

Article

Diagnosing Weak Spatial Autocorrelation to Guide Groundwater Ammonium Risk Mapping at a Chemical Industrial Park

Bin Lu ^{1,2,3,†}, Qihuang Wang ^{4,†} , Ruiyun Li ³, Yaoling He ⁵ , Hua Li ^{3,6,*} and Yijun Yao ^{4,7,*} 

¹ Institute of Resources and Environment Engineering, Shanxi University, Taiyuan 030006, China; lubin3@sxu.edu.cn

² Taiyuan Ecology and Environment Monitoring and Scientific Research Center, Taiyuan 030002, China

³ School of Environmental & Resource Sciences, Shanxi University, Taiyuan 030006, China; liruiyun@sxu.edu.cn

⁴ Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China; qihuangwang@issas.ac.cn

⁵ School of Environmental and Municipal Engineering, Lanzhou Jiaotong University, No. 88 West Anning Rd., Lanzhou 730070, China; heyaoling420@outlook.com

⁶ Shanxi Laboratory for Yellow River, Taiyuan 030006, China

⁷ University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: lihua@sxu.edu.cn (H.L.); yijun_yao@issas.ac.cn (Y.Y.)

† These authors contributed equally to this work.

Abstract

Ammonium nitrogen ($\text{NH}_4^+\text{-N}$) contamination in groundwater beneath chemical industrial parks exhibits extreme spatial heterogeneity, yet the comparative effectiveness of spatial prediction methods under such conditions remains poorly understood. At a chemical industrial park in a region of Shanxi Province, northern China, we analyzed 133 monitoring wells sampled across four campaigns (April–October 2024) at two aquifer depths. Global Moran's I (0.040–0.118) and variogram nugget ratios (>75%) indicated weak spatial autocorrelation. Consequently, on the raw concentration scale, all six geostatistical methods yielded near-zero or negative leave-one-out cross-validation (LOO-CV) R^2 . Evaluated on the $\log_{10}$ scale, machine learning (ML) models achieved positive predictive skills, with Extreme Gradient Boosting (XGBoost) performing best ($R^2 \approx 0.75$). Three hybrid ML–kriging methods produced physically coherent plume surfaces while retaining their predictive skills; the April upper-layer result ($R^2 \approx 0.67$)—the only campaign without retained within-well information—best represents spatial generalization, whereas the higher later-campaign values ($R^2 > 0.97$) are optimistic. Exceedance probability mapping based on XGBoost (area under the ROC curve, $\text{AUC} = 0.959$) revealed a persistent high-risk zone. Because the geostatistical and ML metrics span different response scales and validation schemes, their comparison is indicative rather than a direct ranking. Spatial autocorrelation diagnostics should precede method selection at point-source-dominated industrial sites.

Keywords: groundwater contamination; ammonium nitrogen; exceedance probability mapping; spatial autocorrelation; machine learning; environmental risk assessment



Academic Editor: Fernando
António Leal Pacheco

Received: 10 June 2026

Revised: 14 July 2026

Accepted: 15 July 2026

Published: 21 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 quality degradation by ammonium nitrogen ($\text{NH}_4^+\text{-N}$) is a widespread concern in regions hosting chemical industrial parks, where concentrated manufacturing activities create multiple contamination pathways through leaking storage facilities, waste ponds, and subsurface pipelines [1,2]. In China, thousands of legacy chemical industrial

parks pose groundwater contamination risks, with $\text{NH}_4^+\text{-N}$ frequently detected as a principal pollutant due to its high mobility and persistence under anaerobic conditions [3]. National surveys have revealed that $\text{NH}_4^+\text{-N}$ is among the most frequently exceeded parameters in industrial zones, with contamination levels often spanning several orders of magnitude within a single site [4], highlighting the need for spatial characterization approaches that can accommodate such extreme heterogeneity. Accurate spatial mapping is essential for plume delineation, risk assessment, and remediation planning, yet discrete well-monitoring data inevitably leave substantial gaps that must be filled by prediction methods [5,6].

Geostatistical approaches—particularly kriging variants—have long served as standard tools for environmental spatial interpolation [7,8]. By modeling spatial covariance through variograms, kriging produces unbiased predictions with accompanying uncertainty estimates [9]. Multiple variants exist: Ordinary Kriging (OK) assumes an unknown but constant mean; Simple Kriging (SK) requires a known mean; Co-Kriging (CK) incorporates secondary variables; Indicator Kriging (IK) transforms data into binary indicators for threshold-based mapping; and Empirical Bayesian Kriging (EBK) accounts for variogram uncertainty through Bayesian updating [10,11]. Inverse Distance Weighting (IDW), a deterministic method that does not model spatial autocorrelation explicitly, is frequently included as a benchmark [7]. Despite this diversity, comparative evaluations applying multiple kriging variants simultaneously to a single contamination scenario remain surprisingly scarce, with most studies testing only one or two variants [12].

A critical but often overlooked prerequisite for geostatistical methods is sufficient spatial autocorrelation. When nearby observations are not meaningfully more similar than distant ones, variogram structure becomes unreliable and kriging predictions converge to the global mean [13]. This is particularly relevant in industrial settings, where point-source pollution creates extreme concentration contrasts fundamentally different from the gradual gradients of agricultural non-point-source pollution [14]. Cambardella et al. [15] classified spatial dependence strength using the nugget ratio $C_0/(C_0 + C)$, where values exceeding 75% indicate weak dependence dominated by random variation. Despite the diagnostic importance of this criterion, few studies have systematically evaluated geostatistical performance in settings where weak spatial autocorrelation has been explicitly diagnosed a priori.

Machine learning (ML) methods, such as Random Forest (RF) [16], Gradient Boosting Regression (GBR), and Extreme Gradient Boosting (XGBoost, XGB) [17], have shown increasing promise for groundwater quality prediction [18–20]. Unlike kriging, ML captures non-linear relationships among multiple covariates—including spatial coordinates, depth, and temporal variables—while using this spatial information without imposing the stationarity (constant-mean/variogram) assumption on which kriging relies [21]. Several studies have reported ML outperforming kriging for groundwater nitrate prediction [22–24], though typically where spatial autocorrelation remains reasonably strong (Moran's $I > 0.3$). Whether ML offers a genuine advantage when spatial autocorrelation is specifically weak—as commonly encountered at point-source sites—has not been systematically addressed.

Hybrid approaches combining ML trend estimation with geostatistical residual interpolation represent an emerging strategy. Regression kriging (RK) [25] decomposes the target variable into a deterministic trend (modeled by ML) and a spatially correlated residual (modeled by kriging), potentially combining the strengths of both. Adaptive Regression Kriging (ARK), which applies density-weighted fusion of ML and kriging predictions, offers further refinement but remains largely unexplored for groundwater contamination mapping. The theoretical appeal of these hybrids lies in producing spatially coherent

contamination surfaces while leveraging ML predictive power, yet their practical value under weak spatial structures has not been empirically established.

Beyond concentration prediction, ML-based exceedance probability mapping provides directly actionable risk information for regulatory decision-making [26,27]. By classifying locations as exceeding or not exceeding a regulatory threshold, such maps translate complex predictions into a binary risk metric directly interpretable by environmental managers. This is particularly relevant under China's GB/T 14848-2017 standard [28], which defines the Class III $\text{NH}_4^+\text{-N}$ threshold at 0.5 mg/L. Exceedance probability mapping addresses the challenge of predicting concentrations spanning five orders of magnitude by reframing the problem as a classification task with a single, policy-relevant threshold.

Building on a national-scale groundwater nitrogen contamination map [29], this study addresses site-scale point-source heterogeneity through a systematic comparison at a chemical industrial park in Northern China, with five objectives: (1) characterize pollution heterogeneity and spatial autocorrelation using global statistics and variography; (2) evaluate six geostatistical interpolators; (3) test whether additional covariates (including non-stationary spatial covariates) enable ML to outperform pure kriging; (4) assess whether hybrid ML-kriging produces physically plausible and accurate maps; and (5) generate exceedance probability maps for the Class III threshold. The comparison covers geostatistics, ML, and hybrids under explicitly diagnosed weak spatial autocorrelation—a condition typical of point-source contamination but seldom systematically examined.

2. Materials and Methods

2.1. Study Area and Data

The study area is a chemical industrial park located in a region of Shanxi Province, northern China. The park hosts coking and coal-chemical production, metallurgical smelting, carbon black and pharmaceutical manufacturing, and a centralized wastewater-treatment facility—potential point sources via historical leakage from tanks, treatment stations, tar storage, and pipelines. A river and a highway traverse the zone. Six villages (anonymized as V1–V6) lie within or adjacent to the park. Their untreated supply wells—the nearest of which are located approximately 200–400 m from contaminated monitoring wells—tap the same shallow phreatic aquifer that receives the leakage, which is also used for irrigation.

