yinchuan Region, and a comprehensive review concluded that most well-constrained groundwater nitrate ML studies achieve R^2 values between 0.8 and 0.9 (Haggerty et al., 2023). Against these benchmarks, the present framework performs favorably: with in-situ water quality inputs, the virtual-sample-augmented models attained R^2=0.9622, 0.854, and 0.977 (RMSE=3.03 mg L^{-1}) in the normal, dry, and wet seasons, respectively; even with AEF embeddings alone, inputs carrying no direct hydrochemical information, the R^2 values (0.860, 0.674, and 0.784) remain comparable to or exceed purely geospatial models in the literature (Ransom et al., 2022; Karimanzira et al., 2023).

A legitimate concern regarding quantum machine learning is whether the derived quantum features bear any correspondence to actual hydrochemical processes, or whether they merely constitute an abstract mathematical augmentation. In the present architecture, this correspondence is direct and traceable. The Z-feature map first applies a Hadamard gate followed by a single-qubit rotation $Rz(2\phi(xi))$, so that the resulting Pauli-Z expectation value reduces analytically to $\langle Zi \rangle = \cos(2\phi(xi))$, i.e., each quantum feature is a bounded, smooth, one-to-one nonlinear re-expression of an individual measured hydrochemical variable (Havlíček et al., 2019; Schuld and Killoran, 2019). Consequently, every quantum feature inherits the physical identity and units of its parent variable, and its contribution to the prediction can be unambiguously traced back to that variable through SHAP attribution. Physically, the cosine-type response introduced by the quantum encoding is hydrochemically meaningful: whereas classical decision trees partition samples using hard thresholds on raw concentrations, the quantum features provide smooth, saturation-like response curves that resemble the nonlinear equilibria governing nitrate behavior in groundwater, such as carbonate mineral saturation controlling Ca^{2+}-HCO_3^- buffering, the convex conductivity-concentration relationship under evaporative concentration in the dry season, and the redox thresholds (DO/ORP) that delimit nitrification and denitrification regimes. Moreover, because supervised quantum models of this form are mathematically equivalent to kernel methods, the augmented feature space implicitly measures pairwise similarity among samples in a high-dimensional Hilbert space; in hydrochemical terms, this acts as a facies-similarity measure, consistent with the sample groupings independently observed in the Piper and Gibbs diagrams. The concatenated design $[x, \langle Z \rangle]$ further guarantees that no raw information is discarded: the classical features preserve linear, threshold-based mechanisms, while the quantum features add nonlinear, periodic structure, so the model's mechanistic inferences remain anchored to measurable processes such as evaporative concentration, water-rock interaction, cation exchange, and seasonal dilution (Saberian et al., 2025). We nevertheless acknowledge that quantum features are derived transformations rather than independently measurable physical quantities; their mechanistic interpretation must therefore always proceed through the parent hydrochemical variable, as implemented here via SHAP-based attribution (Pérez-Salinas et al., 2020).

4.5 Limitations and future directions

Although the 66, 65, and 50 sampling points for the normal, dry, and wet seasons adequately covered the 3000 hectare farm, discrete points cannot fully resolve fine scale nitrate heterogeneity driven by microtopography, soil texture transitions, and uneven fertilizer application (Li et al., 2025). The seasonal snapshot design also sacrifices temporal resolution: intervals of four to six months may miss episodic recharge, short term leaching pulses after intense rainfall, and transient biogeochemical shifts during freeze thaw cycles characteristic of the monsoonal North China Plain (Nolte et al., 2025). The single year observation period further precludes assessment of interannual climatic variability and legacy nitrogen depletion, so predictions may be less reliable under extreme hydrological years or future precipitation non stationarity. Three systematic biases warrant acknowledgment: sampling through irrigation wells overrepresents shallow, actively exploited aquifer zones relative to deeper stagnant groundwater (Jia et al., 2025); MixSIAR apportionment relies on literature derived endmember signatures that may not capture local fractionation within the thick vadose zone (Jiang et al., 2026); and virtual sample generation may amplify biases inherent in the small original datasets, particularly the overrepresentation of high concentration outliers during the dry season (Tanner et al., 2022).

Although the hybrid quantum classical framework performed well in the Xiong'an New Area, its validity is intrinsically tied to the North China Plain context, and direct transfer to contrasting settings, such as karst aquifers, humid tropical climates, or fertilizer dominated nitrate regimes, would likely suffer domain shift. The single site design also precludes evaluation of transfer learning benefits that multi site pooling could provide (Clark et al., 2025; Farahani et al., 2025). The t-SNE/GMM/KNN methodology itself is portable, but model parameters require recalibration with local hydrochemical data. Future work should therefore combine leave one region out cross validation, domain adaptation techniques (adversarial training or feature alignment), and meta learning over multi site hydrochemical databases to separate universal nitrate transport principles from site specific behavior, enabling cross regional application without extensive new sampling (Zheng et al., 2025).

