

Article

Data-Driven Risk Identification and Prevention–Control Optimization of Groundwater Nitrate Contamination

Xiangbin Kong ¹, Qun Li ², Jie Wu ^{3,*}, Jing Liu ⁴, Yuanzheng Zhai ^{5,*} and Tianyi Zhang ⁶¹ Handan Ecological Environment Bureau, Handan 056000, China; kxbnvl@163.com² Key Laboratory of Soil Environmental Management and Pollution Control, Nanjing Institute of Environmental Sciences, Ministry of Ecology and Environment of the People's Republic of China, Nanjing 210042, China; liqun@nies.org³ College of Architecture and Civil Engineering, Beijing University of Technology, Beijing 100124, China⁴ China National Environmental Monitoring Centre, Beijing 100012, China; liujing1@cnemc.cn⁵ College of Water Sciences, Beijing Normal University, Beijing 100875, China⁶ China Academy of Urban Planning and Design, Beijing 100044, China; caupde123@163.com

* Correspondence: wujieajmdw@126.com (J.W.); diszyz@163.com (Y.Z.)

Abstract

Groundwater nitrate contamination is delayed, spatially heterogeneous and difficult to screen with sparse monitoring, especially when exceedance probability and exposure concern must be considered together. We analysed 157 shallow groundwater samples from Handan City, China, collected in 2023, and 14 environmental predictors to map NO_3^- exceedance above 50 mg/L. Random forest, support vector machine (SVM) and XGBoost classifiers were evaluated by repeated stratified testing and spatial block cross-validation, with SHAP used to interpret model responses. Ordinary kriging and the USEPA non-carcinogenic risk framework estimated adult and child hazard quotients (HQs), and exceedance probability was cross-classified with child HQ in a probability–HQ matrix. Thirty-one samples (19.7%) exceeded the threshold. Although SVM achieved the highest ROC-AUC, XGBoost was retained for early-warning mapping because it better controlled missed exceedances. High-probability zones were concentrated in central-western and urban-fringe Handan. SHAP responses indicated that hydroclimatic, hydrogeological, land-use, terrain and soil factors jointly shaped model discrimination, without implying source attribution. Children had higher HQs than adults. At $P = 0.5$, general protection, exceedance-warning, concentration-verification and priority-intervention zones occupied 81.26%, 18.40%, 0.01% and 0.33% of the area, respectively. The framework supports targeted monitoring and drinking-water verification rather than fixed contamination boundaries.

Keywords: groundwater nitrate; exceedance probability; machine learning; health risk assessment; probability–HQ matrix; groundwater risk zoning

Academic Editor: Dimitrios E. Alexakis

Received: 12 June 2026

Revised: 21 July 2026

Accepted: 23 July 2026

Published: 24 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Groundwater supports urban and rural water supply, agricultural irrigation and ecosystem functioning, making its quality central to regional water security and public health [1,2]. Nitrate (NO_3^-) is among the most widespread contaminants in shallow groundwater and is commonly associated with fertilizer application, livestock manure, sewage leakage and urban expansion [3–6]. Its high mobility facilitates transport to aquifers through rain-

fall infiltration and irrigation return flow, producing delayed and spatially heterogeneous contamination [5,7,8]. The World Health Organization recommends a drinking-water nitrate limit of 50 mg/L [9], and prolonged exposure may threaten vulnerable populations, particularly children. Groundwater nitrate patterns are shaped by interacting controls, including nitrogen loading, land use, soil properties, groundwater depth, recharge and hydrogeological setting [7,10–12].

Across the North China Plain and adjacent regions, intensive agriculture, rapid urbanisation and long-term groundwater abstraction have substantially altered shallow groundwater systems. Groundwater nitrate has consequently become a major regional water-quality concern [2,13,14]. However, sparse monitoring networks provide only discrete observations and cannot fully resolve continuous patterns of exceedance risk [15,16]. Conventional interpolation relies mainly on spatial autocorrelation and uses environmental information only indirectly. Vulnerability frameworks such as DRASTIC identify pollution-prone areas, but fixed weights and predefined relationships limit their representation of nonlinear controls and interactions [17,18]. Machine learning offers an alternative by integrating multi-source environmental predictors to estimate spatially explicit nitrate exceedance probability [15,19–23].

Despite its promise, machine learning for regional nitrate assessment requires careful evaluation of spatial transferability and interpretability. Random data partitioning can place neighbouring samples in both training and test sets, inflating apparent performance in unsampled areas [24,25]. Spatial cross-validation offers a more rigorous assessment under spatially independent conditions [24,25]. Machine-learning models also identify statistical associations rather than direct causal mechanisms or contaminant sources. Explainable methods such as SHAP quantify the contribution and response direction of individual predictors, thereby improving model transparency [26]. However, SHAP patterns cannot replace isotopic tracers, source inventories or hydrogeochemical evidence for source attribution [5,27–29].

Groundwater nitrate assessment should extend beyond exceedance prediction to consider concentration-derived exposure concern. Non-carcinogenic risk frameworks estimate hazard quotients (HQs) for different population groups using nitrate concentrations and drinking-water exposure parameters [30–32]. Children are generally more susceptible because of their lower body weight and greater water intake per unit body mass [33,34]. Exceedance probability and HQ are related assessment products rather than independent risk dimensions. Exceedance probability provides an early-warning estimate of where NO_3^- is likely to exceed 50 mg/L, whereas child HQ identifies areas where interpolated nitrate concentrations imply greater exposure concern for a vulnerable population. A probability–HQ matrix can therefore be used as a transparent management-classification tool to distinguish areas where these two signals coincide or diverge, without combining them into a single additive composite index [35,36].

Previous studies have made substantial progress in groundwater nitrate assessment from several perspectives. Machine-learning models have been increasingly used to predict nitrate concentrations or identify elevated nitrate risk by integrating land use, soil, climate, topographic and hydrogeological predictors [19–23]. Groundwater vulnerability and pollution-risk mapping have also helped identify areas susceptible to contamination by combining hydrogeological sensitivity and pollution pressure [17,18]. In parallel, health risk assessment studies have estimated non-carcinogenic risks for different population groups based on measured or interpolated nitrate concentrations and exposure parameters [30–34]. More recent studies have further explored probabilistic risk mapping and management-oriented groundwater zoning [35,36]. However, these advances still leave several methodological issues relevant to regional early-warning management. First, many nitrate prediction studies focus mainly on concentration estimation or overall

predictive accuracy, whereas spatially explicit exceedance probability is less often used as a direct early-warning product. Second, model validation is still often based on random partitioning, which may overestimate predictive performance when groundwater samples are spatially clustered; therefore, spatial transferability requires more explicit evaluation. Third, health risk assessments are commonly derived from measured or interpolated nitrate concentrations, but their relationship with machine-learning-derived exceedance probability is not always clearly represented in management zoning. To address these gaps, the methodological innovation lies in a transparent, non-additive probability–HQ matrix that integrates spatially validated, interpretable nitrate-exceedance probability modelling with concentration-derived child HQ. The matrix preserves the distinct management meanings of exceedance likelihood and exposure concern, revealing where the two signals coincide or diverge. These patterns support the delineation of general-protection, exceedance-warning, concentration-verification, and priority-intervention zones for targeted management. We further characterise the uncertainty associated with the resulting zones by quantifying variability in predicted exceedance probabilities, kriging-derived HQ uncertainty, and sensitivity to the probability threshold and the HQ = 1 classification criterion. We apply the framework to 157 shallow groundwater samples and 14 environmental predictors from Handan City, providing a transparent basis for screening, verification, and spatially differentiated groundwater nitrate management.

2. Materials and Methods

2.1. Study Area

Handan City is located in southern Hebei Province, where the eastern foothills of the Taihang Mountains grade into the North China Plain. Elevation generally decreases from west to east, accompanied by a transition from bedrock and karst-fissure groundwater systems in the western hills to Quaternary porous aquifers in the central and eastern plains. In the plain area, shallow groundwater occurs mainly within alluvial sediments of silt, silty clay and sand, and is closely linked to seasonal rainfall, river leakage and irrigation return flow. Deeper groundwater is more confined and is replenished more slowly through lateral flow and leakage between aquifer layers. Groundwater flow is broadly eastward to northeastward. The region has a warm-temperate continental monsoon climate and receives approximately 500–600 mm of precipitation annually. More than 60% of precipitation falls from June to August, while winters are comparatively dry. Concentrated summer rainfall can promote nitrogen leaching into shallow aquifers [10,12], whereas strong evapotranspiration may favour nitrate accumulation within the soil–groundwater continuum. Intensive agriculture, urban development and dense human activity further increase nitrogen inputs through fertilizer use, irrigation return flow, domestic sewage and land development [5,10,14]. The location of the study area and groundwater sampling sites is shown in Figure 1.