The host county covers approximately 600 km² with a resident population on the order of 3.5×10^5 (urbanization $\approx 45\%$; ≈ 570 persons km^{−2}; county statistical communiqué, 2024). For the 78 km² study area (the extent of the 133-well network), WorldPop's UN-adjusted 1 km gridded dataset (Global2, 2024) gives $\approx 66,900$ residents (≈ 850 persons km^{−2}, above the county mean because the area is peri-urban; used as an order-of-magnitude estimate), implying a domestic demand of $\approx 5400\text{--}8000$ m³ d^{−1} ($\approx 2.0\text{--}2.9$ Mm³ yr^{−1}) at 80–120 L capita^{−1} d^{−1}, drawn largely from the contaminated aquifer.

Linking this population to the monitoring network through the same 1 km grid indicates that $\approx 22,000$ residents—about one-third of the population living in the grid cells covered by the monitoring wells—rely on groundwater in which the upper aquifer exceeds the Class III threshold (0.5 mg/L), equivalent to $\approx 1800\text{--}2700$ m³ d^{−1} of domestic demand met by non-compliant groundwater. Ammonium itself has low direct toxicity, but it indicates contaminant ingress at coking/coal chemical sites and is readily nitrified to nitrite and nitrate—the species of established drinking-water health concern [1]—to which this population is exposed by direct ingestion of untreated supplies. These are screening-level estimates and are intended to prioritize rather than to quantify health risk.

Hydrogeologically, quaternary alluvial deposits overlie bedrock, forming an unconfined to semi-confined system. The shallow aquifer comprises silty/fine sand with clay,

whereas the deeper aquifer comprises medium–coarse sand with gravel interbeds. Ground elevation is approximately 730–750 m a.s.l.; the water table is shallow (0.1–4.6 m below ground surface, mean 1.6 m), and regional groundwater flow is directed from NW to SE. The thin, permeable vadose zone (silty and fine to medium sands with clay lenses) offers limited attenuation, favoring rapid vertical migration.

A network of 133 monitoring wells, each screened at upper (5–15 m) and lower (20–35 m) depths, was established across the industrial park [30] following HJ 164-2020 [31] and national guidelines, representing the recharge–runoff–discharge system with three functions: upgradient reference (eastern margin), internal source zone wells within or downgradient of the above facilities, and boundary/diffusion wells along the sides and downgradient of near receptors. Density is highest around known contaminated/high-risk and exceedance zones, and lowest in peripheral background/upgradient areas. The network spans ~13 km (E–W) by 5 km (N–S), with ~40 km² including a ~1 km buffer, and was installed in phases with targeted densification. This point-source design oversamples hotspots, contributing to extreme contrasts and weak spatial autocorrelation. Four campaigns (April, August, September, and October 2024) sampled NH₄⁺-N via Nessler’s reagent spectrophotometry (HJ 535-2009) [32]; non-detects were set to half MDL (0.025 mg/L). The upper-layer monitoring wells and the associated NH₄⁺-N contamination are distributed across the study area (Figure 1). Across the campaigns, the upper aquifer was sampled at all 133 wells (532 observations); the lower aquifer at 133 wells in August–October (399 observations), with only three excluded from modeling in April. Upper-aquifer observations per campaign (133 each) were used for global Moran’s I, variography, and six geostatistical interpolators. ML and hybrid ML–kriging models pooled upper and lower observations plus temporally interpolated virtual samples (upper only, ~2–3 per well per interval); retained sample sizes after $1.5 \times \text{IQR}$ filtering are reported in Section 3.5 and the Supplementary Materials (e.g., $N = 107$ for XGB and $N = 106$ for RK/ARK in the April upper-layer case). The exceedance probability classifier was trained on the same pooled dataset (~70:30 exceed/non-exceed ratio), with SMOTE balancing within each CV fold.

2.2. Descriptive Statistics

For each campaign, basic statistics (mean, median, standard deviation (SD), coefficient of variation (CV), skewness) and exceedance rates against the GB/T 14848-2017 Class III threshold (0.5 mg/L) were computed [28]. Shapiro–Wilk (raw and log₁₀ transformed) and Kolmogorov–Smirnov (K-S, for lognormal fit) tests were applied per campaign to assess distributional assumptions (Table S1), informing data transformation strategies for subsequent geostatistical and ML analyses.

2.3. Spatial Structure Analysis

Global Moran’s I [33] was computed with k-nearest-neighbor spatial weights ($k = 8$) for log₁₀-transformed upper-aquifer concentrations across all four campaigns, and its significance was evaluated using 999 random permutations ($\alpha = 0.05$). Moran’s I was interpreted within a significance-testing framework [34]. As an operational criterion for this study, spatial autocorrelation was regarded as too weak for reliable geostatistical prediction when Moran’s I fell below approximately 0.15 and/or the permutation test was non-significant. Local Indicators of Spatial Association (LISA) [35] were computed to identify statistically significant spatial clusters and outliers (Figure S1). Moran scatter plots were constructed to visualize the relationship between observed values and their spatial lags (Figure S5).

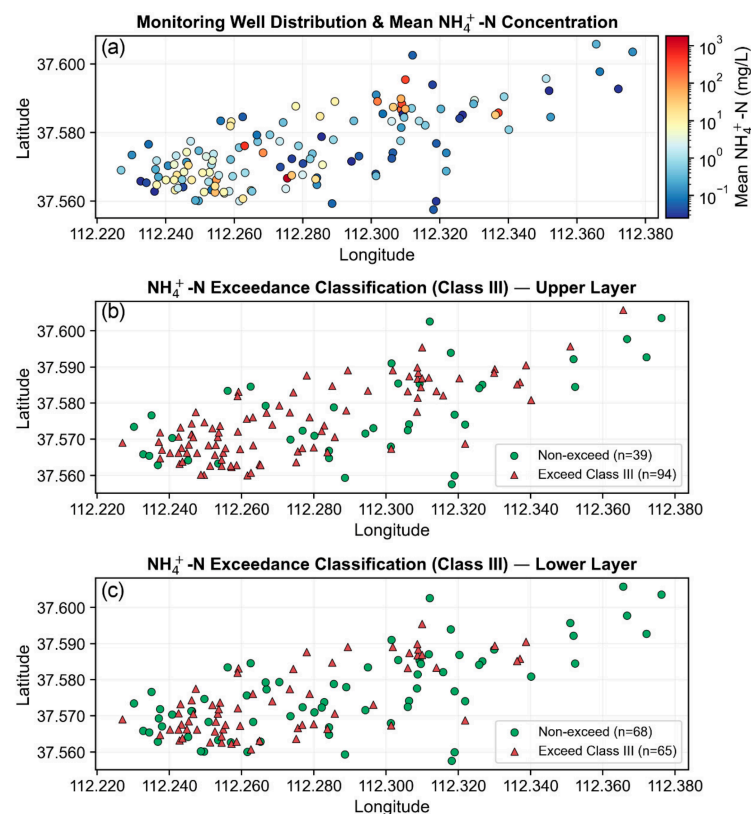


Figure 1. Distribution of the 133 monitoring wells and $\text{NH}_4^+\text{-N}$ contamination characteristics across both aquifer layers. (a) Spatial distribution of the wells colored by mean upper-aquifer $\text{NH}_4^+\text{-N}$ concentration (mg/L, log scale). (b) Class III exceedance classification in the upper aquifer: non-exceeding wells (green circles, $n = 39$) and Class III-exceeding wells (red triangles, $n = 94$). (c) Class III exceedance classification in the lower aquifer: non-exceeding wells (green circles, $n = 68$) and Class III-exceeding wells (red triangles, $n = 65$). Exceedance is defined by the temporal mean concentration relative to the Class III threshold (0.5 mg/L; GB/T 14848-2017).

Experimental variograms were computed for $\log_{10}$ -transformed concentrations and fitted with three theoretical models—spherical, exponential, and Gaussian—for each campaign (Figure S2). The best-fitting model was selected by minimizing the root mean square error (RMSE) between experimental and theoretical semivariograms. The nugget ratio $C_0/(C_0 + C)$ was used to classify spatial dependence strength following Cambardella et al. [15]: values <25% indicate strong spatial dependence, 25–75% moderate dependence, and >75% weak dependence dominated by random variation. This dual-diagnostic approach—combining global autocorrelation statistics with variogram-based classification—provides a robust assessment of whether the spatial structure is sufficient to support geostatistical prediction. Global Moran's I and variogram analyses were performed for the upper aquifer only, because it is the primary receptor of surface industrial leakage and has the most complete and densely sampled record across all four campaigns. The lower aquifer, sampled less completely, is characterized through Class III exceedance classification (Figure 1c) rather than variogram modeling. A full observation inventory per analysis is given at the end of Section 2.1.