Models using only AEF embeddings achieve R^2 values roughly 10 to 20% lower than those using measured water quality parameters, because field parameters directly reflect the groundwater chemical state and nitrate transformation processes, whereas remote sensing semantics offer only indirect characterization (Alam et al., 2025). Admittedly, predicting nitrate from field parameters has limited value if every parameter must be measured anyway; however, these parameters can be acquired on site within minutes using portable meters, whereas nitrate requires laboratory analysis, making the field parameter model practical for preliminary screening and rapid decision making when laboratory results are unavailable. After tenfold virtual expansion, the AEF model still attains R^2 above 0.67 (dry), 0.85 (normal), and 0.78 (wet season), supporting its use as a rapid, large scale screening tool, especially in unmonitored areas. Seasonal shifts in dominant embeddings (A05/A00, A08/A06, A02/A03) plausibly encode crop residue or soil organic matter, bare soil exposure, and vegetation or runoff potential, respectively, consistent with MixSIAR apportionment and Bayesian driving factors (Alvarez et al., 2025). Although causal inference remains indirect, the global coverage and annual updates of AEF make it a powerful supplement to, rather than substitute for, monitoring networks, and ground truth validation remains essential for linking remote sensing signals to subsurface water quality (Cai et al., 2025; Tollefson et al., 2025).

5. Conclusion

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.

Code and data availability. All code and data used in this study will be provided in a GitHub repository (https://github.com/sherlockjjobs/Quantum-RF-framework, Junjie Xu, 2026).

Author contributions. **JX**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing (original draft preparation), Writing (review and editing). **XW**: Validation, Formal analysis, Investigation, Resources, Visualization, Writing (original draft preparation). **YY**: Funding acquisition, Project administration, Supervision, Writing-review and editing. **LY** and **XS**: Resources, Supervision. **YZ**: Funding acquisition. **CL**: Investigation, Supervision.

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

interests.

Financial support. 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).

References

Wang, S., Chen, J., Zhang, S., Zhang, X., Chen, D., and Zhou J.: Hydrochemical evolution characteristics, controlling factors, and high nitrate hazards of shallow groundwater in a typical agricultural area of Nansi Lake Basin, North China, Environ. Res., 223, 115430, https://doi.org/10.1016/j.envres.2023.115430, 2023.

Zhang, M., Zhi, Y.Y., Shi, J.C., and Wu, L.S.: Apportionment and uncertainty analysis of nitrate sources based on the dual isotope approach and a Bayesian isotope mixing model at the watershed scale, Sci. Total. Environ., 639, 1175-1187, https://doi.org/10.1016/j.scitotenv.2018.05.239, 2018.

Gu, B.J., Ge, Y., Chang, S.X., Luo, W.D., and Chang, J.: Nitrate in groundwater of China: Sources and driving forces, Global. Environ. Chang., 23(5), 1112-1121, https://doi.org/10.1016/j.gloenvcha.2013.05.004, 2013.

Han, D.M., Currell, M.J., and Cao, G.L.: Deep challenges for China's war on water pollution, Environ. Pollut., 218, 1222-1233, https://doi.org/10.1016/j.envpol.2016.08.078, 2016.

Wang, S.Q., Zheng, W.B., and Kong, X.L.: Spatial distribution characteristics of nitrate in shallow groundwater of the agricultural area of the North China Plain, Chinese Journal of Eco-Agriculture, 26(10), 1476-1482, https://dx.doi.org/10.13930/j.cnki.cjea.180639, 2018.

Sebestyen, S.D., Shanley, J.B., Boyer, E.W., Kendall, C., and Doctor, D.H.: Coupled hydrological and biogeochemical processes controlling variability of nitrogen species in streamflow during autumn in an upland forest, Water. Resour. Res., 50(2), 1569-1591, https://doi.org/10.1002/2013WR013670, 2014.

Thunyawatcharakul, P., Cho, K.H., and Chotpantarat, S.: Predicting Arsenic Speciation in Coastal Aquifers Using Machine Learning: A Case Study of the Chonburi and Rayong Groundwater Basins, Thailand, ACS. EST. Water., 5(9), 5011-5024, https://doi.org/10.1021/acsestwater.4c01082, 2025.

Gao, H., Yang, L., Song, X., Guo, M., Li, B., and Cui, X.: Sources and hydrogeochemical processes of groundwater under multiple water source recharge condition, Sci. Total. Environ., 903, 166660, https://doi.org/10.1016/j.scitotenv.2023.166660, 2023.

Deng, Y., Ye, X., and Du, X.: Predictive modeling and analysis of key drivers of groundwater nitrate pollution based on machine learning, J.Hydrol., 624, 129934, https://doi.org/10.1016/j.jhydrol.2023.129934, 2023.

Cai, C., Zhao, H., Zhang, H., Wang, C., Wang, Z., Liu, M., Chen, J., and Zhang, H.: Timely assessment of maize lodging severity with limited samples using multi-temporal Sentinel-1 and Sentinel-2 data across large spatial extents, Comput. Electron. Agr., 237, 110671, https://doi.org/10.1016/j.compag.2025.110671, 2025.

Feng, D., Liu, J., Lawson, K., and Shen, C.: Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy, Water. Resour. Res., 58(10), https://doi.org/10.1029/2022WR032404, 2022.

Chen, Q., Yang, H., Cui, R., Hu, W., Wang, C., Chen, A., and Zhang, D.: Shallow groundwater table fluctuations: A driving force for accelerating the migration and transformation of phosphorus in cropland soil, Water. Res., 275, 123209, https://doi.org/10.1016/j.watres.2025.123209, 2025.