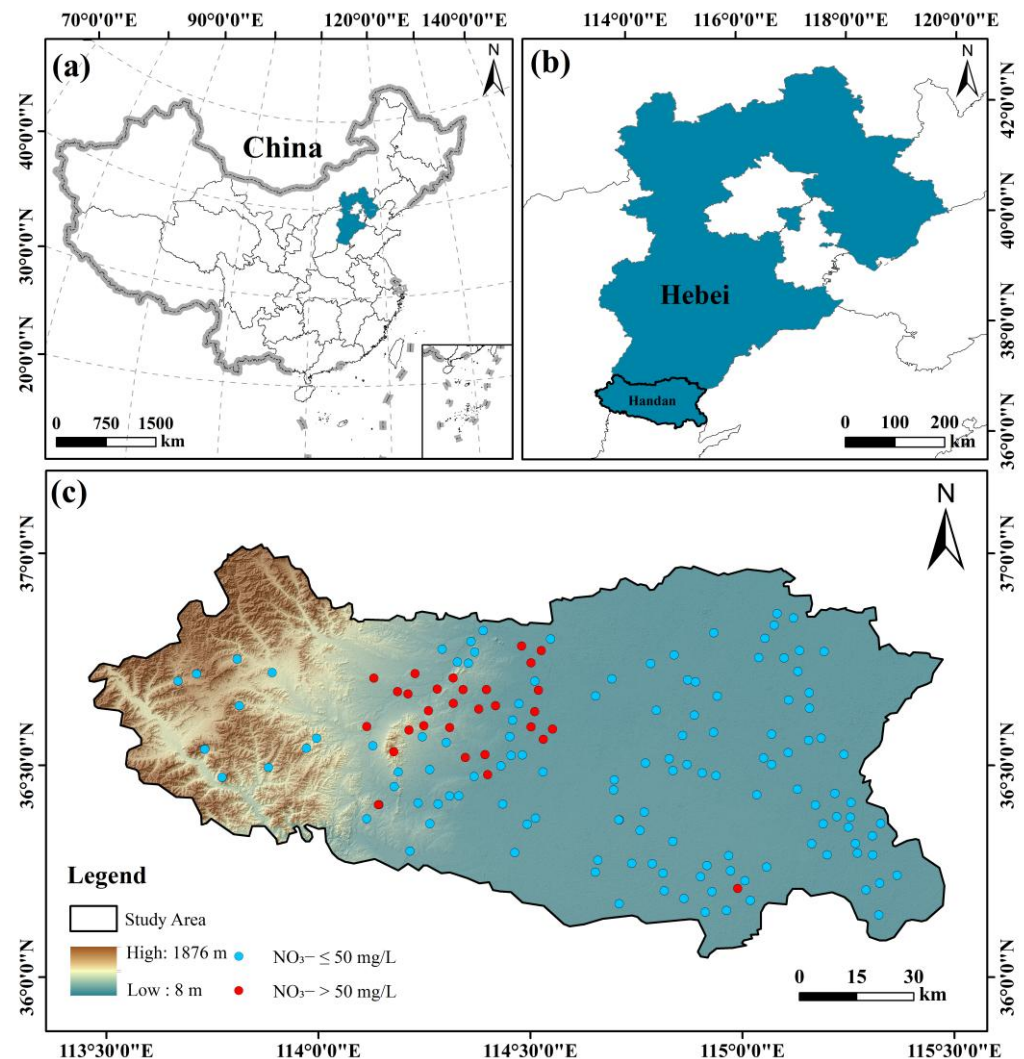


Figure 1. Location of the study area and groundwater sampling sites: (a) China; (b) Hebei Province; (c) Handan City.

2.2. Data Sources and Environmental Predictors

The core dataset comprised 157 shallow groundwater samples collected across Handan City in 2023. Sampling locations covered diverse geomorphic, hydrogeological and anthropogenic settings to represent spatial variation in groundwater NO_3^- . Exceedance was defined as $\text{NO}_3^- > 50 \text{ mg/L}$ [9], and samples were coded as 1 for exceedance and 0 otherwise. Thirty-one samples exceeded the threshold, giving an exceedance rate of 19.7%. To prevent target leakage, NO_3^- concentration was used only to define exceedance labels and estimate health risk. It was excluded from the machine-learning predictor set.

We selected 14 environmental predictors based on both the known hydrogeochemical controls on groundwater nitrate and variables commonly used in previous nitrate-prediction and groundwater-vulnerability studies [7,10–12,17–23]. The predictors were chosen to represent four key processes affecting nitrate occurrence in shallow groundwater: nitrogen input, leaching and transport, dilution or enrichment, and hydrogeological susceptibility. Selection also considered spatial coverage, data quality, representativeness, limited redundancy and availability at the regional scale. Soil predictors included clay content, pH, bulk density and cation-exchange capacity, which influence infiltration, solute transport and nitrogen transformation. Human-activity predictors comprised population density and built-up, cropland and vegetation fractions, representing domestic and agricultural nitrogen inputs and land-use pressure. Climate variables included mean an-

nual precipitation, potential evapotranspiration and temperature, reflecting recharge, evaporative concentration and hydrothermal conditions. Elevation, slope and groundwater table depth represented topographic and hydrogeological controls on runoff, infiltration, vadose-zone thickness and aquifer susceptibility.

Multi-year or long-term averages were used for soil clay content, pH, bulk density, cation-exchange capacity, precipitation, potential evapotranspiration, temperature and groundwater table depth. Elevation and slope were treated as static terrain attributes. Population density from 2022 represented human activity near the sampling period, and 2021 land-cover data were used to derive built-up, cropland and vegetation fractions. Table 1 summarises the data sources, spatial resolutions and temporal coverage.

Table 1. Environmental predictors and data sources.

Predictor Category	Predictor	Spatial Resolution	Time Period	Data Source
Soil property	Soil clay content	250 m	2000–2020	SoilGrids, ISRIC [37]
	Soil pH	250 m	2000–2020	SoilGrids, ISRIC
	Soil bulk density	250 m	2000–2020	SoilGrids, ISRIC
	Cation exchange capacity	250 m	2000–2020	SoilGrids, ISRIC
Human activity	Population density	100 m	2022	WorldPop [38]
	Built-up land fraction	10 m	2021	ESA WorldCover 2021 [39]
	Cropland fraction	10 m	2021	ESA WorldCover 2021
	Vegetation cover fraction	10 m	2021	ESA WorldCover 2021
Meteorological condition	Mean annual precipitation	1 km	2000–2020	TerraClimate [40]
	Mean annual potential evapotranspiration	1 km	2000–2020	TerraClimate
	Mean annual temperature	1 km	2000–2020	TerraClimate
Topographic condition	DEM	30 m	—	ASTER GDEM V3 [41]
	Slope	30 m	—	ASTER GDEM V3
Hydrogeological condition	Groundwater table depth	1 km	2005–2022	National Tibetan Plateau Data Center [42]

Figure 2 summarises the analytical workflow. We first derived NO_3^- exceedance labels from the 2023 groundwater samples and extracted multi-source environmental predictors. Machine-learning models were then used to map nitrate exceedance probability, and spatial cross-validation was applied to assess model transferability. SHAP analysis was used to interpret the principal environmental responses captured by the model. Finally, child HQ derived from interpolated NO_3^- concentrations was cross-classified with exceedance probability using a probability–HQ matrix to delineate management-oriented groundwater nitrate risk zones.

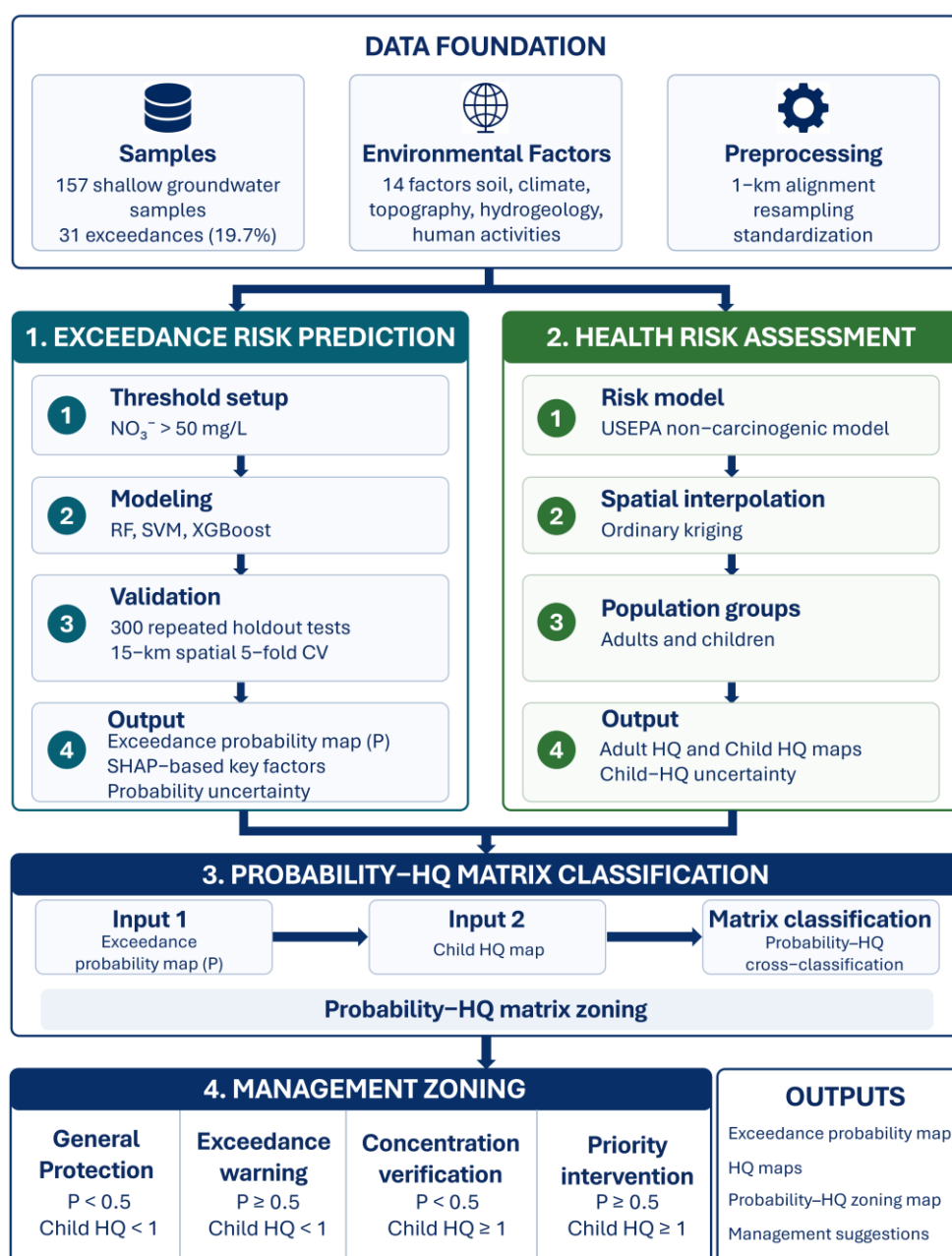


Figure 2. Analytical workflow of the study.

2.3. Data Preprocessing and Site-Specific Predictor Extraction

Before model development, groundwater samples, environmental predictors and land-cover data were harmonised within a common spatial framework. All datasets were referenced to GCS_WGS_1984, clipped to the study area, resampled to a common resolution and aligned to a consistent raster grid. Continuous rasters were resampled to an approximately 1 km analytical scale using bilinear interpolation. Categorical land-cover data were not interpolated. Instead, class proportions were calculated within local buffers around sampling sites to avoid artificial mixing of categorical information. The processed raster layers for the 14 environmental predictors are provided in Figure A1.

Continuous predictors describing soil, climate, topography and groundwater table depth were extracted at each sampling location. Built-up, cropland and vegetation fractions were calculated from 2021 land-cover data within a 1000 m circular buffer around each site. Population density was extracted from the 2022 raster and log-transformed to reduce skewness and the influence of extreme values. The final site-level dataset con-

tained sample identifiers, NO_3^- concentrations, binary exceedance labels and 14 environmental predictors.