2.4. Geostatistical Interpolation

Six interpolation methods were evaluated: IDW, OK, SK, CK (using surface elevation as the secondary variable), IK (with the Class III threshold as the indicator cutoff), and EBK. All geostatistical analyses were performed using ArcGIS Pro 2.5.0 (Esri Inc., Redlands, CA, USA). For kriging-based methods, the variogram model yielding the lowest RMSE

was applied for each campaign. Spatial interpolation surfaces were generated for each method–campaign–layer combination on a regular grid covering the study area.

Prediction accuracy was assessed by leave-one-out cross-validation (LOO-CV), in which each observation was sequentially withheld and predicted from the remaining data. This approach is the standard validation method for geostatistical interpolation, as the limited spatial sample sizes ($n = 133$ per campaign) preclude the use of independent holdout sets without substantial information loss. Performance was evaluated using the coefficient of determination (R^2) computed on raw (untransformed) concentrations, enabling direct assessment of predictive accuracy on the original measurement scale.

2.5. Machine Learning Prediction

We constructed a spatiotemporal prediction dataset combining upper- and lower-aquifer observations across all four campaigns. To augment training data, temporal interpolation between consecutive campaigns generated virtual samples at intermediate time points (Supplementary Methods S1): cubic spline interpolation for wells with all four measurements, and linear interpolation for those with fewer observations. This augmentation, applied only to the upper aquifer, yielded approximately 2–3 virtual samples per well per inter-campaign interval, under the rationale that NH_4^+ -N concentrations are expected to vary smoothly between campaigns, making interpolated values physically plausible representations of intermediate states.

The base feature set for ML model comparison included eight variables: longitude; latitude; screen elevation (Z); month; two cyclic temporal encodings ($\text{time_sin} = \sin(2\pi \times \text{month}/12)$ and $\text{time_cos} = \cos(2\pi \times \text{month}/12)$); aquifer-layer code (0 = upper, 1 = lower); and Euclidean distance to well centroid (dist_center). The cyclic encodings ensures December and January are treated as adjacent months, while dist_center proxies the spatial gradient from the centrally concentrated contamination source zone. These eight features were selected a priori based on their mechanism and availability: coordinates and dist_center encode the source to periphery gradient; screen elevation and layer code capture vertical stratification; month with cyclic encodings represents seasonal recharge dynamics.

Six regression models were evaluated: RF, XGB, GBR, Support Vector Regression (SVR), k-nearest-neighbor regression (KNN), and AdaBoost regression (ADB). Detailed hyperparameters for all models are provided in Table S2. Hyperparameters were selected through preliminary grid search experiments and are representative of well-tuned configurations for the dataset characteristics. Model performance was evaluated using 10-fold cross-validation with three metrics— R^2 , RMSE, and mean absolute error (MAE)—computed on $\log_{10}$ -transformed concentrations. R^2 , RMSE, and MAE were all computed from the same set of 10-fold out-of-fold predictions, using identical folds across all models, so that the three metrics were mutually consistent and the models were compared on identical partitions. The $\log_{10}$ transformation was applied to stabilize variance across the five-order-of-magnitude concentration range and to improve the symmetry of the error distribution. All ML models were implemented using scikit-learn [36] (version 1.3) and XGBoost [17] (version 2.0) in Python 3.10.

2.6. Hybrid ML–Kriging Methods

Three hybrid approaches combining ML prediction with geostatistical smoothing were implemented [37,38].

The three hybrids span an increasing degree of ML–geostatistics coupling: (i) “ML Improved” isolates the effect of an enriched spatial feature set on pure XGB; (ii) regression kriging (RK) is the canonical trend-plus-residual decomposition, which tests whether kriging the residuals recovers spatial coherence lost by tree-based outputs; and (iii) Adaptive

Regression Kriging (ARK) adds density-weighted fusion to guard against extrapolation in data-sparse areas.

(1) ML Improved applies pure XGB prediction on the interpolation grid using an extended feature set (ten total), with distance-to-center replaced by three spatial interaction terms (longitude $\times$ latitude, $Z \times$ longitude, and $Z \times$ latitude). These spatial covariates allow the tree ensemble to represent anisotropic, location-dependent trends without kriging's stationarity assumption and capture non-linear spatial dependencies beyond simple coordinates (Supplementary Methods S1).

(2) RK decomposes the target into an ML trend and a kriged residual (Equation (1)) [25]:

$$\hat{Z}(S) = \hat{m}_{ML}(s) + \hat{e}_{OK}(s) \quad (1)$$

where $\hat{m}_{ML}(s)$ is the XGB trend, and $\hat{e}_{OK}(s)$ is the ordinary kriging of the ML residuals. This decomposition smooths step-like discontinuities inherent in tree-based predictions while preserving XGB accuracy.

(3) ARK applies a density-weighted fusion, whose weight is defined by Equation (2):

$$w(s) = 1 / (1 + \exp(-k(d(s) - D_O))) \quad (2)$$

where the weight depends on the distance to the nearest monitoring well $d(s)$. In data-dense areas, ML receives a higher weight; in data-sparse areas, the kriging weight increases to avoid extreme extrapolation.

LOO-CV was performed under a “global mode” scheme (only the test observation were withheld; other campaigns of the same well were retained), which may overestimate spatial generalization, particularly for later campaigns (see Section 4.5). R^2 and RMSE were computed on $\log_{10}$ -transformed concentrations. For outlier handling, $\log_{10}$ residuals were screened using the $1.5 \times \text{IQR}$ rule; points outside $[Q1 - 1.5\text{IQR}, Q3 + 1.5\text{IQR}]$ were excluded as ultra-high-leverage outliers. This criterion flagged one extreme residual in hybrid (RK and ARK) predictions and none in XGB, yielding retained sample sizes of $N = 107$ for XGB and $N = 106$ for RK/ARK (reported in Section 3.5). Recomputing metrics without removal changed values only marginally and did not alter method ranking or any conclusion.

2.7. Exceedance Probability Mapping

XGB was adapted as a binary classifier to predict the probability of $\text{NH}_4^+\text{-N}$ exceeding the Class III threshold (0.5 mg/L). Five additional classifiers were evaluated for comparison: Random Forest Classifier (RFC), Support Vector Classifier (SVC), k-Nearest Neighbors Classifier (KNNC), AdaBoost Classifier (ADBC), and Gradient Boosting Classifier (GBC). Class imbalance ($\sim 7:3$ ratio of exceeding to non-exceeding observations) was addressed using the Synthetic Minority Oversampling Technique (SMOTE) [39,40] applied within each fold of the stratified cross-validation to avoid data leakage. Model discrimination was assessed using receiver operating characteristic (ROC) analysis and the area under the ROC curve (AUC; Figure S3). The best-performing classifier was then applied to the interpolation grid to generate spatial exceedance probability maps for each campaign.

3. Results

3.1. Pollution Characteristics

$\text{NH}_4^+\text{-N}$ concentrations in the upper aquifer exhibited extreme heterogeneity across all four campaigns (Table S3). Mean concentrations ranged from 27.62 mg/L (September 2024) to 40.62 mg/L (April 2024), while median concentrations were dramatically lower (0.082–1.290 mg/L), reflecting the dominance of a small number of heavily contaminated

wells. CV values ranged from 3.70 to 7.92 and skewness from 4.77 to 11.16, confirming severe right-skewness. Overall, concentrations spanned from below the detection limit to 3682 mg/L. The Shapiro–Wilk test rejected normality for raw data in all campaigns ($p < 0.001$; Table S1), supporting $\log_{10}$ transformation. The K-S test indicated that only April 2024 passed the lognormal distribution test ($p = 0.368$), while the other three campaigns were rejected ($p < 0.001$; Table S1), suggesting that $\log_{10}$ transformation does not fully normalize the distributions in all cases.