Liu, X., Yue, F.J., Li, L., Zhou, F., Wen, H., Yan, Z., Wang, L., Wong, W.W., Liu, C.Q., and Li, S.L.: Chronic nitrogen legacy in the aquifers of China, Commun. Earth. Environ., 6(1), 58, https://doi.org/10.1038/s43247-025-02016-7, 2025.

Wu, Y., Wang, J., Liu, Z., Li, C., Niu, Y., and Jiang, X.: Seasonal nitrate input drives the spatiotemporal variability of regional surface water-groundwater interactions, nitrate sources and transformations, J. Hydrol., 655, 132973, https://doi.org/10.1016/j.jhydrol.2025.132973, 2025.

Luo, Y., Yan, J., McClure, S.C., and Li, F.: Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model[J]. Environ. Sci. Pollut. Res., 29(22), 33205-33217, https://doi.org/10.1007/s11356-021-17513-3, 2022.

Wang, C., Yang, J., Zhang, B.: A fault diagnosis method using improved prototypical network and weighting similarity-Manhattan distance with insufficient noisy data, Measurement, 226, 114171, https://doi.org/10.1016/j.measurement.2024.114171, 2024.

Tang, W., Carey, S.K.: Classifying annual daily hydrographs in Western North America using t-distributed stochastic neighbour embedding, Hydrol. Process., 36(1), e14473, https://doi.org/10.1002/hyp.14473, 2022.

Tung, P.Y., Sheikh, H.A., Ball, M., Nabiei, F., and Harrison, R.: SIGMA: Spectral interpretation using gaussian mixtures and autoencoder, Geochem. Geophy. Geosy., 24(1), https://doi.org/10.1029/2022GC010530, 2023.

Farnia, F., Wang, W.W., Das, S., and Jadbabaie.: Gat–gmm: Generative adversarial training for gaussian mixture models, Siam. Math. Data. Sci., 5(1), 122-146, https://doi.org/10.1137/21M1445831, 2023.

Wang, N., Zhou, Q., Gao, J., and Wang, Z.: Evaluating the efficacy of PCA and t-SNE in optimizing input features for groundwater level simulation using machine learning models, Environ. Earth. Sci., 84(12), 336, https://doi.org/10.1007/s12665-025-12357-3, 2025.

Tollefson, J.: Google AI model creates maps of Earth ‘at any place and time’, Nature, 644, 313, https://doi.org/10.1038/d41586-025-02412-1, 2025.

Li, X., Yu, L.: 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, Earth. Syst. Sci. Data., 18, 5601–5626, https://doi.org/10.5194/essd-18-5601-2026, 2026.

Hong, Y.Y., Lopez, D.J.D.: A Review on Quantum Machine Learning in Applied Systems and Engineering, IEEE Access, 13, 144607-144631, https://doi.org/10.1109/ACCESS.2025.3599147, 2025.

Oliveira, Santos.V., Costa, Rocha.P.A., Thé, J.V.G., and Gharabaghi, B.: Optimizing the Architecture of a Quantum–Classical Hybrid Machine Learning Model for Forecasting Ozone Concentrations: Air Quality Management Tool for Houston, Texas, Atmosphere, 16(3), 255, https://doi.org/10.3390/atmos16030255, 2025.

Gujju, Y., Matsuo, A., Raymond, R.: Quantum machine learning on near-term quantum devices: Current state of supervised and unsupervised techniques for real-world applications, Phys. Rev. Appl., 21(6), 067001, https://doi.org/10.1103/PhysRevApplied.21.067001, 2024.

Tian, D., Zhao, X., Gao, L., Jiang, T., Liang, Z., Yang, Z., Zhang, P., Wu, Q., Ren, K., Yang, C., Li, R., Li, S., Cao, Y., Xuan, Y., Chen, J., and Zhu, A.: A framework for tracing the sources of nitrate in surface water through remote sensing data coupled with machine learning, npj Clean Water, 8(1), 43, https://doi.org/10.1038/s41545-025-00473-3, 2025.

Alam, S.M.K., Li, P., Rahman, M., Fida, M., and Elumalai, V.: 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, Environ. Pollut., 364, 125336, https://doi.org/10.1016/j.envpol.2024.125336, 2025.

Liu, M., Geng, D., Wu, L., Min, L., Wang, S., and Shen, Y.: The impact of agricultural land use change on water and nitrate fluxes in the deep vadose zone, the North China Plain, J. Hydrol-Reg. Stud., 62, 102914, https://doi.org/10.1016/j.ejrh.2025.102914, 2025.

Xu, G., Su, X., Yuan, Z., Ji, L., Li, N., and Liang, H.: Nitrogen behavior during artificial groundwater recharge through ponds: A case study in Xiong’an New Area, Environ. Geochem. Heal., 44(8), 2545-2561, https://doi.org/10.1007/s10653-021-01041-7, 2021.

Li, X., Liu, M., Min, L., and Shen, Y.: Nitrogen transport and transformation processes in the typical deep vadose zone in the central North China Plain, Chinese Journal of Eco-Agriculture, 33(12), 2359-2370, https://doi.org/10.12357/cjea.20250148, 2025.