After extraction, variables were checked for unit consistency, plausible ranges and completeness. The few missing predictor values were imputed using neighbourhood medians, allowing all 157 samples to be retained. Continuous predictors were screened for outliers, adjusted for extreme values where necessary and standardised using Z-score normalisation. Random forest, SVM and XGBoost were trained and evaluated with the same processed predictor set to ensure comparability.

All spatial preprocessing, statistical analysis, machine-learning modelling and visualization were performed using Python version 3.12.2 with pandas version 2.2.2, NumPy version 1.26.4, scikit-learn version 1.3.2, xgboost version 2.0.3, imbalanced-learn version 0.11.0, shap version 0.47.2 and matplotlib version 3.9.0rc2. Spatial mapping and raster processing were conducted using ArcGIS Desktop version 10.8.

2.4. Machine-Learning Modelling, Spatial Prediction and SHAP Interpretation

2.4.1. Model Development and Performance Assessment

Groundwater nitrate exceedance was formulated as a binary classification problem. The 14 environmental predictors served as model inputs, and NO_3^- exceedance above 50 mg/L was the response. The resulting probability, P , represented the likelihood of nitrate exceedance at each location. Three widely used classifiers were evaluated: random forest (RF), support vector machine (SVM) and extreme gradient boosting (XGBoost) [43–45]. Samples were repeatedly divided into training and test sets at an 80:20 ratio using stratified random sampling. Three hundred repetitions reduced dependence on any single split. Within each repetition, hyperparameters were optimised by random search with stratified five-fold cross-validation using only the training data [46]. Performance and stability were summarised as the mean and standard deviation across the 300 independent test evaluations.

Because exceedance samples were relatively scarce, the synthetic minority over-sampling technique (SMOTE) was applied during training to improve sensitivity to the minority class [47]. To prevent data leakage, SMOTE was restricted to the training set or to training folds within cross-validation. Test sets and validation folds retained their original class distributions. Performance was assessed using ROC-AUC, accuracy, precision, recall and F1-score [48]. Model selection considered both overall discrimination and the early-warning objective of limiting false negatives. Metrics were reported at probability thresholds of 0.5 and 0.7 to compare conventional and high-confidence decision settings.

2.4.2. Spatial Cross-Validation and Exceedance-Probability Mapping

Spatial transferability was evaluated using five-fold spatial block cross-validation with 15 km × 15 km grid-based groups [24,25]. The study area was first divided into regular spatial blocks, and each groundwater sample was assigned to a block according to its geographic coordinates. Blocks, rather than individual samples, were then allocated to five folds. In each fold, samples located in the held-out blocks were used for validation, whereas samples in the remaining blocks were used for model training and hyperparameter optimisation. SMOTE was applied only to the training samples within each fold and was not applied to the held-out validation samples.

For spatial prediction, the classifier retained after model evaluation was retrained in 300 repeated runs and applied to the full predictor raster stack. In each run, the model was trained using the available labelled groundwater samples and then used to predict nitrate exceedance probability for each raster pixel. The pixel-wise mean probability from the 300 runs was used as the final exceedance-probability map, while the pixel-wise standard deviation was used as a supplementary prediction-variability layer.

2.4.3. SHAP-Based Model Interpretation

SHAP analysis was applied to the final spatial prediction model to improve interpretability. Based on game theory, SHAP decomposes a model output into predictor-level contributions [26]. It therefore quantifies how individual environmental variables influence the predicted probability of nitrate exceedance.

Mean absolute SHAP values quantified the overall contribution of each predictor and identified the variables most influential to model discrimination. Predictor values were also examined against their SHAP values to assess response direction and nonlinearity. These analyses describe the environmental patterns used by the model to distinguish potential nitrate exceedance. Because several predictors were spatially structured and potentially collinear, SHAP results were interpreted only as model-response patterns rather than evidence of independent causal drivers.

2.5. Health Risk Assessment and Probability–HQ Matrix Construction

2.5.1. Health Risk Assessment

The USEPA non-carcinogenic risk framework was used to estimate hazard quotients (HQs) for adults and children exposed to groundwater NO_3^- through drinking-water ingestion [30–32]. The average daily dose (ADD) was calculated from nitrate concentration, water intake, exposure frequency, exposure duration, body weight and averaging time:

$$\text{ADD} = \frac{C \times IR \times EF \times ED}{BW \times AT} \quad (1)$$

where ADD denotes the average daily dose, $\text{mg}/(\text{kg} \cdot \text{d})$; C is nitrate concentration expressed as NO_3^- , mg/L . To maintain unit consistency, both nitrate concentration and the oral reference dose were expressed on a nitrate-ion basis. Measured NO_3^- concentrations were therefore used directly for HQ calculation, and the RfD was set to $7.09 \text{ mg NO}_3^-/(\text{kg} \cdot \text{d})$, equivalent to $1.6 \text{ mg NO}_3^-/\text{N}/(\text{kg} \cdot \text{d})$. IR is daily drinking-water intake, L/d ; EF is exposure frequency, d/a ; ED is exposure duration, a ; BW is average body weight, kg ; and AT is averaging time, d .

The non-carcinogenic hazard quotient (HQ) was then calculated as follows:

$$\text{HQ} = \frac{\text{ADD}}{\text{RfD}} \quad (2)$$

where RfD denotes the oral reference dose for nitrate, $\text{mg}/(\text{kg} \cdot \text{d})$. An HQ value greater than 1 indicates a potential non-carcinogenic health risk [30,32]. Separate exposure parameters were assigned to adults and children to account for differences in water intake, body weight and exposure duration. Given their higher water intake per unit of body mass and greater susceptibility to environmental exposure, children were considered the priority vulnerable population in assessing groundwater nitrate health risk [33,34]. Parameter values are summarized in Table 2.

Table 2. Exposure parameters used in the health risk assessment.

Parameter	Symbol	Unit	Adult	Child	Source
Drinking water ingestion rate	IR	L/d	1.8	1	Chinese Exposure Factors Handbook [49]
Exposure frequency	EF	d/a	365	365	Annual exposure
Exposure duration	ED	a	30	6	Population characteristics
Body weight	BW	kg	60	20	Chinese Exposure Factors Handbook [49]
Averaging time	AT	d	$ED \times 365$		Non-carcinogenic effect
Oral reference dose	RfD *	$\text{mg NO}_3^-/(\text{kg} \cdot \text{d})$	7.09	7.09	USEPA IRIS [32]

Notes: * The RfD was expressed on a nitrate-ion basis. The value of $7.09 \text{ mg NO}_3^-/(\text{kg} \cdot \text{d})$ is equivalent to $1.6 \text{ mg NO}_3^-/\text{N}/(\text{kg} \cdot \text{d})$.

Sample NO_3^- concentrations were interpolated by ordinary kriging to produce a continuous regional concentration surface [34,50]. Given the right-skewed distribution of NO_3^- concentrations, experimental semivariograms were carefully examined before interpolation. Spherical, exponential and Gaussian variogram models were compared, and the final model was selected based on leave-one-out cross-validation. Model selection considered a mean error close to zero, a lower root mean square error and a standardized root mean square error close to one. Kriging standard deviation was retained as the interpolation uncertainty layer. Adult- and child-specific exposure parameters were then applied to estimate the spatial distribution of HQ for each population group.

2.5.2. Probability–HQ Matrix Zoning

Exceedance probability and child HQ were cross-classified using a probability–HQ risk matrix rather than combined into a single additive score. The two layers were treated as related assessment products rather than independent likelihood and consequence dimensions. Child HQ was classified using the non-carcinogenic risk criterion $\text{HQ} = 1$ according to the USEPA health risk assessment framework [30,32]. The exceedance probability map was classified using $P = 0.5$ as the primary warning threshold, and $P = 0.7$ was used for sensitivity analysis. This matrix-based classification defines an intersection-based priority footprint and a broader warning or verification footprint. Four management zones were defined to support differentiated groundwater risk screening and management [35,36,51]: general protection, exceedance-warning, concentration-verification and priority-intervention zones. To account for zoning uncertainty, pixel-wise probability standard deviation from 300 repeated final-model training-and-prediction runs, child-HQ kriging standard deviation and threshold-sensitive pixels around $P = 0.5$ and child $\text{HQ} = 1$ were mapped as supplementary uncertainty layers. The classification criteria for the four management zones are summarized in Table 3.

Table 3. Probability–HQ risk matrix used for management-oriented zoning.

Child HQ/Exceedance Probability	$P < 0.5$	$P \geq 0.5$
Child $\text{HQ} < 1$	General protection zone	Exceedance-warning zone
Child $\text{HQ} \geq 1$	Concentration-verification zone	Priority-intervention zone

The priority-intervention zone represents the conservative coincidence footprint, whereas the union of exceedance-warning, concentration-verification and priority-intervention zones represents the broader warning and verification footprint [51].

3. Results

3.1. Spatial Characteristics of Groundwater Nitrate Concentrations and Exceedances

Groundwater NO_3^- concentrations varied markedly among the 157 samples, ranging from 0.009 to 270.673 mg/L. The mean, median and standard deviation were 24.405, 5.626 and 39.541 mg/L, respectively. Most samples contained relatively low nitrate concentrations, while a small number of high values strongly influenced the distribution. This pattern indicates pronounced spatial heterogeneity and localised nitrate enrichment in shallow groundwater.

Using the $\text{NO}_3^- > 50$ mg/L threshold, 31 samples were classified as exceedances and 126 as non-exceedances. The corresponding exceedance rate was 19.7%. Thus, nitrate contamination was characterised by localised hotspots rather than widespread regional exceedance. Nevertheless, nearly one-fifth of samples exceeded the drinking-water limit, indicating a clear need for risk screening and water-safety reassessment. Exceedance and

high-concentration samples clustered mainly in central-western Handan and around parts of the urban core. Eastern and northeastern areas contained comparatively few exceedances.

3.2. Model Performance and Selection for Spatial Prediction

Across 300 repeated hold-out validations, RF, SVM and XGBoost showed distinct performance profiles (Table 4). SVM achieved the highest ROC-AUC at 0.903, compared with 0.893 for XGBoost and 0.884 for RF. This result indicated a modest advantage for SVM in overall discrimination.

Threshold-dependent metrics showed that XGBoost detected exceedance samples more effectively. At the $P = 0.5$ threshold, XGBoost increased recall from 0.840 for SVM to 0.917 and reduced the mean number of false negatives from 0.96 to 0.50. At the stricter $P = 0.7$ threshold, XGBoost increased recall from 0.676 to 0.809 and reduced the mean number of false negatives from 1.95 to 1.15. Because false negatives could leave potential nitrate-exceedance areas unidentified, model selection prioritised recall and missed-case control. XGBoost was therefore selected for spatial probability mapping despite the slightly higher ROC-AUC of SVM.