The spatial distribution of well monitoring (Figure 1a) revealed elevated concentrations in the central–southeastern sector and low concentrations in the periphery. Among 133 wells, 94 (70.7%) exceeded the Class III threshold based on the temporal mean concentration (Figure 1b). In the lower aquifer, 65 of 133 wells (48.9%) exceeded the Class III threshold based on the temporal mean concentration (Figure 1c), indicating that contamination is not confined to the shallow layer. The exceedance rate peaked at 62.4% in April 2024, declined to 35.3% in August, and stabilized around 40–42% in September–October (Table S3), suggesting partial dilution from summer rainfall recharge followed by stabilization. The interquartile range of $\log_{10}$ -transformed concentrations was broadly stable across the campaigns despite seasonal shifts in the median (Figure S4).

3.2. Weak Spatial Structure: A Diagnostic Finding

Both Moran’s I and variogram fitting revealed consistently weak spatial dependence (Table 1). Global Moran’s I ranged from 0.040 (August) to 0.118 (October)—well below the 0.3–0.7 range typical of agricultural non-point-source studies [14,41]. Only September and October reached statistical significance ($p < 0.01$); April ($p = 0.068$) and August ($p = 0.216$) were not significant at $\alpha = 0.05$, indicating that spatial clustering was not reliably detectable above random noise.

Table 1. Spatial structure diagnosis for $\log_{10}$ -transformed NH_4^+ -N in the upper aquifer: global Moran’s I, permutation test statistics, and variogram nugget ratio of the best-fit model.

Campaign	Moran’s I	Z-Score	<i>p</i> -Value	Nugget Ratio $C_0/(C_0 + C)$ (%)	Best Variogram Model	Spatial Dependence
April 2024	0.063	1.83	0.068	83.8	Gaussian	Weak
August 2024	0.040	1.24	0.216	104.8	Gaussian	Weak
September 2024	0.103	2.87	0.004	75.1	Exponential	Weak
October 2024	0.118	3.26	0.001	86.5	Gaussian	Weak

Notes: Moran’s I computed with k-nearest-neighbor spatial weights and 999 random permutations ($\alpha = 0.05$). Nugget ratio = $C_0/(C_0 + C) \times 100\%$ of the best-fit variogram model (minimum RMSE; the full fitting parameters for all three models are in Table S4). August is a pure-nugget case (nugget ratio $> 100\%$), so spatially structured variograms could not be fitted. All four campaigns indicate weak spatial dependence.

Variogram analysis corroborated this diagnosis: nugget ratios of 75.1–104.8% (Table S4) uniformly indicate weak spatial dependence by Cambardella’s classification [15]. For the August campaign the variogram could not be reliably fitted (nugget ratio $\geq 100\%$), precluding a meaningful spatial-dependent estimate for that campaign. The three fitted variogram models produced similar RMSE values within each campaign (Table S4; Figure S2), confirming that the poor spatial structure is intrinsic rather than an artifact of model selection.

LISA analysis (Figure S1) identified localized clusters—14 high–high hot spots and 16 low–low cold spots—but $>75\%$ of wells were classified as non-significant, consistent with weak global autocorrelation. Moran scatter plots (Figure S5) showed a diffuse cloud with a minimal positive spatial lag correlation.

This weak autocorrelation reflects the inherent nature of point-source industrial contamination; heavily contaminated wells adjacent to leaking facilities create extreme contrasts against a predominantly low-concentration background, producing localized plumes with

sharp boundaries rather than the broad gradients generated by diffuse agricultural inputs. This suggests that the geostatistical framework—which assumes meaningful spatial dependence—may be fundamentally mismatched to industrial groundwater contamination.

3.3. Failure of Geostatistical Methods

The weak spatial structure translated directly into poor geostatistical prediction performance. LOO-CV revealed that all six methods achieved negative or near-zero R^2 values across all campaign–layer combinations (Table S5). Among the kriging variants, OK yielded R^2 values of -0.01 to -0.07 , indicating predictions no more accurate than the global mean. SK, CK, and EBK performed comparably to OK, while IDW performed substantially worse ($R^2 = -0.62$ to -0.81) because its distance-weighted averaging was strongly influenced by extreme outlier values. CK with elevation offered no improvement due to the minimal topographic relief across the study area, which provided no useful secondary information for predicting contamination patterns. IK showed the best relative performance among geostatistical methods ($R^2 = -0.01$ to 0.00), which is expected given that IK operates on binary indicators rather than extreme concentration values, thereby mitigating the influence of outliers.

Despite this quantitative failure, the interpolation maps (Figure 2) provided qualitatively reasonable visual representations of the contamination plume, with all methods identifying the high-concentration zone in the central–southeastern sector. This illustrates an important practical distinction: geostatistical methods may retain utility for approximate plume delineation even when their point-prediction accuracy is demonstrably poor, because the visual interpolation surface captures the broad spatial pattern even if individual predictions are unreliable. We emphasize that a raw-scale R^2 across five orders of magnitude is near-zero or negative for almost any interpolator—as the IDW results ($R^2 = -0.62$ to -0.81) illustrate—so these values reflect the difficulty of raw-scale point prediction rather than a lack of intrinsic method skill. The geostatistical (raw scale, LOO-CV) and ML ($\log_{10}$ -scale, 10-fold CV) R^2 values are therefore indicative and not directly comparable. Results for other campaigns and layers are shown in Figures S6 and S7.

3.4. Machine Learning Improvement

The six ML models achieved markedly higher prediction accuracies than the geostatistical methods when evaluated on $\log_{10}$ -transformed concentrations via 10-fold cross-validation (Figure 3). Tree-based ensemble methods—XGB, GBR, and RF—consistently outperformed the remaining models across all three metrics. XGB achieved the highest R^2 (0.75 ± 0.04), the lowest RMSE (0.55 ± 0.03), and the lowest MAE (0.30 ± 0.02), and was therefore selected as the best-performing model for subsequent hybrid prediction and exceedance probability mapping. Although XGBoost, GBR, and RF rank closely in R^2 , paired per-fold tests showed that XGB and GBR were statistically indistinguishable (paired t -test $p = 0.62$; Wilcoxon $p = 0.38$; $\Delta R^2 = 0.008$), whereas both significantly outperformed RF (paired t -test $p \leq 0.002$). XGB was therefore selected not on the basis of a statistically decisive margin over GBR, but because it attained the highest mean R^2 together with the lowest mean RMSE and MAE on identical folds, and because it provides native handling of the exceedance probability classification task used subsequently. SVR and KNN showed intermediate performance ($R^2 \approx 0.60$ – 0.65), while ADB displayed the lowest R^2 (~ 0.50) among the six models. Feature interpretability analysis (Figure S12) confirmed that distance-to-centroid was the dominant predictor, followed by latitude, longitude and well-screen elevation, with SHAP values and permutation importance giving a consistent ranking; the predominantly negative SHAP contribution of distance-to-centroid at large values indicates that NH_4^+ -N decreases away from the central source zone.

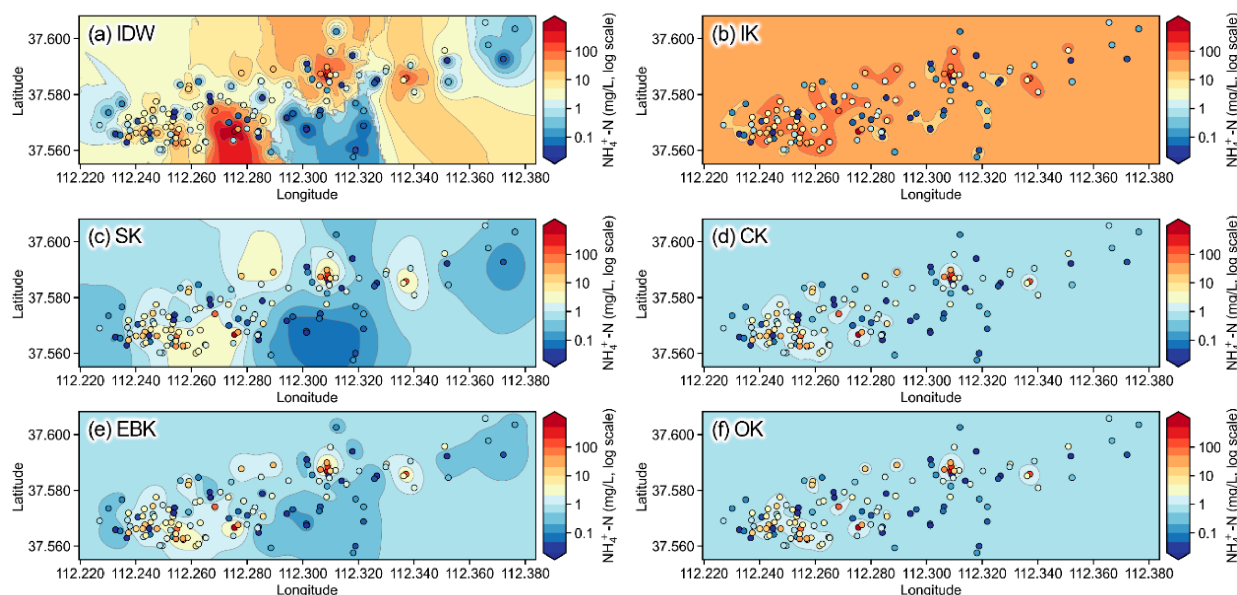


Figure 2. Spatial interpolation maps of upper-aquifer $\text{NH}_4^+\text{-N}$ for April 2024 using six geostatistical methods: (a) IDW; (b) IK; (c) SK; (d) CK (secondary variable: well head elevation Z via external drift, interpolated to grid by IDW from same wells); (e) EBK; and (f) OK. Colored surfaces show $\log_{10}$ concentrations (common color bar); overlaid circles are wells colored by measured values; and gray lines are isolines. All methods identify the central–southeastern high-concentration zone; IDW is strongly affected by extremes; CK offers no improvement given minimal relief; and IK, operating on indicators, is more robust to outliers. Corresponding results for other campaigns and layers are provided in the Supplementary Materials (Figures S6 and S7). Primary variable and residuals were kriged in log space with spherical variogram.