Li, J., Zhu, D., Zhang, S., Yang, G., Zhao, Y., Zhou, C., Lin, Y., and Zou, S.: 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, Ecotox. Environ. Safe., 231, 113205, https://doi.org/10.1016/j.ecoenv.2022.113205, 2022.

Huang, P., Chen, J.: Recharge sources and hydrogeochemical evolution of groundwater in the coal-mining district of Jiaozuo, China, Hydrogeol. J., 20(4), 739-754, https://doi.org/10.1007/s10040-012-0836-4, 2012.

Stock, B.C., Jackson, A.L., Ward, E.J., Parnell, A.C., Phillips, D.L., and Semmens, B.X.: Analyzing mixing systems using a new generation of Bayesian tracer mixing models, PeerJ, 6, https://doi.org/10.7717/peerj.5096, 2018.

Mao, H., Wang, G., Liao, F., Shi, Z., Zhang, H., Chen, X., Qiao, Z., Li, B., and Bai, Y.: Spatial variability of source contributions to nitrate in regional groundwater based on the positive matrix factorization and Bayesian model, J. Hazard. Mater., 445, 130569, https://doi.org/10.1016/j.jhazmat.2022.130569, 2023.

Torres-Martínez, J.A., Mora, A., Mahlknecht, J., Daessle, L.W., Cervantes-Avilés, P.A., and Ledesma-Ruiz, R.L.: 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, Environ. pollut., 269, 115445, https://doi.org/10.1016/j.envpol.2020.115445, 2021.

Jamshidi, E.J, Yusup, Y., Kayode, J.S., and Kamaruddin, M.A.: Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature, Ecol. Inform., 69, 101672, https://doi.org/10.1016/j.ecoinf.2022.101672, 2022.

Islam, M.T., Zhou, Z., Ren, H., Khuzani, M.B., Kapp, D., Zou, J., Tian, L., Liao, J.C., and Xing, L.: Revealing hidden patterns in deep neural network feature space continuum via manifold learning, Nature. Commun., 14(1), 8506, https://doi.org/10.1038/s41467-023-43958-w, 2023.

Liu, H., Yang, J., Ye, M., James, S.C., Tang, Z., Dong, J., and Xing, T.: Using t-distributed Stochastic Neighbor Embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data, J. Hydrol., 597, 126146, https://doi.org/10.1016/j.jhydrol.2021.126146, 2021.

Jia, B., Zhou, J., Tang, Z., Xu, Z., Chen, X., and Fang, W.: Effective stochastic streamflow simulation method based on Gaussian mixture model, J. Hydrol., 605, 127366, https://doi.org/10.1016/j.jhydrol.2021.127366, 2022.

Yan, R., Huang, J.J.: Confident learning-based Gaussian mixture model for leakage detection in water distribution networks, Water. Res., 247, 120773, https://doi.org/10.1016/j.watres.2023.120773, 2023.

Ghodba, A., Richelle, A., McCready, C., Ricardez-Sandoval, L., and Budman, H.: A novel dynamic flux balance analysis for modeling CHO cell fed-batch cultures with pH and temperature shifts, J. Biotechnol., 408, 61-71, https://doi.org/10.1016/j.jbiotec.2025.08.010, 2025.

Niu, H., McCallum, G.B., Chang, A.B., Khan, K., and Azam, S.: Exploring unsupervised feature extraction algorithms: tackling high dimensionality in small datasets, Sci. Rep., 15(1), 21973, https://doi.org/10.1038/s41598-025-07725-9, 2025.

Razavi-Termeh, S.V., Sadeghi-Niaraki, A., Razavi, S., and Choi, S.M.: Enhancing flood-prone area mapping: fine-tuning the K-nearest neighbors (KNN) algorithm for spatial modelling, Int. J. Digit. Earth., 17(1), 2311325, https://doi.org/10.1080/17538947.2024.2311325, 2024.

Abderzak, M., Zeghmar, A., Leila, B., Aziz, M., Velibor, S., Lizny, J., mohamed, K., Fernanda, H., and Shuraik, K.: Ensemble learning-driven optimization of coagulant dosing for drinking water treatment plants using a scalable framework for smart and sustainable process control, Environ. Res., 288, 123229, https://doi.org/10.1016/j.envres.2025.123229, 2025.

Kaur, H., Bansod, B.S., Khungar, P., Dhawan, C.: Combining clustering and ensemble learning for groundwater quality monitoring: a data-driven framework for sustainable water management, Environ. Sci. Pollut. Res., 32, 13862-13903, https://doi.org/10.1007/s11356-025-36477-2, 2025.

Naresh, V.S., Reddi, S.: Quantum-enhanced predictive analytics in healthcare: benchmarking QSVM and QNN on medical datasets, Measurement, 258, 119099, https://doi.org/10.1016/j.measurement.2025.119099, 2025.

Lamichhane, P., Rawat, D.B.: Quantum Machine Learning: Recent Advances, Challenges and Perspectives, IEEE Access, 19, 94057-94105, https://doi.org/10.1109/ACCESS.2025.3573244, 2025.

Vedavyasa, K.V., Kumar, A.: Classification Analysis of Transition Metal Compounds Using Quantum Machine Learning, Adv. Quantum. Technol., 8(5), 2400081, https://doi.org/10.1002/qute.202400081, 2025.