Five-fold spatial block cross-validation yielded a mean ROC-AUC of 0.867 for XGBoost. This value was lower than the result from random partitioning, indicating some performance inflation under random validation. Nevertheless, XGBoost retained sufficient spatial discrimination to resolve broad regional patterns of nitrate exceedance probability.

Table 4. Model performance at probability thresholds of 0.5 and 0.7.

Threshold	Model	AUC	Accuracy	Precision	Recall	F1-Score	Mean FN
0.5	SVM	0.903 ± 0.045	0.825 ± 0.057	0.535 ± 0.111	0.840 ± 0.151	0.646 ± 0.102	0.96
0.5	XGBoost	0.893 ± 0.048	0.804 ± 0.058	0.500 ± 0.089	0.917 ± 0.118	0.642 ± 0.084	0.5
0.5	Random Forest	0.884 ± 0.051	0.802 ± 0.059	0.496 ± 0.091	0.890 ± 0.139	0.631 ± 0.094	0.66
0.7	SVM	0.903 ± 0.045	0.846 ± 0.052	0.597 ± 0.140	0.676 ± 0.174	0.620 ± 0.122	1.95
0.7	XGBoost	0.893 ± 0.048	0.811 ± 0.057	0.511 ± 0.105	0.809 ± 0.157	0.618 ± 0.101	1.15
0.7	Random Forest	0.884 ± 0.051	0.819 ± 0.053	0.538 ± 0.176	0.542 ± 0.218	0.516 ± 0.155	2.75

3.3. Spatial Distribution of Groundwater Nitrate Exceedance Probability

The final exceedance-probability map was represented by the pixel-wise mean probability from 300 repeated XGBoost training-and-prediction runs. Values ranged from 0 to 1 and represented the probability that groundwater NO_3^- exceeded 50 mg/L in each spatial unit.

Predicted exceedance probability varied markedly across Handan, with higher values in the central-western region and lower values in the east (Figure 3). Principal high-probability zones occurred in Fengfeng Mining District, Fuxing District, northern Cixian County and central-eastern Wu'an. These areas were identified as potential exceedance hotspots under the observed environmental conditions. Eastern and northeastern Handan generally showed lower predicted probabilities.

Mean predicted probabilities also differed among administrative units. Relatively high values occurred in Fengfeng Mining District, Fuxing District, Cixian County, Hanshan District, Congtai District and parts of Wu'an. Eastern and northeastern counties generally had lower values. These patterns confirm that nitrate exceedance potential is spatially uneven and concentrated in localised clusters.

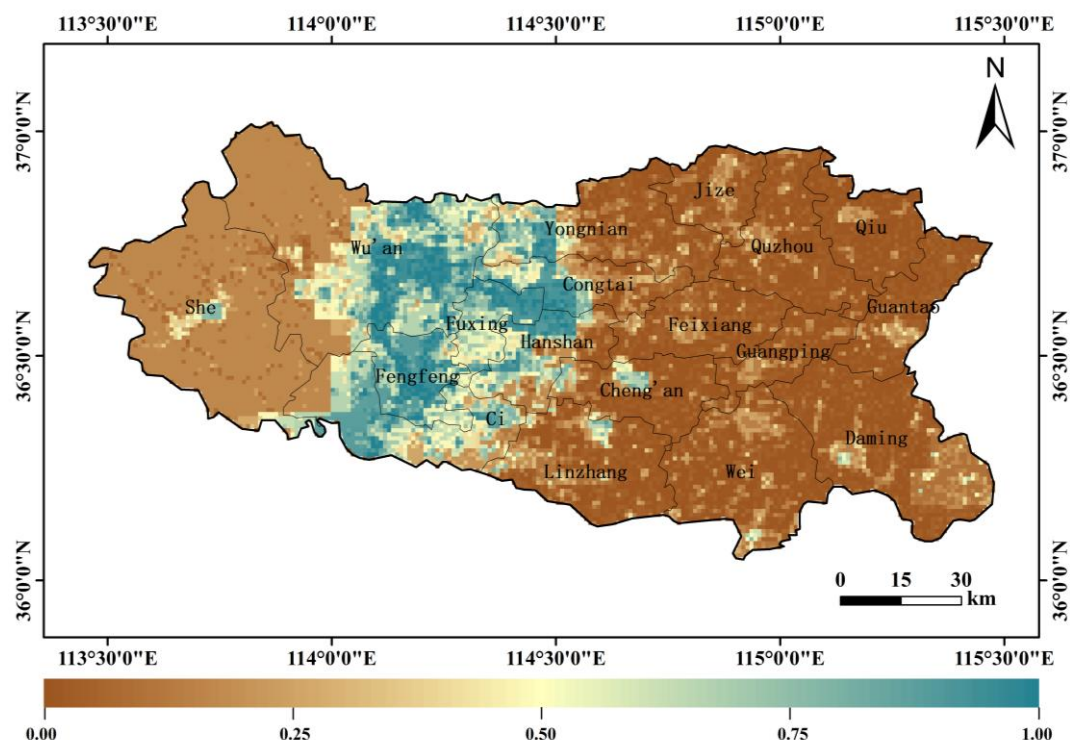


Figure 3. Predicted groundwater nitrate exceedance probability map for Handan City.

3.4. SHAP Responses of Key Environmental Predictors

SHAP analysis was used to interpret how XGBoost differentiated nitrate exceedance probability. Mean annual potential evapotranspiration, built-up land fraction, cropland fraction, slope and soil clay content were the five most influential predictors. Their mean absolute SHAP values were 1.4267, 0.5933, 0.3048, 0.3000 and 0.2607, respectively. These results indicate that model discrimination reflected the combined influence of climatic water balance, anthropogenic pressure, land-use structure and terrain-soil conditions (Figure 4).

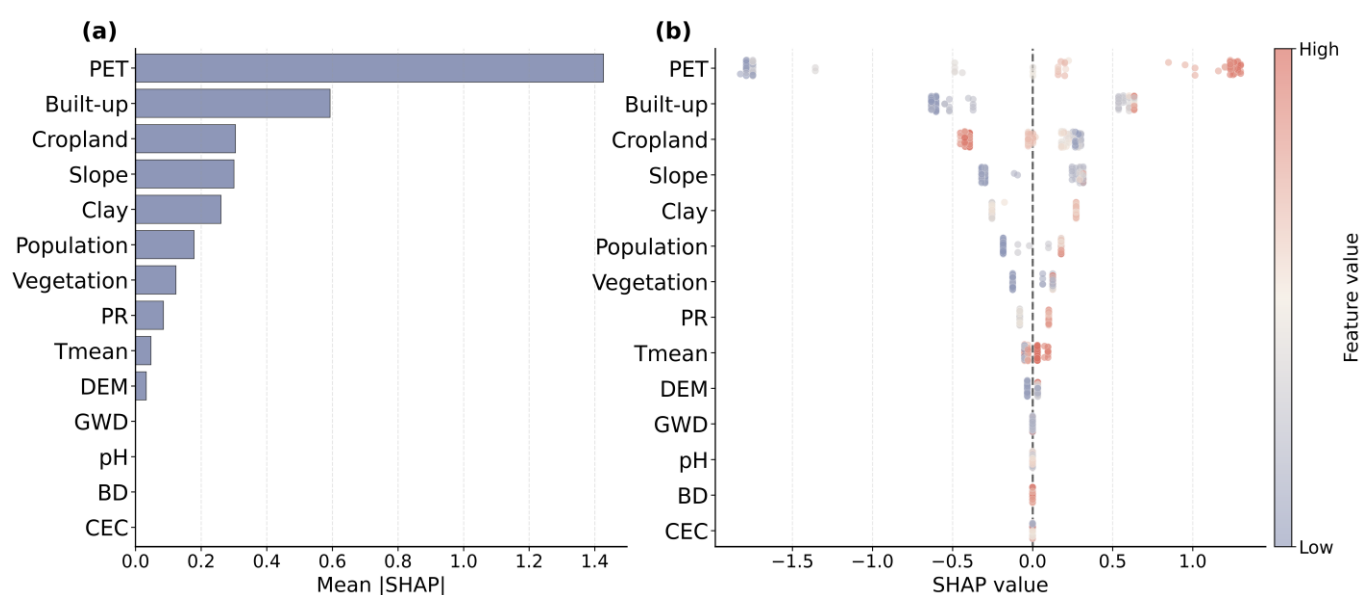


Figure 4. SHAP-based interpretation of the XGBoost model: (a) predictor importance; (b) SHAP summary plot.

Mean annual potential evapotranspiration had the largest mean absolute SHAP value in the final XGBoost model. However, this result should be interpreted as a model-response pattern rather than evidence that PET independently controls nitrate exceedance. PET is a smooth climate variable and is broadly aligned with the west–east topographic and hydroclimatic gradient of Handan. Therefore, its high SHAP importance may partly reflect the model’s use of PET as a proxy for broader spatial gradients.

Cropland fraction, slope and soil clay content showed nonlinear SHAP responses (Figure 5). Cropland fraction did not increase monotonically with SHAP values, indicating a context-dependent association with model output. Slope and soil clay content also showed segmented or nonlinear responses across their observed ranges. Overall, no single predictor governed the modelled probability. Instead, model discrimination reflected the combined effects of climate, human activity, land use and terrain-soil conditions.