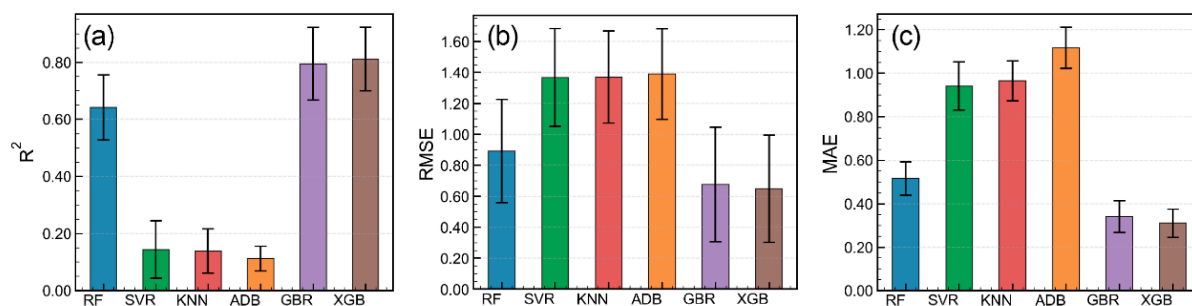


Figure 3. Comparison of the predictive performance of six machine learning regression models for groundwater $\text{NH}_4^+\text{-N}$ concentration prediction. (a) R^2 , (b) RMSE, and (c) MAE evaluated via 10-fold cross-validation on $\log_{10}$ -transformed concentrations. Error bars denote ± 1 standard deviation across the 10 cross-validation folds. RF = Random Forest; SVR = Support Vector Regression; KNN = k-nearest-neighbor; ADB = AdaBoost; GBR = Gradient Boosting Regression; XGB = XGBoost.

The jump from near-zero R^2 (geostatistics on raw concentrations) to positive R^2 (ML on $\log_{10}$ -transformed concentrations) highlights the value of covariates beyond kriging's reach: distance-to-center (used without stationarity) captures the source-zone gradient, well-screen elevation reflects vertical stratification, cyclic temporal variables encode seasonality, and the layer code distinguishes aquifer-specific distributions. However, direct equivalence between geostatistical and ML metrics is precluded by differing scales (raw vs. log-transformed) and validation schemes (LOO-CV vs. 10-fold CV). ML and hybrids showed higher apparent log-scale accuracy, but given these discrepancies, the comparison indicates directional differences between paradigms rather than a direct ranking.

3.5. Hybrid ML–Kriging Prediction

LOO-CV results for the three hybrids showed strong predictive skill across all campaign–layer combinations (Table 2). In the hardest case—April 2024’s upper layer, the first campaign with minimal training history—XGB achieved $R^2 = 0.673$ (RMSE = 0.601), while RK and ARK both scored $R^2 = 0.661$ (RMSE = 0.599) (Figure 4). Across campaigns, this April 2024 upper-layer result offers the most representative estimate of spatial generalization to unmonitored sites, as it was the only campaign with no prior within-well information retained in training. In contrast, the $R^2 > 0.97$ for August–October (Table 2) should be viewed as in-sample (optimistic) estimates, because same-well records from other campaigns and temporally interpolated virtual samples derived from them remained in the training set under the global-mode LOO-CV scheme (Section 4.5). Results for additional campaigns and layers are shown in Figures S8 and S9.

Table 2. LOO-CV accuracy of three hybrid ML–kriging methods (log₁₀ scale).

Campaign	Layer	XGB R^2	XGB RMSE	RK R^2	RK RMSE	ARK R^2	ARK RMSE
April 2024	Upper	0.673	0.601	0.661	0.599	0.661	0.599
August 2024	Upper	0.987	0.135	0.988	0.130	0.988	0.130
September 2024	Upper	0.998	0.055	0.998	0.059	0.998	0.059
October 2024	Upper	0.997	0.074	0.995	0.090	0.995	0.090
August 2024	Lower	0.975	0.204	0.972	0.219	0.972	0.219
September 2024	Lower	0.994	0.100	0.993	0.112	0.993	0.112
October 2024	Lower	0.995	0.097	0.994	0.107	0.994	0.107

Notes: The April 2024 upper-layer values (bold) best represent spatial generalization to unmonitored locations, as it is the only campaign trained without retaining any within-well information from prior campaigns. The high R^2 values for August–October partly reflect retention of same-well records (and their temporally interpolated virtual samples) from other campaigns under the global-mode LOO-CV scheme (Section 4.5), and are thus in-sample (optimistic) estimates. XGB = XGBoost; RK = regression kriging; ARK = Adaptive Regression Kriging; RMSE on log₁₀ scale. RK and ARK yield identical metrics because the near-structureless residual (nugget ratio > 75%) reduces ARK to RK.

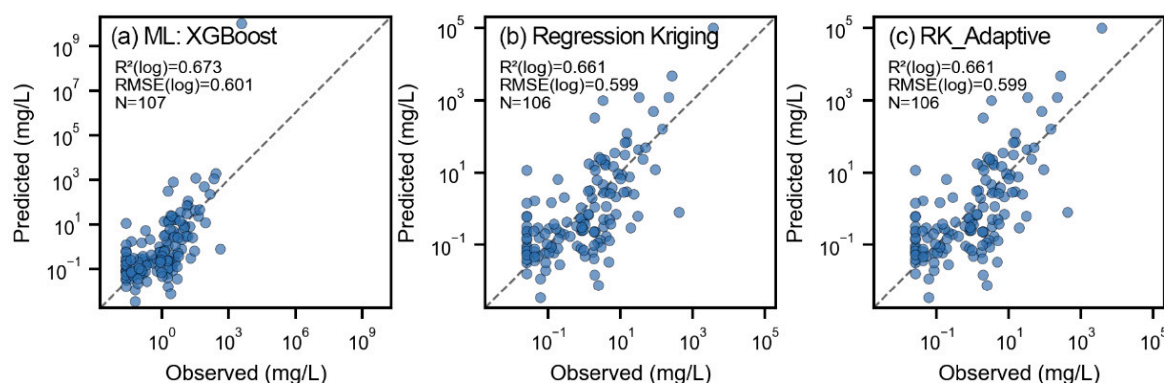


Figure 4. Leave-one-out cross-validation results for groundwater NH_4^+ -N concentration prediction in April 2024 for the upper layer using three hybrid modeling approaches: (a) XGB, (b) regression kriging, and (c) Adaptive Regression Kriging. Scatter plots compare predicted and observed concentrations (mg/L) on logarithmic scales. The dashed line represents the 1:1 line. R^2 , RMSE, and N (number of wells retained after applying a $1.5 \times \text{IQR}$ rule to log₁₀ residuals; see Section 2.6) are shown in each panel.

The scatter plots (Figure 4) demonstrate reasonable agreement between observed and predicted values across approximately five orders of magnitude. Notably, the slightly lower R^2 of the augmented kriging methods (RK, ARK) compared to pure XGB reflects the weak residual spatial structure, where the kriged residual component adds marginal noise rather

than systematic spatial information, resulting in negligibly reduced accuracy rather than the improvement typically expected from hybrid approaches.