Khalil, M., Zhang, C., Ye, Z., and Zhang, P.: PegasosQSVM: A Quantum Machine Learning Approach for Accurate Fake News Detection, Appl. Artif. Intell., 39(1), 2457207, https://doi.org/10.1080/08839514.2025.2457207, 2025.

Liao, H., Wang, D.S., Sitdikov, I., Salcedo, C., Seif, A., and Minev, Z.K.: Machine learning for practical quantum error mitigation, Nat. Mach. Intell., 6(12), 1478-1486, https://doi.org/10.1038/s42256-024-00927-2, 2024.

Cowlessur, H., Alpcan, T., Thapa, C., Camtepe, S., and Kundu, N.K.: A Qubit-Efficient Hybrid Quantum Encoding Mechanism for Quantum Machine Learning, arXiv preprint arXiv:2506.19275, https://doi.org/10.48550/arXiv.2506.19275, 2025.

Merabet, K., Di, Nunno.F., Granata, F., Kim, S., Adnan, R.M., Heddam, S., Kisi, O., and Zounemat-Kermani, M.: Predicting water quality variables using gradient boosting machine: global versus local explainability using SHapley Additive Explanations (SHAP), Earth. Sci. Inform., 18(3), 298, https://doi.org/10.1007/s12145-025-01796-y, 2025.

Alam, G.M.I., Arfin, Tanim.S., Sarker, S.K., Watanobe, Y., Islam, R., Mridha, M.F., and Nur, K.: Deep learning model based prediction of vehicle CO2 emissions with eXplainable AI integration for sustainable environment, Sci. Rep., 15(1), 3655, https://doi.org/10.1038/s41598-025-87233-y, 2025.

Li, R., Feng, K., An, T., Cheng, P., Wei, L., Zhao, Z., Xu, X., and Zhu, L.: Enhanced insights into effluent prediction in wastewater treatment plants: Comprehensive deep learning model explanation based on shap, ACS. EST. Water., 4(4), 1904-1915, https://doi.org/10.1021/acsestwater.4c00040, 2024.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., and Hutter, F.: Accurate predictions on small data with a tabular foundation model, Nature., 637(8045), 319-326, https://doi.org/10.1038/s41586-024-08328-6, 2025.

Ren, M., Sun, W., Chen, S.: Combining machine learning models through multiple data division methods for PM2.5 forecasting in Northern Xinjiang, China, Environ. Monit. Assess., 193(8), 476, https://doi.org/10.1007/s10661-021-09233-5, 2021.

Gul, N., Khan, A.U., Ullah, B., Khan, B.N., Almalki, H.M., Banga, A.S., Kumar, K.: Optimizing LSTM for sediment load prediction in the Swat River Basin, Pakistan: Evaluation of optimizers and activation functions, Phys. Chem. Earth., 140, 104019, https://doi.org/10.1016/j.pce.2025.104019, 2025.

Alvarez, C.I., Ulloa, Vaca.C.A., Echeverria, Llumipanta.N.A.: Machine learning for urban air quality prediction using Google AlphaEarth Foundations satellite embeddings: A case study of Quito, Ecuador, Remote. Sens., 17(20), 3472, https://doi.org/10.3390/rs17203472, 2025.

Zhu, J.J., Yang, M., Ren, Z.J.: Machine learning in environmental research: common pitfalls and best practices, Environ. Sci. Technol., 57(46), 17671-17689, https://doi.org/10.1021/acs.est.3c00026, 2023.

Saha, G.K., Rahmani, F., Shen, C., Li, L., and Cibin, R.: A deep learning-based novel approach to generate continuous daily stream nitrate concentration for nitrate data-sparse watersheds, Sci. Total. Environ., 878, 162930, https://doi.org/10.1016/j.scitotenv.2023.162930, 2023.

Li, J., Liu, G., Shen, Z.: Integrating spatiotemporal variability of non-point source pollution and best management practice efficiency to improve adaptive watershed management, Water. Res., 284, 124042, https://doi.org/10.1016/j.watres.2025.124042, 2025.

Bigler, M.C., Brusseau, M.L., Guo, B., Jones, S.l., Pritchard, J.C., Higgins, C.P., and Hatton, J.: High-resolution depth-discrete analysis of PFAS distribution and leaching for a vadose-zone source at an AFFF-Impacted site, Environ. Sci. Technol., 58(22), 9863-9874, https://doi.org/10.1021/acs.est.4c01615, 2024.

Udu, A.G., Salman, M.T., Ghalati, M.K., Lecchini-Visintini, A., Siddle, D., and Dong, H.: Emerging SMOTE and GAN-variants for Data Augmentation in Imbalance Machine Learning Tasks: A Review, IEEE Access, 13, 113838-113853, https://doi.org/10.1109/ACCESS.2025.3584532, 2025.

Silva, R., Melo-Pinto, P.: t-SNE: A study on reducing the dimensionality of hyperspectral data for the regression problem of estimating oenological parameters, Artif. Intell. Agr., 7: 58-68, https://doi.org/10.1016/j.aiia.2023.02.003, 2023.