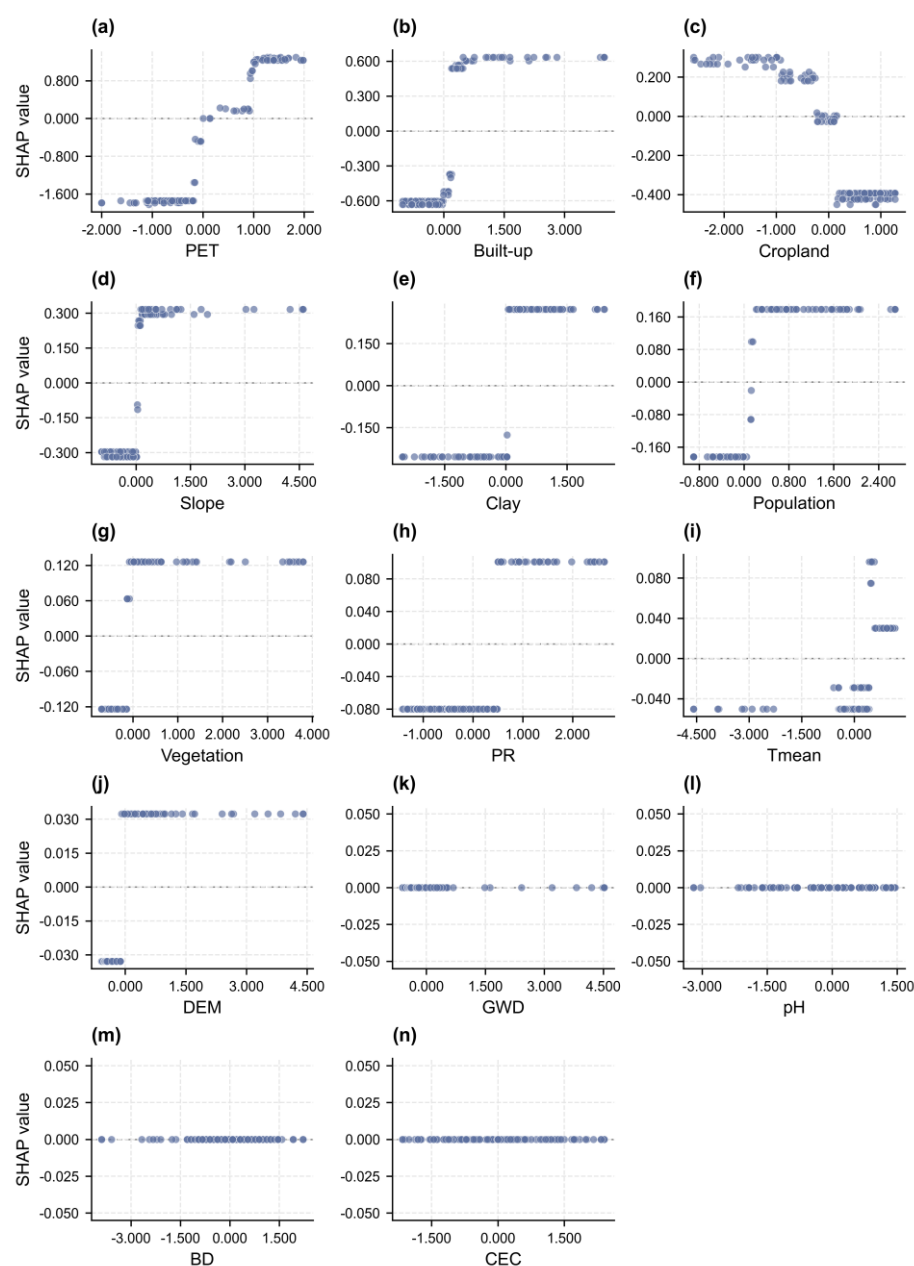


Figure 5. SHAP dependence plots for the 14 environmental predictors: (a) potential evapotranspiration (PET); (b) built-up land fraction; (c) cropland fraction; (d) slope; (e) soil clay content; (f) population density; (g) vegetation fraction; (h) mean annual precipitation (PR); (i) mean annual

temperature (Tmean); (j) elevation (DEM); (k) groundwater table depth (GWD); (l) soil pH; (m) bulk density (BD); and (n) cation-exchange capacity (CEC).

3.5. Health Risk Assessment and Probability–HQ Matrix Zoning

3.5.1. Health Risk Assessment for Adults and Children

Ordinary kriging was used to spatialise NO_3^- concentrations, after which population-specific exposure parameters were applied to calculate drinking-water HQs. Groundwater nitrate health risk was generally low across the study area. However, child HQs were consistently higher than adult HQs. The Gaussian variogram model provided the best cross-validation performance, with a nugget of 1170.89 (mg/L)², sill of 3185.98 (mg/L)² and range of 193.06 km. Leave-one-out cross-validation yielded a mean error of 0.11 mg/L, RMSE of 32.56 mg/L and standardized RMSE of 0.94. These results indicate little overall bias and acceptable interpolation performance for regional health-risk screening. The relatively high RMSE mainly reflected the influence of a small number of high-concentration samples in the right-skewed nitrate dataset.

Adult HQs ranged from 0.00004 to 1.14027, with a mean of 0.0948 and a standard deviation of 0.1147. Child HQs ranged from 0.00006 to 1.90044, with a mean of 0.1580 and a standard deviation of 0.1912. The maximum, mean and variability were all higher for children. Mean child HQ was approximately 1.67 times the adult value.

Areas with $\text{HQ} > 1$ accounted for approximately 0.03% of the study area for adults and 0.35% for children. Adult and child hotspots showed broadly similar spatial patterns, with higher values in western and central Handan and lower values in the east and south. Elevated-risk areas occurred mainly in central-eastern Wu'an and localised parts of the urban core (Figure 6). Child HQ was therefore used as the concentration-derived health-concern indicator in the probability–HQ risk matrix.

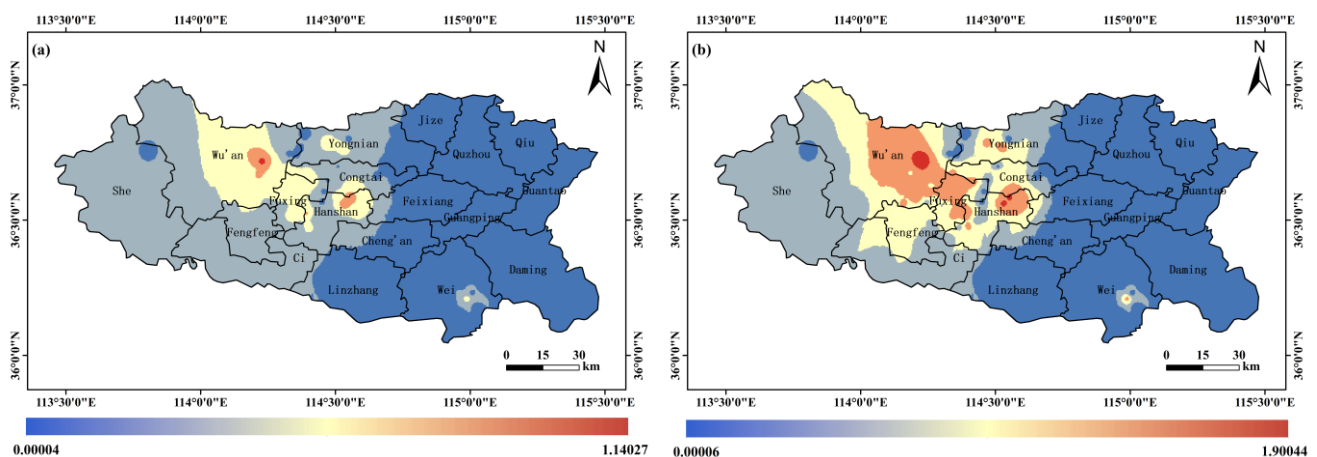


Figure 6. Spatial distribution of nitrate hazard quotients (HQs) for (a) adults and (b) children.

3.5.2. Management-Oriented Probability–HQ Matrix Zoning

The probability–HQ matrix reclassified the study area into four management zones. Under the primary $P = 0.5$ matrix, general protection and exceedance-warning zones covered 81.26% and 18.40% of the study area, respectively. The concentration-verification zone was negligible, covering 0.01%, and the priority-intervention zone covered 0.33% (Table 5; Figure 7).

The matrix zones contained most observed exceedance samples. Among the 31 samples exceeding the nitrate threshold, 26 were located in exceedance-warning zones and 3 in priority-intervention zones. Thus, 29 exceedance samples, corresponding to 93.55%, fell within the broader warning and verification footprint. Only two exceedance samples were

located in general protection zones. The broader warning and verification footprint covered 18.74% of the study area.

When the probability threshold was increased to $P = 0.7$, the exceedance-warning zone decreased to 8.79%, whereas the priority-intervention zone remained similar at 0.29%. Uncertainty analysis showed that threshold-sensitive transition areas accounted for 8.06% of the mapped study area, including 7.24% of pixels sensitive to $P = 0.5$ and 0.97% sensitive to child $HQ = 1$ (Figure A2; Table A1). Because the two masks partly overlapped, the combined transition area was smaller than their sum. These transition areas occurred mainly near the boundary between general protection and exceedance-warning zones, indicating that local zone boundaries remain uncertain although the regional zoning pattern is stable.

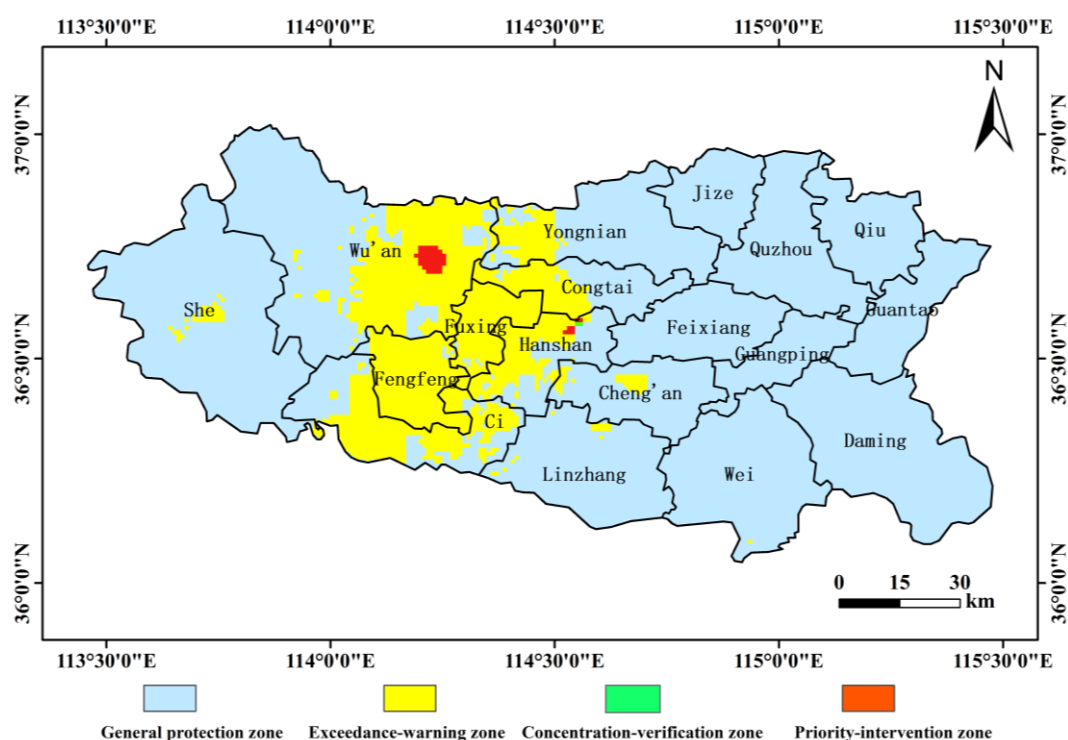


Figure 7. Probability–HQ matrix zoning of groundwater nitrate risk.