The spatial prediction maps (Figure 5) illustrate the visual characteristics of each approach. XGB (Figure 5a) exhibited step-like discontinuities characteristic of tree-based models, producing “blocky” concentration boundaries. RK (Figure 5b) produced smoother, more physically plausible surfaces by attenuating these artifacts through the kriged residual component, which distributes prediction errors spatially even when the residual structure is weak. ARK (Figure 5c) showed a similar smoothing pattern, with seamless transitions between ML-dominated (data-dense) and kriging-dominated (data-sparse) zones. All three methods consistently reproduced the high-concentration zone in the central–southeastern sector. Results for other campaigns and layers are provided in Figure S10.

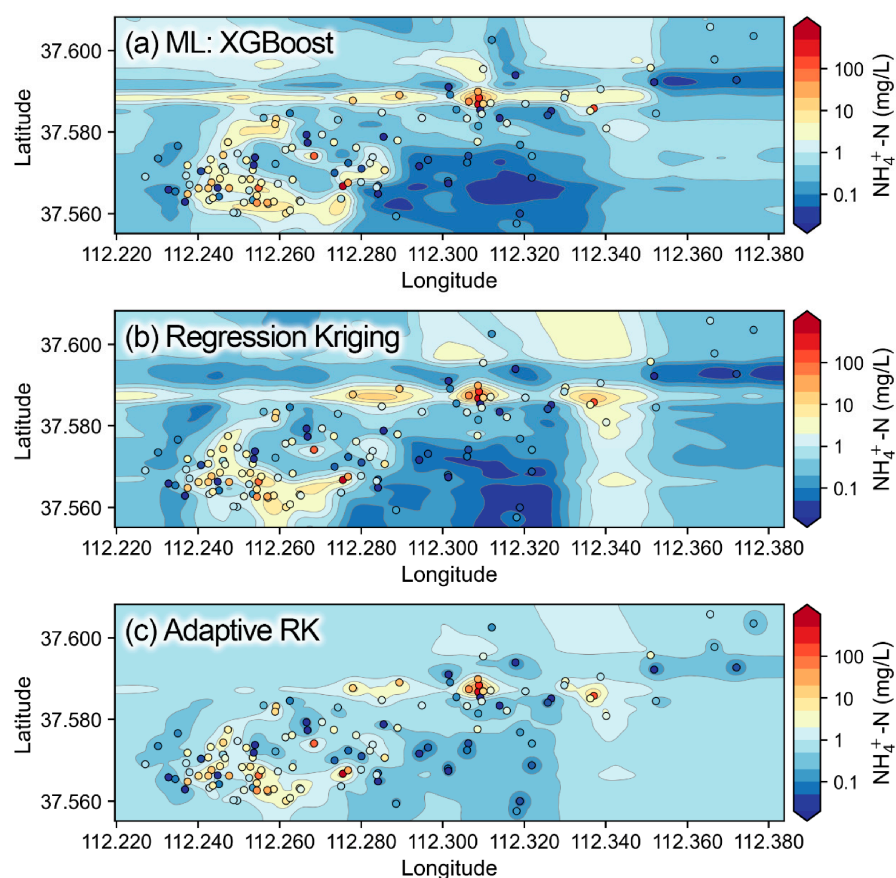


Figure 5. Spatial prediction maps of groundwater $\text{NH}_4^+\text{-N}$ concentrations for April 2024 in the upper layer: (a) XGB, (b) regression kriging, and (c) Adaptive Regression Kriging. The colored surfaces represent predicted concentrations (mg/L) on a common logarithmic color scale. Overlaid circles indicate monitoring wells colored by measured concentrations; gray contour lines denote concentration isolines. Corresponding results for other campaigns and layers are provided in the Supplementary Materials (Figure S10).

3.6. Exceedance Probability Mapping

The XGB classifier achieved excellent discriminatory performance for predicting Class III exceedance, with an AUC of 0.959, closely followed by GBC (AUC = 0.958) and RFC (AUC = 0.943; Table 3; Figure S3). The remaining models showed lower but acceptable discrimination (AUC 0.758–0.803).

The exceedance probability maps (Figure 6) revealed a spatially persistent high-risk zone in the central–southeastern sector across all four campaigns. In April 2024 (Figure 6a), the high-probability zone (≥ 0.8) was most extensive, consistent with the highest exceedance

rate observed in this campaign. In subsequent months (Figure 6b–d), the core high-risk zone persisted while its spatial extent showed seasonal contraction and recovery. Non-exceeding wells (green circles) were consistently located in peripheral low-probability zones, and exceeding wells (red triangles) fell within or near the high-probability contours, confirming spatial concordance between observed and predicted exceedances.

Table 3. Classification performance of six ML models for Class III exceedance prediction.

Model	Accuracy	AUC	F1	Precision	Recall
XGB	0.904	0.959	0.904	0.907	0.901
GBC	0.908	0.958	0.908	0.907	0.909
RFC	0.880	0.943	0.878	0.892	0.865
KNNC	0.711	0.803	0.734	0.680	0.797
SVC	0.734	0.800	0.744	0.718	0.772
ADBC	0.699	0.758	0.685	0.719	0.654

Notes: Binary classification— $\text{NH}_4^+\text{-N} > 0.5$ mg/L (Class III, GB/T 14848-2017). AUC = area under the ROC curve. Ranked by AUC: XGB = XGBoost Classifier; GBC = Gradient Boosting Classifier; RFC = Random Forest Classifier; KNNC = k-Nearest Neighbors Classifier; SVC = Support Vector Classifier; ADBC = AdaBoost Classifier. SMOTE was applied for class balancing (see Supplementary Methods S2).

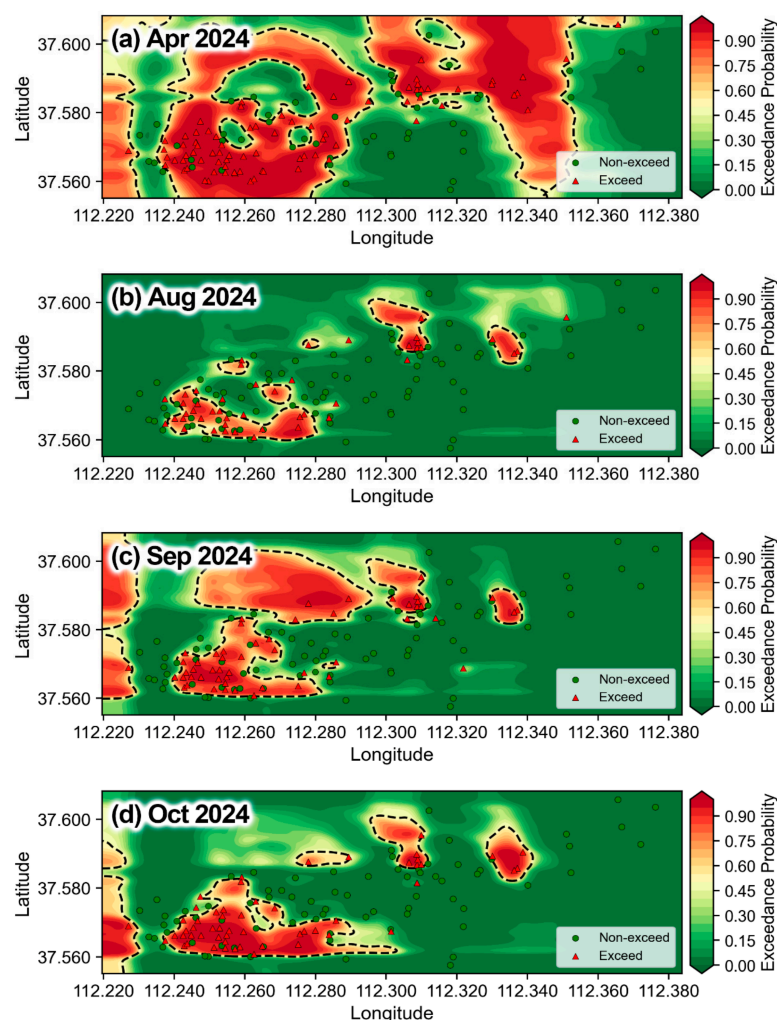


Figure 6. Predicted exceedance probability (XGB classifier) for (a) April, (b) August, (c) September, and (d) October 2024, color-scaled for $\text{NH}_4^+\text{-N}$ exceeding the Class III standard (0.5 mg/L, GB/T 14848-2017). Black dashed contours denote the $p = 0.8$ isoline, delineating high-risk zones ($\geq 80\%$ exceedance probability). Green circles and red triangles indicate observed non-exceedances and exceedances, respectively.