Peng, D., Gui, Z., Wei, W., Li, F., Gui, J., Wu, H., and Gong, J.: Sampling-enabled scalable manifold learning unveils the discriminative cluster structure of high-dimensional data, Nat. Mach. Intell., 7, 1669–1684, https://doi.org/10.1038/s42256-025-01112-9, 2025.

Ranga, D., Rana, A., Prajapat, S., Kumar, P., Kumar, K., and Vasilakos, A.V.: Quantum machine learning: Exploring the role of data encoding techniques, challenges, and future directions, Mathematics, 12(21), 3318, https://doi.org/10.3390/math12213318, 2024.

Alam, M.M.T., Milas, A.S.: Dimensionality Optimized Machine Learning Retrieval of Canopy Chlorophyll, Nitrogen, and Phosphorus from Google Satellite Embeddings, Smart. Agr. Technol., 12, 101601, https://doi.org/10.1016/j.atech.2025.101601, 2025.

Sun, H., Zheng, W., Wang, S., Ma, L., Min, L., and Shen, Y.: Variation of nitrate sources affected by precipitation with different intensities in groundwater in the piedmont plain area of alluvial-pluvial fan, J. Environ. Manage., 367, 121885, https://doi.org/10.1016/j.jenvman.2024.121885, 2024.

Wang, J., Hao, X., Liu, X., Ouyang, W., Li, T., Cui, X., Pei, J., Zhang, S., Zhu, W., and Jin, R.: Groundwater–surface water exchange affects nitrate fate in a seasonal freeze–thaw watershed: Sources, migration and removal, J. Hydrol., 654, 132803, https://doi.org/10.1016/j.jhydrol.2025.132803, 2025.

Wang, J., Wang, M., Zhang, C., Fan, J., and Lv, F.: Vertical Migration Characteristics and Effect Factors of BaP in Contaminated Soil Under Rainwater Infiltration, Water. Air. Soil. Pollut., 236(14), 915, https://doi.org/10.1007/s11270-025-08578-8, 2025.

Wu, H., Song, F., Min, L., Li, J., Shen, Y., Huang, Y., Fan, H., Liu, J., and Fu, S.: Exploring recharge mechanisms of soil water in the thick unsaturated zone using water isotopes in the

North China Plain, Catena, 234, 107615, https://doi.org/10.1016/j.catena.2023.107615, 2024.

Liu, S., Hao, Y., Wang, H., Zheng, X., Yu, X., Meng, X., Qiu, Y., Li, S., and Zheng, T.: Bidirectional potential effects of DON transformation in vadose zones on groundwater nitrate contamination: Different contributions to nitrification and denitrification, J. Hazard. Mater., 448, 130976, https://doi.org/10.1016/j.jhazmat.2023.130976, 2023.

Ahmed, M.A., Abdel, Samie.S.G., Badawy, H.A.: Factors controlling mechanisms of groundwater salinization and hydrogeochemical processes in the Quaternary aquifer of the Eastern Nile Delta, Egypt, Environ. Earth. Sci., 68(2), 369-394, https://doi.org/10.1007/s12665-012-1744-6, 2013.

Zhang, W., Xin, C., Yu, S.: A review of heavy metal migration and its influencing factors in karst groundwater, Northern and Southern China, Water., 15(20), 3690, https://doi.org/10.3390/w15203690, 2023.

Zhang, J., Zhang, L., Zheng, T., Jin, M., Kang, F., Jiang, J., Yuan, Z., and Luo, J.: Tracing Nitrate Contamination Sources and Transformations in a Rural-Urban Karst Groundwater System in North China Using Multiple Isotopes and Simmr Modeling, Water. Res. Res., 61(10), https://doi.org/10.1029/2025WR040156, 2025.

Kobak, D., Berens, P.: The art of using t-SNE for single-cell transcriptomics, Nat. commun., 10(1), 5416, https://doi.org/10.1038/s41467-019-13056-x, 2019.

Di, D., He, C., Chen, F., and Giovannini, L.: ML-AMPSIT: Machine learning-based automated multi-method parameter sensitivity and importance analysis tool, Geosci. Model. Dev, 18, 433-459, https://doi.org/10.5194/gmd-18-433-2025, 2025.

Huang, C., Tong, J., Ye, M.: Global sensitivity analysis for a prediction model of soil solute transfer into surface runoff, J. Hydrol., 599, 126342, https://doi.org/10.1016/j.jhydrol.2021.126342, 2021.

Anderson, G.J., Lucas, D.D.: Machine learning predictions of a multiresolution climate model ensemble, Geophys. Res. Lett., 45(9), 4273-4280, https://doi.org/10.1029/2018GL077049, 2018.

Dega, S., Dietrich, P., Schroen, M., and Paasche, H.: 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, Front. Environ. Sci., 11, 1009191, https://doi.org/10.3389/fenvs.2023.1009191, 2023.

Xiong, X., Hao, G., Xu, L., Li, Z., Zhang, X., Lu, J., and Zhang, Z.: Characteristics of spatial distributions for nitrate and traceability of pollution in shallow groundwater of the Qingshui River Basin, China, Hydrol. Res., 56(9): 920-936, https://doi.org/10.2166/nh.2025.047, 2025.