Table 5. Probability–HQ matrix zoning results and observed exceedance samples.

Zone	Criteria	Area Proportion	Observed Exceedance Samples
General protection zone	$P < 0.5$, child $HQ < 1$	81.26%	2/31
Exceedance-warning zone	$P \geq 0.5$, child $HQ < 1$	18.40%	26/31
Concentration-verification zone	$P < 0.5$, child $HQ \geq 1$	0.01%	0/31
Priority-intervention zone	$P \geq 0.5$, child $HQ \geq 1$	0.33%	3/31

4. Discussion

4.1. Machine-Learning Prediction and Key Environmental Responses

Recent machine-learning studies have predicted groundwater nitrate concentration or elevated nitrate risk using land-use, soil, climatic, topographic and hydrogeological variables [15,16,19–23]. Consistent with these studies, our results show that nitrate exceedance risk is spatially heterogeneous and reflects the combined influence of natural conditions and human activities. The difference is that this study focused on exceedance probability and early-warning screening rather than concentration prediction alone; therefore,

model selection emphasised recall and missed-exceedance control in addition to overall discrimination.

The three machine-learning models captured nonlinear associations between nitrate exceedance and multi-source environmental predictors, but their performance profiles differed. Although SVM achieved the highest ROC-AUC, XGBoost showed stronger recall and fewer false negatives at both probability thresholds. For early-warning purposes, identifying as many potential exceedances as possible is more important than maximising overall classification accuracy. XGBoost was therefore selected because it better matched the practical objective of reducing missed exceedance areas.

The reduction in performance under spatial cross-validation indicates that sample location influenced model extrapolation. Random partitioning can overstate generalisation when nearby samples occur in both training and test sets [24,25]. Spatial cross-validation provides a more conservative estimate of performance in unsampled areas and is therefore important for regional groundwater risk mapping.

High-probability zones occurred mainly in central-western Handan and along the transition between the piedmont plain and urbanised areas. This distribution broadly corresponds to the regional gradient from western mountains and hills to central-eastern alluvial-proluvial plains. Variations in terrain, aquifer properties, recharge–runoff–discharge processes and human activity may jointly affect nitrate input, leaching, transport and local accumulation. These combined gradients may explain why the model identified central-western Handan as a potential hotspot.

SHAP analysis identified potential evapotranspiration, built-up land fraction, cropland fraction, slope and soil clay content as the leading predictors of model response. These variables are broadly consistent with previous nitrate studies showing that land use, anthropogenic nitrogen input, hydroclimatic conditions, soil properties and groundwater-related transport processes jointly affect nitrate accumulation [7,10–12,19–23]. The high SHAP importance of PET should be interpreted cautiously because PET covaries with the regional west–east elevation and hydroclimatic gradient. It may represent a broader climatic–topographic context rather than an independently verified evapotranspiration mechanism. Built-up land fraction reflects urbanisation and possible domestic nitrogen inputs. The non-monotonic response of cropland fraction suggests that agricultural effects also depend on irrigation, fertilizer practices, soil texture and groundwater depth. Slope and clay content may further influence runoff, infiltration and solute transport. These interpretations describe plausible environmental contexts for the model response, but they do not establish causality or identify nitrate sources, which would require isotope evidence, hydrochemical indicators or source inventories [27–29].

4.2. Interpreting Probability–HQ Matrix Zoning

The probability–HQ matrix translates two related assessment products into management zones. Child HQ reflects concentration-derived exposure concern, whereas exceedance probability captures environmental conditions associated with nitrate exceedance. In Handan, almost all pixels with child HQ ≥ 1 were located within areas where $P \geq 0.5$, while the concentration-verification zone was negligible. In this case, the off-diagonal disagreement was one-sided: high-probability but low-HQ areas were common, whereas high-HQ but low-probability areas were absent. The matrix therefore separated a small priority-intervention footprint from a broader exceedance-warning zone. The priority-intervention zone represents a narrow conservative target where child-health concern and exceedance susceptibility overlap. This zone covered only 0.33% of the study area and remained similar at 0.29% when the stricter $P \geq 0.7$ threshold was used, indicating that the highest-priority footprint was relatively stable.

For management, the most informative class was the exceedance-warning zone. Although child HQ remained below 1 in this zone, it contained 26 of the 31 observed nitrate exceedance samples. These areas are therefore more suitable for intensified monitoring, source-load prevention and repeated water-quality checks than for immediate intervention. General protection zones can remain under routine monitoring, but the two observed exceedance samples in this class indicate that local verification is still needed where site-specific conditions differ from regional model patterns. Therefore, the zoning boundaries should be interpreted as screening and verification boundaries rather than deterministic contamination boundaries, especially in threshold-sensitive transition areas.

4.3. Uncertainty and Study Limitations

Several uncertainties constrain how the predicted zones should be used. The sample set was small and imbalanced, with nitrate exceedances accounting for 19.7% of observations. Although SMOTE was restricted to the training folds to reduce leakage, minority-class patterns may still be under-represented. The probability surface should therefore be interpreted as a relative indicator of regional exceedance susceptibility rather than as a site-specific estimate for operational decisions.

Predictors also differed in temporal coverage, spatial resolution and measurement quality. For example, groundwater depth was represented by a multi-year mean, capturing the broad hydrogeological setting rather than conditions at the sampling date. These scale mismatches make the results more appropriate for regional screening and trend interpretation than for parcel-scale boundary delineation. Higher-frequency groundwater, nitrogen-loading and land-use data would improve sensitivity to short-term variability.

Interpretation of the probability–HQ matrix depends on both model and interpolation uncertainty. SHAP analysis described associations within the selected predictor set, but these associations should not be treated as causal evidence or source attribution. Similarly, HQ estimates depended on kriged NO_3^- concentrations. Despite low overall cross-validation bias, the right-skewed distribution and fitted variogram may smooth local concentration peaks. Repeated XGBoost runs, kriging standard deviation and threshold-sensitive masks all indicate that local boundaries remain uncertain, especially near $P = 0.5$ and child HQ = 1. The matrix should therefore guide monitoring, drinking-water verification and management prioritisation rather than define fixed contamination or intervention boundaries. Source attribution and targeted management would require repeated monitoring, independent validation, isotopic and hydrochemical evidence, source inventories and more detailed hydrogeological information.

5. Conclusions

Among 157 shallow groundwater samples collected in 2023, 31 exceeded the NO_3^- threshold, giving an exceedance rate of 19.7%. Exceedances were spatially clustered, indicating that nitrate risk in Handan is dominated by localised hotspots. Although SVM achieved the highest ROC-AUC, XGBoost showed stronger recall and fewer false negatives, making it more suitable for early-warning mapping. High-probability areas were mainly located in central-western and urban-fringe areas, including Fengfeng Mining District, Fuxing District, northern Cixian County and central-eastern Wu'an.

SHAP analysis indicated that the XGBoost model mainly used potential evapotranspiration, built-up land fraction, cropland fraction, slope and soil clay content to discriminate exceedance probability. These variables reflect combined hydroclimatic, land-use, terrain and soil conditions, but should be interpreted as model-response patterns rather than direct evidence of nitrate sources. Non-carcinogenic health risk was generally low, but children had higher HQs than adults and should remain the priority receptor group in nitrate-related drinking-water risk assessment.

Under the primary $P = 0.5$ threshold, the probability–HQ matrix classified 81.26%, 18.40%, 0.01% and 0.33% of the study area as general protection, exceedance-warning, concentration-verification and priority-intervention zones, respectively. The broader warning and verification footprint covered 18.74% of the study area and contained 29 of the 31 observed exceedance samples. These results support the use of the matrix as a transparent screening and verification tool rather than a deterministic contamination boundary or an independent composite risk index. Future work should incorporate multi-year monitoring, denser sampling in transition zones and applications in other hydrogeological settings to validate the framework. The integration of isotope tracing, source inventories, groundwater-flow information and seasonal hydrochemical data would further help distinguish statistical risk patterns from nitrate sources and transport processes.

Author Contributions: Conceptualization, J.W. and Y.Z.; methodology, J.W. and Y.Z.; software, J.W.; validation, X.K., Q.L., J.L. and T.Z.; formal analysis, J.W.; investigation, X.K. and J.L.; resources, X.K., Q.L., Y.Z. and T.Z.; data curation, X.K. and J.W.; writing—original draft preparation, X.K. and J.W.; writing—review and editing, X.K., Q.L., J.L., Y.Z. and T.Z.; visualization, J.W.; supervision, Y.Z. and Q.L.; project administration, X.K. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number U2444219.