The temporal persistence of the high-risk zone across all four campaigns, despite seasonal variations in concentration magnitude, strongly suggests a continuous, active contamination source rather than a historical release undergoing natural attenuation. This finding has direct implications for remediation prioritization, indicating that source control in the central–southeastern sector should take precedence over downstream plume management.

4. Discussion

4.1. Interpreting the Contrast Between Geostatistical and Machine Learning Performance

The consistently negative R^2 values across all geostatistical methods (Table S5) are a direct consequence of the weak spatial autocorrelation diagnosed in Section 3.2. With Moran's I values of only 0.040–0.118 and nugget ratios uniformly exceeding 75%, kriging's fundamental assumption of spatial correlation is barely supported. This is a predictable outcome of point-source industrial contamination, which produces highly localized plumes with extreme concentration contrasts—a well adjacent to a leaking waste pond may exceed 1000 mg/L while its neighbor 200 m away reads below 0.1 mg/L—creating spatial variance dominated by the nugget component.

The accuracy improvement achieved by tree-based ML models stems from three complementary factors: (1) integration of covariates that carry predictive information—including spatial covariates (coordinates, distance-to-center, coordinate interaction terms) used without a stationarity assumption, together with well-screen elevation and temporal encodings—which pure kriging cannot exploit; (2) inherent robustness of ensemble tree methods to extreme outliers and non-linear relationships; and (3) the cumulative benefit of multi-campaign training data augmented by temporal interpolation, as evidenced by the improvement from $R^2 \approx 0.67$ (April) to >0.99 (September–October; Table 2). These findings are consistent with recent studies reporting ML superiority for groundwater quality prediction [18,23,42,43]. In contrast to these studies—many of which operate under stronger spatial autocorrelation, as noted in the Introduction (Moran's $I > 0.3$)—the present dataset exhibits explicitly diagnosed weak autocorrelation, which is the distinctive condition under which our comparison is made; our XGB's skill ($R^2 \approx 0.75$ on $\log_{10}$; classification AUC = 0.959) is comparable while obtained under this harder condition.

4.2. Value of the Hybrid Approach

The hybrid ML–kriging methods addressed a practical limitation of tree-based outputs: spatial incoherence. While XGB achieves the highest point-prediction accuracy, its grid predictions exhibit physically implausible step-like discontinuities (Figure 5a). RK resolved this by decomposing predictions into a smooth ML trend and kriged residual (Figure 5b), producing surfaces more consistent with contaminant transport expectations. ARK (Figure 5c) further prevented extreme extrapolation in data-sparse areas by down-weighting ML predictions where monitoring density is low. Because regression kriging matched XGB in predictive skill (Table 2), the limiting factor for pure kriging here is the missing spatial-trend component rather than the interpolation step: once the ML trend supplies that component, the kriged residual adds little because the residual structure is weak.

Notably, RK and ARK produced identical validation metrics (Table 2) because ARK reduces to RK when the kriged residual carries negligible spatial information: with nugget ratios $> 75\%$, the density-adaptive weight multiplies an essentially structureless residual, so its effect on predictions is below the reported precision. ARK would diverge from RK only under moderate-to-strong residual autocorrelation. This suggests that ARK's added complexity may not be justified when spatial structure is very weak. Under moderate spatial autocorrelation, the adaptive weighting would be expected to provide additional benefit.

Our results suggest a tiered recommendation: pure ML provides the highest accuracy for numerical predictions at unmonitored locations, while hybrid methods offer the best visual plausibility for contamination maps used in stakeholder communication and regulatory reporting. The choice should be guided by the intended application.

4.3. Exceedance Probability for Risk Assessment

The exceedance probability maps (Figure 6) translate ML predictions into a directly interpretable risk metric [44], with a high AUC of 0.959 (Table 3) confirming reliable discrimination. The persistent high-risk zone in the central–southeastern sector indicates a continuous source requiring priority remediation, whereas peripheral low-probability zones can be more confidently classified as low risk. Exceedance probabilities in the 0.4–0.6 band near the 0.5 mg/L threshold carry the largest uncertainty and should be treated as “watch” zones rather than definitive exceedances; nevertheless, this high-risk zone remained spatially stable across all four campaigns (April–October 2024), supporting its use for prioritizing continuous source control. This persistent high-risk zone lies within the study area’s $\approx 66,900$ -resident domain (Section 2.1), where village supply wells sit only approximately 200–400 m from contaminated monitoring wells in the same shallow aquifer. The sustained Class III exceedance thus indicates a credible ingestion exposure pathway for groundwater-dependent communities, supporting source control and, where appropriate, an alternative water supply.

This approach offers several regulatory advantages over traditional concentration mapping: exceedance probabilities are directly interpretable to non-technical stakeholders; the binary classification framework converts a wide-range regression problem into a single-threshold classification, naturally handling extreme heterogeneity; and the temporal consistency of the high-risk zone provides stronger evidence for remediation targeting than any single-campaign map. These merits make exceedance probability mapping especially suitable for regulatory communication and remediation prioritization at contaminated industrial sites.

4.4. Practical Recommendations

Based on our findings, we offer the following methodological recommendations for spatial prediction of groundwater contamination at industrial sites, summarized as a two-step decision table (Table 4).

First, diagnostic assessment of spatial autocorrelation should precede method selection. If Moran’s I falls below approximately 0.15 and/or the nugget ratio exceeds 75%, ML-based approaches should be prioritized over pure geostatistics. This step is particularly important where point-source contamination is suspected.

Second, for multi-campaign monitoring programs, hybrid ML–kriging methods provide the best combination of predictive accuracy and spatial coherence. The temporal augmentation strategy effectively leverages multi-campaign data, with accuracy improving substantially as additional campaigns are incorporated.

Third, for regulatory risk assessment, exceedance probability mapping using XGB classification provides directly actionable spatial information that is more robust and interpretable than deterministic concentration maps under extreme heterogeneity.

Fourth, method selection should reflect the intended end-use: hybrid RK for regulatory reporting requiring smooth maps, pure XGB for quantitative assessment requiring maximum point-prediction accuracy, and exceedance probability mapping for stakeholder communication requiring intuitive risk metrics.

Although demonstrated here for $\text{NH}_4^+\text{-N}$, this workflow is governed by the spatial structure of the data rather than by any specific contaminant, and is therefore applicable to

other point-source pollutants—such as nitrate, trace metals, or organic compounds—where extreme heterogeneity limits conventional interpolation.

Table 4. Decision table for selecting a spatial prediction method from spatial autocorrelation diagnostics. Step 1 chooses the method family from the diagnosed spatial structure; Step 2 refines the choice according to the intended output.

Diagnostic Condition	Spatial Structure Interpretation	Recommended Method
Step 1—Diagnose		
Moran’s I significant ($p < 0.05$) AND nugget ratio low ($< \sim 50\%$)	Strong/moderate spatial autocorrelation; stationarity plausible	Ordinary or regression kriging (OK/RK) is defensible
Moran’s I marginal OR nugget ratio moderate ($\sim 50\text{--}75\%$)	Weak-to-moderate structure; kriging unstable on its own	ML-kriging hybrid (RK/ARK) preferred over pure kriging
Moran’s I non-significant OR nugget ratio high ($> \sim 75\%$)—as at this site ($I = 0.040\text{--}0.118$; nugget $75\text{--}105\%$)	Weak spatial autocorrelation; point-source-dominated	ML or ML-kriging hybrid using spatial covariates without a stationarity assumption; pure kriging degrades to the global mean
Step 2—Match output to purpose		
Smooth, physically coherent surface required (maps, stakeholder/regulatory reporting)	Visual coherence prioritized over point accuracy	Regression kriging (RK)/Adaptive RK (ARK)
Only point prediction/maximum accuracy at unmonitored locations required	Point accuracy prioritized over visual smoothness	XGBoost (XGB)
Threshold-based regulatory risk communication required	Binary exceedance more robust than regression under extreme heterogeneity	XGB exceedance probability classification

Notes: Thresholds ($\sim 50\%$ and $\sim 75\%$ nugget ratio) are practical guides and should be adapted to the site; the study reported here falls in the bottom Step 1 row (Moran’s $I = 0.040\text{--}0.118$; nugget ratio $75\text{--}105\%$), for which ML and ML-kriging hybrids are preferred over pure kriging.

4.5. Limitations and Future Directions

Several limitations should be acknowledged. First, the monitoring period is limited to seven months (four campaigns); longer-term monitoring would be needed to assess interannual variability and ML prediction stability.