Hou, X., Peng, L., Zhang, Y., Zhang, Y., Wang, Y., Feng, W., and Yang, H.: A Data-Driven

Method for Determining DRASTIC Weights to Assess Groundwater Vulnerability to Nitrate: Application in the Lake Baiyangdian Watershed, North China Plain, Appl. Sci., 15(5), 2866, https://doi.org/10.3390/app15052866, 2025.

Liu, N., Chen, M., Gao, D., Wu, Y., and Wang, X.: Identification of hydrogeochemical processes in shallow groundwater using multivariate statistical analysis and inverse geochemical modeling, Environ. Monit. Assess, 197(2), 135, https://doi.org/10.1007/s10661-024-13528-8, 2025.

Li, X., Zhou, J., Zhou, W., Mao, L., Wang, C., Hao, Y., and Bian, P.: Hydrochemical Characteristics of Shallow Groundwater and Analysis of Vegetation Water Sources in the Ulan Buh Desert, Water., 17(21), 3058, https://doi.org/10.3390/w17213058, 2025.

Nolte, A., Heudorfer, B., Bender, S., and Hartmann, J.: Multi-site deep learning for groundwater level prediction across global datasets: toward scalable applications under data scarcity, J. Hydroinform., 27(10), 1632-1651, https://doi.org/10.2166/hydro.2025.095, 2025.

Jia, H., Qian, H.: Groundwater nitrate response to hydrogeological conditions and socioeconomic load in an agriculture dominated area, Sci. Rep., 15(1), 1315, https://doi.org/10.1038/s41598-024-84318-y, 2025.

Jiang, H., Geng, D., Fu, J., Liu, M., Qi, Y., Min, L., Wang, S., and Shen, Y.: Quantifying nitrate dynamics in deep vadose zone under irrigated farmland in the North China Plain: Insights from continuous in-situ monitoring, Agr. Ecosyst. Environ., 396, 109956, https://doi.org/10.1016/j.agee.2025.109956, 2026.

Tanner, K.B., Cardall, A.C., Williams, G.P.: A spatial long-term trend analysis of estimated chlorophyll-a concentrations in Utah Lake using Earth observation data, Remote. Sens., 14(15), 3664, https://doi.org/10.3390/rs14153664, 2022.

Farahani, M.A., Wood, A.W., Tang, G., and Mizukami, N.: Calibrating a large-domain land/hydrology process model in the age of AI: the SUMMA CAMELS emulator experiments, Hydrol. Earth. Syst. Sci., 29(18), 4515-4537, https://doi.org/10.5194/hess-29-4515-2025, 2025.

Clark, S.R., Jaffres, J.B.D.: Associations between deep learning runoff predictions and hydrogeological conditions in Australia, J. Hydrol., 651, 132569, https://doi.org/10.1016/j.jhydrol.2024.132569, 2025.

Su, Y., Wu, Z., Zheng, X., Qiu, Y., Ma, Z., Ren, Y., and Bai, Y.: Harmonizing remote sensing and ground data for forest aboveground biomass estimation, Ecol. Inform., 86, 103002, https://doi.org/10.1016/j.ecoinf.2025.103002, 2025.

Ma, Z., Ye, C., Lu, C., Wei, Q., Zhu, R., Xie, X., Zhong, S., Chu, W., and Xu, Z.: A Robust Gaussian Process Paradigm for Predictive Modeling on Small Data sets in Environmental

Science: A Case Study in Ballasted Flocculation, Environ. Sci. Technol., 60(1), 748-759, https://doi.org/10.1021/acs.est.5c12617, 2025.

Charizanos, G., Demirhan, H.: Bayesian prediction of wildfire event probability using normalized difference vegetation index data from an Australian forest, Ecol. Inform., 73, 101899, https://doi.org/10.1016/j.ecoinf.2022.101899, 2023.

Addy, J.W.G., MacLaren, C., Lang, R.: A Bayesian approach to analyzing long-term agricultural experiments, Eur. J. Agron., 159, 127227, https://doi.org/10.1016/j.eja.2024.127227, 2024.

Weller, D.L., Murphy, C.M., Johnson, S., Green, H., Michalenko, E., Love, T.M., and Strawn, L.K.: Land use, weather, and water quality factors associated with fecal contamination of northeastern streams that span an urban-rural gradient, Front. Water., 3, 741676, https://doi.org/10.3389/frwa.2021.741676, 2022.

Liu, M., Min, L., Wu, L., Pei, H., and Shen, Y.: Evaluating nitrate transport and accumulation in the deep vadose zone of the intensive agricultural region, North China Plain, Sci. Total. Environ., 825, 153894, https://doi.org/10.1016/j.scitotenv.2022.153894, 2022.

Miao, P., Zhu, X., Zhang, S., Li, W., Zhou, J., and Chen, Z.: 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, Agr. Ecosyst. Environ., 357, 108682, https://doi.org/10.1016/j.agee.2023.108682, 2023.

Williams, M.R., Ford, W.I., Mumbi, R.C.K.: Preferential flow in the shallow vadose zone: Effect of rainfall intensity, soil moisture, connectivity, and agricultural management, Hydrol. Process., 37(12), 15057, https://doi.org/10.1002/hyp.15057, 2023.