Data Availability Statement: The data are available from the authors upon reasonable request due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RF	Random Forest
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
ROC-AUC	Receiver Operating Characteristic–Area Under the Curve
HQ	Hazard Quotient

Appendix A

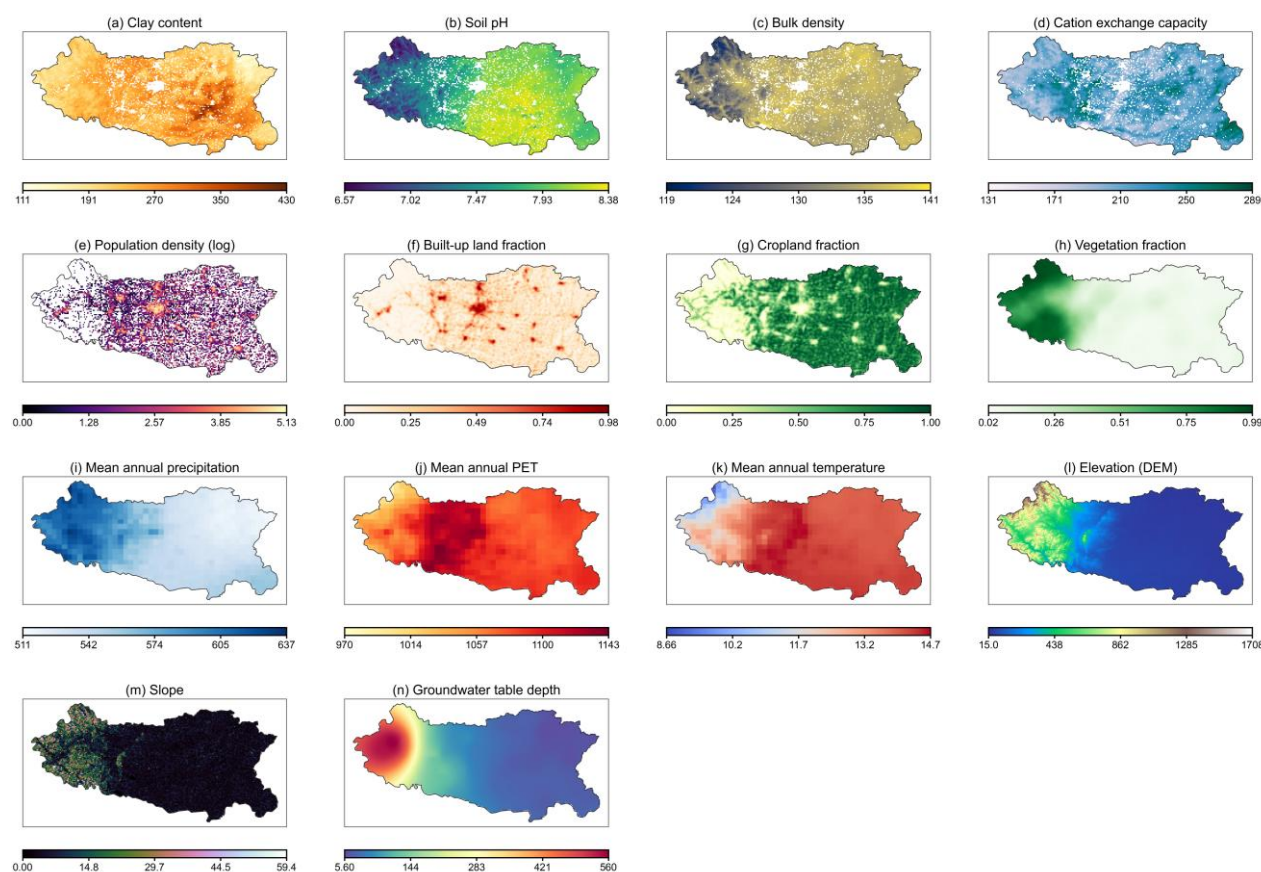


Figure A1. Spatial distribution of processed environmental predictor rasters used for groundwater nitrate exceedance-probability modelling: (a) clay content; (b) soil pH; (c) bulk density; (d) cation exchange capacity; (e) population density; (f) built-up land fraction; (g) cropland fraction; (h) vegetation fraction; (i) mean annual precipitation; (j) mean annual potential evapotranspiration; (k) mean annual temperature; (l) elevation; (m) slope; and (n) groundwater table depth.

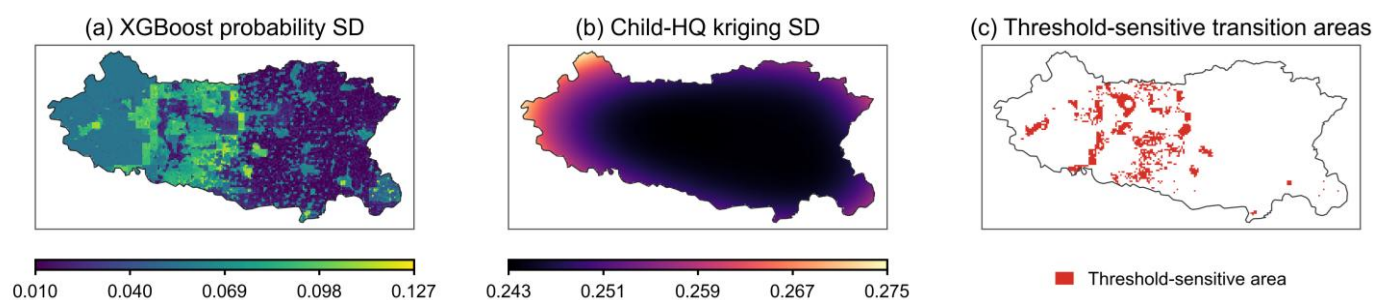


Figure A2. Spatial uncertainty associated with probability-HQ management zoning: (a) uncertainty of predicted exceedance probability; (b) uncertainty of child HQ derived from kriging; and (c) threshold-sensitive transition areas around $P = 0.5$ or child HQ = 1.

Table A1. Area statistics of threshold-sensitive transition areas.

Uncertainty or Transition Layer	Criterion	Area Proportion
$P = 0.5$ threshold-sensitive area	$ P_{\text{mean}} - 0.5 \leq P_{\text{sd}}$	7.24%
Child HQ = 1 threshold-sensitive area	$ HQ_{\text{child}} - 1 \leq HQ_{\text{child, sd}}$	0.97%
Combined threshold-sensitive transition area	P -sensitive or HQ-sensitive	8.06%

References

1. Xu, X.; Zou, Y.; Pan, H.; Zhang, R.; Gu, B. Safeguarding Groundwater Nitrate within Regional Boundaries in China. *Environ. Sci. Technol.* **2025**, *59*, 467–477. <https://doi.org/10.1021/acs.est.4c08197>.
2. Zhou, Q.; Zhang, J.; Xing, K.; Wei, J.; Yao, Y. Groundwater Nitrate Contamination in China: Spatial Distribution, Temporal Trend, and Driver Analysis. *Environ. Res.* **2024**, *262*, 119932. <https://doi.org/10.1016/j.envres.2024.119932>.
3. Richa, A.; Touil, S.; Fizir, M. Recent Advances in the Source Identification and Remediation Techniques of Nitrate-Contaminated Groundwater: A Review. *J. Environ. Manag.* **2022**, *316*, 115265. <https://doi.org/10.1016/j.jenvman.2022.115265>.
4. Cui, R.; Zhang, D.; Hu, W.; Zhao, X.; Yan, H.; Liu, G.; Chen, A. Nitrogen in Soil, Manure and Sewage Has Become a Major Challenge in Controlling Nitrate Pollution in Groundwater around Plateau Lakes, Southwest China. *J. Hydrol.* **2023**, *620*, 129541. <https://doi.org/10.1016/j.jhydrol.2023.129541>.
5. Wakida, F.T.; Lerner, D.N. Non-Agricultural Sources of Groundwater Nitrate: A Review and Case Study. *Water Res.* **2005**, *39*, 3–16. <https://doi.org/10.1016/j.watres.2004.07.026>.
6. Spalding, R.F.; Exner, M.E. Occurrence of Nitrate in Groundwater: A Review. *J. Environ. Qual.* **1993**, *22*, 392–402. <https://doi.org/10.2134/jeq1993.00472425002200030002x>.
7. Mao, H.; Wang, G.; Liao, F.; Shi, Z.; Rao, Z.; Zhang, H.; Qiao, Z.; Bai, Y.; Chen, X.; Yan, X.; et al. Spatiotemporal Variation of Groundwater Nitrate Concentration Controlled by Groundwater Flow in a Large Basin: Evidence from Multi-Isotopes. *Water Resour. Res.* **2024**, *60*, e2023WR035299. <https://doi.org/10.1029/2023WR035299>.
8. Rivett, M.O.; Buss, S.R.; Morgan, P.; Smith, J.W.N.; Bemment, C.D. Nitrate Attenuation in Groundwater: A Review of Biogeochemical Controlling Processes. *Water Res.* **2008**, *42*, 4215–4232. <https://doi.org/10.1016/j.watres.2008.07.020>.
9. World Health Organization. *Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First and Second Addenda*; World Health Organization: Geneva, Switzerland, 2022. Available online: <https://www.who.int/publications/i/item/9789240045064> (accessed on 12 June 2026).
10. Wu, R.; Guo, Z.; Zhan, Y.; Cao, G.; Chen, K.; Chang, Z.; Li, H.; He, X.; Zheng, C. Impacts of Land Surface Nitrogen Input on Groundwater Quality in the North China Plain. *ACS ES&T Water* **2024**, *4*, 2369–2381. <https://doi.org/10.1021/acsestwater.3c00712>.
11. Zhang, Y.; Chen, Z.; Huang, G.; Yang, M. Origins of Groundwater Nitrate in a Typical Alluvial-Pluvial Plain of North China Plain: New Insights from Groundwater Age-Dating and Isotopic Fingerprinting. *Environ. Pollut.* **2023**, *316*, 120592. <https://doi.org/10.1016/j.envpol.2022.120592>.
12. Raji-Hoffman, I.; Dahan, O.; Dahlke, H.E.; Harter, T.; Kisekka, I. Assessing Nitrate Leaching during Drought and Extreme Precipitation. *Water Resour. Res.* **2024**, *60*, e2024WR037973. <https://doi.org/10.1029/2024WR037973>.
13. Ju, X.T.; Kou, C.L.; Zhang, F.S.; Christie, P. Nitrogen Balance and Groundwater Nitrate Contamination: Comparison among Three Intensive Cropping Systems on the North China Plain. *Environ. Pollut.* **2006**, *143*, 117–125. <https://doi.org/10.1016/j.envpol.2005.11.005>.
14. He, S.; Li, P.; Su, F.; Wang, D.; Ren, X. Identification and Apportionment of Shallow Groundwater Nitrate Pollution in Weinig Plain, Northwest China, Using Hydrochemical Indices, Nitrate Stable Isotopes, and the New Bayesian Stable Isotope Mixing Model (MixSIAR). *Environ. Pollut.* **2022**, *298*, 118852. <https://doi.org/10.1016/j.envpol.2022.118852>.
15. Zhan, Y.; Guo, Z.; Ruzzante, S.; Gleeson, T.; Andrews, C.B.; Babovic, V.; Zheng, C. Assessment of Spatiotemporal Risks for Nationwide Groundwater Nitrate Contamination. *Sci. Total Environ.* **2024**, *947*, 174508. <https://doi.org/10.1016/j.scitotenv.2024.174508>.
16. Liang, Y.; Zhang, X.; Gan, L.; Chen, S.; Zhao, S.; Ding, J.; Kang, W.; Yang, H. Mapping Specific Groundwater Nitrate Concentrations from Spatial Data Using Machine Learning: A Case Study of Chongqing, China. *Heliyon* **2024**, *10*, e27867. <https://doi.org/10.1016/j.heliyon.2024.e27867>.
17. Bera, A.; Mukhopadhyay, B.P.; Chowdhury, P.; Ghosh, A.; Biswas, S. Groundwater Vulnerability Assessment Using GIS-Based DRASTIC Model in Nangasai River Basin, India with Special Emphasis on Agricultural Contamination. *Ecotoxicol. Environ. Saf.* **2021**, *214*, 112085. <https://doi.org/10.1016/j.ecoenv.2021.112085>.
18. Zhao, X.; Wang, D.; Xu, H.; Ding, Z.; Shi, Y.; Lu, Z.; Cheng, Z. Groundwater Pollution Risk Assessment Based on Groundwater Vulnerability and Pollution Load on an Isolated Island. *Chemosphere* **2022**, *289*, 133134. <https://doi.org/10.1016/j.chemosphere.2021.133134>.
19. He, S.; Wu, J.; Wang, D.; He, X. Predictive Modeling of Groundwater Nitrate Pollution and Evaluating Its Main Impact Factors Using Random Forest. *Chemosphere* **2022**, *290*, 133388. <https://doi.org/10.1016/j.chemosphere.2021.133388>.

20. Ransom, K.M.; Nolan, B.T.; Stackelberg, P.E.; Belitz, K.; Fram, M.S. Machine Learning Predictions of Nitrate in Groundwater Used for Drinking Supply in the Conterminous United States. *Sci. Total Environ.* **2022**, *807*, 151065. <https://doi.org/10.1016/j.scitotenv.2021.151065>.
21. Sarkar, S.; Mukherjee, A.; Gupta, S.D.; Bhanja, S.N.; Bhattacharya, A. Predicting Regional-Scale Elevated Groundwater Nitrate Contamination Risk Using Machine Learning on Natural and Human-Induced Factors. *ACS ES&T Eng.* **2022**, *2*, 689–702. <https://doi.org/10.1021/acsestengg.1c00360>.
22. Mahboobi, H.; Shakiba, A.; Mirbagheri, B. Improving Groundwater Nitrate Concentration Prediction Using Local Ensemble of Machine Learning Models. *J. Environ. Manag.* **2023**, *345*, 118782. <https://doi.org/10.1016/j.jenvman.2023.118782>.
23. Sajedi-Hosseini, F.; Malekian, A.; Choubin, B.; Rahmati, O.; Cipullo, S.; Coulon, F.; Pradhan, B. A Novel Machine Learning-Based Approach for the Risk Assessment of Nitrate Groundwater Contamination. *Sci. Total Environ.* **2018**, *644*, 954–962. <https://doi.org/10.1016/j.scitotenv.2018.07.054>.
24. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Aroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* **2017**, *40*, 913–929. <https://doi.org/10.1111/ecog.02881>.
25. Meyer, H.; Pebesma, E. Machine Learning-Based Global Maps of Ecological Variables and the Challenge of Assessing Them. *Nat. Commun.* **2022**, *13*, 2208. <https://doi.org/10.1038/s41467-022-29838-9>.
26. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774. Available online: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (accessed on 12 June 2026).
27. Kendall, C.; Aravena, R. Nitrate Isotopes in Groundwater Systems. In *Environmental Tracers in Subsurface Hydrology*; Cook, P.G.; Herczeg, A.L., Eds.; Springer: Boston, MA, USA, 2000; pp. 261–297. https://doi.org/10.1007/978-1-4615-4557-6_9.
28. Qi, S.; Feng, Q.; Zhu, M.; Shu, H.; Liu, W.; Yang, L.; Yin, Z.; Zhang, C. Source Apportionment of Nitrates in Different Aquifers in an Arid Region, Northwestern China. *J. Clean. Prod.* **2022**, *374*, 133969. <https://doi.org/10.1016/j.jclepro.2022.133969>.
29. Liu, J.; Qiao, S.; Zhao, S.; Chen, H.; Wu, Y.; Li, D.; Liu, P.; Li, L. Quantifying the Sources and Health Risks of Groundwater Nitrate via Dual N-O Isotopes and Monte Carlo Simulations in a Developed Planting-Breeding Area. *Ecotoxicol. Environ. Saf.* **2025**, *290*, 117778. <https://doi.org/10.1016/j.ecoenv.2025.117778>.
30. U.S. Environmental Protection Agency. *Risk Assessment Guidance for Superfund, Volume I: Human Health Evaluation Manual, Part A*; EPA/540/1-89/002; Office of Emergency and Remedial Response: Washington, DC, USA, 1989. Available online: https://www.epa.gov/sites/default/files/2015-09/documents/rags_a.pdf (accessed on 12 June 2026).
31. U.S. Environmental Protection Agency. *Exposure Factors Handbook: 2011 Edition*; EPA/600/R-09/052F; U.S. Environmental Protection Agency: Washington, DC, USA, 2011. Available online: <https://www.epa.gov/expobox/exposure-factors-handbook-2011-edition> (accessed on 12 June 2026).
32. U.S. Environmental Protection Agency. *Integrated Risk Information System (IRIS): Chemical Assessment Summary: Nitrate*; CASRN 14797-55-8; U.S. Environmental Protection Agency: Washington, DC, USA, 1991. Available online: https://iris.epa.gov/static/pdfs/0076_summary.pdf (accessed on 12 June 2026).
33. Lin, L.; St Clair, S.; Gamble, G.D.; Crowther, C.A.; Dixon, L.; Bloomfield, F.H.; Harding, J.E. Nitrate Contamination in Drinking Water and Adverse Reproductive and Birth Outcomes: A Systematic Review and Meta-Analysis. *Sci. Rep.* **2023**, *13*, 563. <https://doi.org/10.1038/s41598-022-27345-x>.
34. Wang, H.; Lu, K.; Shen, C.; Song, X.; Hu, B.; Liu, G. Human Health Risk Assessment of Groundwater Nitrate at a Two Geomorphic Units Transition Zone in Northern China. *J. Environ. Sci.* **2021**, *110*, 38–47. <https://doi.org/10.1016/j.jes.2021.03.013>.
35. Qiao, X.; Cheng, T.; Zhang, L.; Sun, N.; Ding, Z.; Shi, Z.; Wang, G.; Zhang, Z. Zoning Method for Groundwater Pollution Risk Control in Typical Industrial–Urban Integration Areas in the Middle Reaches of the Yangtze River. *Water* **2025**, *17*, 2249. <https://doi.org/10.3390/w17152249>.
36. He, X.; Yan, X.; Tang, H.; Liu, M.; Ma, L. An Integrated Framework for Urban Groundwater Risk Management: Linking Probabilistic Risk, Source Apportionment, and Hotspot Analysis in Zhengzhou, China. *Environ. Technol. Innov.* **2025**, *40*, 104665. <https://doi.org/10.1016/j.eti.2025.104665>.
37. Hengl, T.; Mendes de Jesus, J.; Heuvelink, G.B.M.; Ruiperez Gonzalez, M.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B.; et al. SoilGrids250m: Global Gridded Soil Information Based on Machine Learning. *PLoS ONE* **2017**, *12*, e0169748. <https://doi.org/10.1371/journal.pone.0169748>.

38. Bondarenko, M.; Priyatikanto, R.; Tejedor Garavito, N.; Zhang, W.; McKeen, T.; Cunningham, A.; Woods, T.; Hilton, J.; Cihan, D.; Nosatiuk, B.; et al. *Constrained Estimates of 2015–2030 Total Number of People per Grid Square at a Resolution of 3 Arc (Approximately 100 m at the Equator), R2025A Version v1*; WorldPop, School of Geography and Environmental Science, University of Southampton: Southampton, UK, 2025. <https://doi.org/10.5258/SOTON/WP00839>.
39. Zanaga, D.; Van De Kerchove, R.; Daems, D.; De Keersmaecker, W.; Brockmann, C.; Kirches, G.; Wevers, J.; Cartus, O.; Santoro, M.; Fritz, S.; et al. *ESA WorldCover 10 m 2021 v200*; Zenodo: Geneva, Switzerland, 2022. <https://doi.org/10.5281/zenodo.7254221>.
40. Abatzoglou, J.T.; Dobrowski, S.Z.; Parks, S.A.; Hegewisch, K.C. TerraClimate, a High-Resolution Global Dataset of Monthly Climate and Climatic Water Balance from 1958–2015. *Sci. Data* **2018**, *5*, 170191. <https://doi.org/10.1038/sdata.2017.191>.
41. Abrams, M.; Yamaguchi, Y.; Crippen, R. ASTER Global DEM (GDEM) Version 3. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2022**, *XLIII-B4-2022*, 593–598. <https://doi.org/10.5194/isprs-archives-XLIII-B4-2022-593-2022>.
42. Wang, M.; Yao, J.; Chang, H.; Liu, R.; Cao, Y.; Zhao, Y. *Monthly Groundwater Level Grid Dataset of China Region (2005–2022)*; National Tibetan Plateau/Third Pole Environment Data Center: Beijing, China, 2024. <https://doi.org/10.11888/Terre.tpdc.301342>.
43. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. <https://doi.org/10.1023/A:1010933404324>.
44. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. <https://doi.org/10.1007/BF00994018>.
45. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
46. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305. Available online: <https://www.jmlr.org/papers/v13/bergstra12a.html> (accessed on 12 June 2026).
47. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. <https://doi.org/10.1613/jair.953>.
48. Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>.
49. Ministry of Environmental Protection of the People’s Republic of China. *Exposure Factors Handbook of Chinese Population: Adults*; China Environmental Science Press: Beijing, China, 2013.
50. Qiu, H.; Gui, H.; Xu, H.; Cui, L.; Li, Z.; Yu, H. Quantifying Nitrate Pollution Sources of Shallow Groundwater and Related Health Risks Based on Deterministic and Monte Carlo Models: A Study in Huaibei Mining Area, Huaibei Coalfield, China. *Ecotoxicol. Environ. Saf.* **2023**, *249*, 114434. <https://doi.org/10.1016/j.ecoenv.2022.114434>.
51. Ministry of Ecology and Environment of the People’s Republic of China; Ministry of Water Resources of the People’s Republic of China; Ministry of Natural Resources of the People’s Republic of China. *Technical Guideline for Delineating Key Areas for Groundwater Pollution Prevention and Control (Trial)*; Ministry of Ecology and Environment of the People’s Republic of China: Beijing, China, 2023. Available online: https://www.mee.gov.cn/xxgk/xxgk/xxgk06/202309/t20230913_1040809.html (accessed on 12 June 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.