Yang, W., Long, D., Scanlon, B.R., Burek, P., Zhang, C., Han, Z., Butler, J., Pan, Y., Lei, X., and Wada, Y.: Human intervention will stabilize groundwater storage across the North China Plain, Water. Res. Res., 58(2), https://doi.org/10.1029/2021WR030884, 2022.

Feng, W., Wang, S., Tan, K., Ma, L., and Hu, C.: Simulation of spatial and temporal variation of nitrate leaching in the vadose zone of alluvial regions on a large regional scale, Sci. Total. Environ., 916, 170114, https://doi.org/10.1016/j.scitotenv.2024.170114, 2024.

Saberian, M., Zafarmomen, N., Neupane, A., Panthi, K., and Samadi, V.: HydroQuantum: A new quantum-driven Python package for hydrological simulation, Environ. Modell. Softw, 195, 106736, https://doi.org/10.1016/j.envsoft.2025.106736, 2025.

Zheng, Y., Zhang, X., Zhou, Y., Zhang, Y., Zhang, T., and Farmani, R.: Deep representation learning enables cross-basin water quality prediction under data-scarce conditions[J]. NPJ Clean Water, 8(1), 33, https://doi.org/10.1038/s41545-025-00466-2, 2025.

Havlíček, V., Córcoles, A.D., Temme, K., Harrow, A., Kandala, A., Chow, J.M., and Gambetta, J.M.: Supervised learning with quantum-enhanced feature spaces, Nature, 567(7747), 209-212, https://doi.org/10.1038/s41586-019-0980-2, 2019.

Schuld, M., Killoran, N.: Quantum machine learning in feature Hilbert spaces, Phys. rev. lett., 122(4), 040504, https://doi.org/10.1103/PhysRevLett.122.040504, 2019.

Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E., and Latorre, J.: Data re-uploading for a universal quantum classifier, Quantum, 4, 226, https://doi.org/10.22331/q-2020-02-06-226, 2020.

Li, Y., Jiang, Z., Wu, S., Gao, J., and Sharma, A.: Systematic evaluation and correction of extreme water level in global storm surge simulation, Geophys. Res. Lett., 53(6), https://doi.org/10.1029/2025GL119793, 2026.

Michalek, A.T., Villarini, G., Kim, T., Quintero, F., Krajewski, W., and Scoccimarro, E.: Evaluation of CMIP6 HighResMIP for hydrologic modeling of annual maximum discharge in Iowa, Water. Res. Res., 59(8), https://doi.org/10.1029/2022WR034166, 2023.

Li, X., Wang, Y., Xue, B., A, Y., Zhang, X., and Wang, G.: Attribution of runoff and hydrological drought changes in an ecologically vulnerable basin in semi-arid regions of China, Hydrol. Process, 37(10), e15003, https://doi.org/10.1002/hyp.15003, 2023.

Van, P., Fox-Kemper, B., Seroussi, H., Nowicki, S., and Bergen, K.: A variational LSTM emulator of sea level contribution from the Antarctic ice sheet, J. Adv. Model. Earth. Sy., 15(12), https://doi.org/10.1029/2023MS003899, 2023.

Balogun, E., Rajagopal, R., Majumdar, A.: TemperatureGAN: generative modeling of regional atmospheric temperatures, Environ. Data. Sci., 3, https://doi.org/10.1017/eds.2024.21, 2024.

Kalinin, A.A., Arevalo, J., Serrano, E., Vulliard, L., Tsang, H., Bornholdt, M., Muñoz, A., Sivagurunathan, S., Rajwa, B., Carpenter, A., Way, G., and Singh, S.: A versatile information retrieval framework for evaluating profile strength and similarity, Nat. Commun., 16(1), 5181, , 2025.

Li, S., Zheng, T., Farchi, A., Bocquet, M., and Gentine, P.: Probabilistic data assimilation for ensemble distribution projections with generative machine learning: A Lorenz'96 proof-of-concept, Geophys. Res. Lett, 52(12), https://doi.org/10.1029/2024GL112523, 2025.

An, B., Zhang, Z., Ren, J., and Zhang, W.: Recurrent adversarial learning for geo-technical time-series augmentation: application to slope instability forecasting in open-pit mines, Environ. Earth. Sci., 84, 559, https://doi.org/10.1007/s12665-025-12566-w, 2025.

Zhao, Y., Liu, J., Zhang, X., Li, Q., and Wu, J.: Integrated Machine Learning and Health Risk Assessment for Groundwater Nitrate Contamination in Handan City, China, Water., 18, 1174, https://doi.org/10.3390/w18101174, 2026.

Haggerty, R., Sun, J., Yu, H., and Li, Y.: Application of machine learning in groundwater quality modeling-A comprehensive review, Water. Res., 233, 11974, https://doi.org/10.1016/j.watres.2023.119745, 2023.

Ransom, K.M., Nolan, B.T., Stackelberg, P.E., Belitz, K., and Fram, M.S.: Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States, Sci. Total. Environ., 807, 151065, https://doi.org/10.1016/j.scitotenv.2021.151065, 2022.

Karimanzira, D., Weis, J., Wunsch, A., Ritzau, L., Liesch, T., and Ohmer, M.: Application of machine learning and deep neural networks for spatial prediction of groundwater nitrate concentration to improve land use management practices, Front. Water., 5, 1193142, https://doi.org/10.3389/frwa.2023.1193142, 2023.