Second, the cross-validation schemes differ between geostatistical (LOO-CV on raw concentrations) and ML methods (10-fold CV on $\log_{10}$ -transformed concentrations for model comparison; LOO-CV for hybrid methods). While this reflects standard practice and does not affect the directional conclusion, future studies should adopt a unified evaluation framework—ideally the same scale and validation scheme—for maximal comparability.

Third, the hybrid LOO-CV used a global-mode scheme withholding only the test observation, while same-well data from other campaigns remained in training—potentially overestimating spatial generalization, especially for later campaigns with $R^2 > 0.97$ [45]. Spatial-block CV withholding entire wells [46] would give more conservative estimates: temporally interpolated virtual samples from the same well allow near-exact reconstruction of a single withheld observation, directly producing $R^2 > 0.99$. We did not rerun models under leave-one-well-out or block schemes, since the leakage-free April 2024 upper-layer campaign already served this purpose—as the only campaign without within-well information from other campaigns, its $R^2 \approx 0.67$ provides the conservative baseline against which the optimistic later-campaign values should be judged. A dedicated leave-one-well-out validation remains worthwhile for future work.

Fourth, the temporal augmentation strategy assumes smooth inter-campaign variation in $\text{NH}_4^+\text{-N}$, which may not hold during rapid hydrological changes, such as the April–August recharge events; interpolated samples from this period should be interpreted with caution.

Fifth, hydrogeological covariates (e.g., hydraulic conductivity, aquifer thickness, depth to water table) were unavailable and could improve accuracy, especially for April 2024 ($R^2 \approx 0.67$); applying similar interpretability analysis (Figure S12) with such covariates at other sites would help guide future data collection [47].

Sixth, this single-site study requires multi-site validation across varied contamination sources, geological settings, and monitoring densities; future work should extend the monitoring period, incorporate hydrogeological parameters, adopt spatial-block cross-validation, and validate across multiple sites.

Population was estimated from a 1 km gridded product (WorldPop Global2, UN-adjusted); township- or village-level census data would refine the exposed-population and water-demand figures. Site-specific epidemiological data were also unavailable, so linking high-risk zones to health outcomes via household surveys and health screening remains an important future direction.

5. Conclusions

This study systematically compared geostatistical, machine learning, and hybrid approaches for spatial prediction of $\text{NH}_4^+\text{-N}$ in groundwater beneath a chemical industrial park under weak spatial autocorrelation. $\text{NH}_4^+\text{-N}$ was extremely heterogeneous (CV 3.7–7.9; 0.025–3682 mg/L), with 62.4% of upper-layer wells exceeding the Class III standard in April 2024. Spatial autocorrelation was weak throughout (Moran's $I = 0.040\text{--}0.118$; nugget ratios 75.1–104.8%), reflecting point-source contamination. Consequently, on the raw scale, all six geostatistical methods degraded toward the global mean (near-zero or negative R^2). On the $\log_{10}$ scale, XGBoost was the best ML model ($R^2 \approx 0.75$), and hybrid ML–kriging produced spatially coherent maps while retaining comparable skill; the leakage-free April value ($R^2 \approx 0.67$) represents spatial generalization accuracy, whereas later-campaign values are optimistic, and because the two families use different scales and validation schemes, the comparison is directional. Exceedance probability mapping (AUC = 0.959) identified a persistent high-risk zone in the central–southeastern sector. In practice and for society, these results provide environmental managers with directly actionable, threshold-based risk information, as exceedance probability maps are interpretable to non-technical stakeholders, can pinpoint where continuous source control should be prioritized to protect nearby communities relying on groundwater, and support cost-effective allocation of monitoring and remediation resources. More broadly, the “diagnose spatial autocorrelation first, then select methods” workflow transfers to other point-source pollutants and contaminated-site programs.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/w18141761/s1>, Supplementary Methods S1: Temporal Data Augmentation and Feature Engineering; Supplementary Methods S2: LOO-CV Outlier Filtering and Classification Details; Figure S1: LISA cluster maps; Figure S2: Variogram fitting; Figure S3: ROC curves; Figure S4: Temporal boxplots; Figure S5: Moran scatter plots; Figures S6 and S7: Geostatistical interpolation maps (other campaigns); Figures S8 and S9: LOO-CV scatter plots (other campaigns); Figure S10: Spatial prediction maps (other campaigns); Figure S11: Overall research framework and workflow of this study; Figure S12: Feature importance and SHAP analysis; Table S1: Normality test results for $\text{NH}_4^+\text{-N}$ concentrations in the upper aquifer; Table S2: Hyperparameters of the six ML regression models evaluated in Section 2.5; Table S3: Descriptive statistics of $\text{NH}_4^+\text{-N}$ concentrations (mg/L) in the upper aquifer; Table S4: Variogram fitting parameters for all three

theoretical models across four campaigns; Table S5: LOO-CV R^2 for six geostatistical methods across all campaign–layer combinations.

Author Contributions: B.L.: Writing—original draft preparation, Methodology, Investigation, and Data curation. Q.W.: Writing—original draft preparation, Methodology, Software, and Formal analysis. R.L.: Validation and Visualization. Y.H.: Validation and Visualization. H.L.: Writing—review and editing, Supervision, and Funding acquisition. Y.Y.: Writing—review and editing, Conceptualization, and Project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Program of Shanxi Province (grant number 202303021221057). The authors thank the park operator for facilitating the field sampling campaigns.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors due to confidentiality requirements of the industrial park.

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

Abbreviations

NH₄⁺-N, Ammonium Nitrogen; ML, Machine Learning; XGB, XGBoost (Extreme Gradient Boosting); OK, Ordinary Kriging; SK, Simple Kriging; CK, Co-Kriging; IK, Indicator Kriging; EBK, Empirical Bayesian Kriging; IDW, Inverse Distance Weighting; RF, Random Forest; GBR, Gradient Boosting Regression; SVR, Support Vector Regression; KNN, K-Nearest Neighbor; ADB, AdaBoost; RK, Regression Kriging; ARK, Adaptive Regression Kriging; LOO-CV, Leave-One-Out Cross-Validation; RMSE, Root Mean Square Error; MAE, Mean Absolute Error; AUC, Area Under the ROC Curve; ROC, Receiver Operating Characteristic; SMOTE, Synthetic Minority Oversampling Technique; LISA, Local Indicators of Spatial Association; CV, Coefficient of Variation.

References

- Li, P.; He, X.; Guo, W. Spatial groundwater quality and potential health risks due to nitrate ingestion through drinking water: A case study in Yan'an City on the Loess Plateau of northwest China. *Hum. Ecol. Risk Assess. Int. J.* **2019**, *25*, 11–31. [\[CrossRef\]](#)
- Han, D.; Currell, M.J.; Cao, G. Deep challenges for China's war on water pollution. *Environ. Pollut.* **2016**, *218*, 1222–1233. [\[CrossRef\]](#) [\[PubMed\]](#)
- Liu, J.; Yuan, J.; Zhang, Y.; Zhang, H.; Luo, Y.; Su, Y. Identification of ammonium source for groundwater in the piedmont zone with strong runoff of the Hohhot Basin based on nitrogen isotope. *Sci. Total Environ.* **2023**, *882*, 163650. [\[CrossRef\]](#) [\[PubMed\]](#)
- Li, P.; Tian, R.; Xue, C.; Wu, J. Progress, opportunities, and key fields for groundwater quality research under the impacts of human activities in China with a special focus on western China. *Environ. Sci. Pollut. Res.* **2017**, *24*, 13224–13234. [\[CrossRef\]](#) [\[PubMed\]](#)
- Webster, R.; Oliver, M.A. *Geostatistics for Environmental Scientists*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
- David, M. *Geostatistical Ore Reserve Estimation*; Elsevier: Amsterdam, The Netherlands, 1977.
- Cressie, N. *Statistics for Spatial Data*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
- Goovaerts, P. *Geostatistics for Natural Resources Evaluation*; Oxford University Press: Oxford, UK, 1997.
- Chiles, J.-P.; Delfiner, P. *Geostatistics: Modeling Spatial Uncertainty*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
- Adhikary, P.P.; Dash, C.J. Comparison of deterministic and stochastic methods to predict spatial variation of groundwater depth. *Appl. Water Sci.* **2017**, *7*, 339–348. [\[CrossRef\]](#)
- Pilz, J.; Spöck, G. Why do we need and how should we implement Bayesian kriging methods. *Stoch. Environ. Res. Risk Assess.* **2008**, *22*, 621–632. [\[CrossRef\]](#)
- Belkhir, L.; Tiri, A.; Mouni, L. Spatial distribution of the groundwater quality using kriging and Co-kriging interpolations. *Groundw. Sustain. Dev.* **2020**, *11*, 100473. [\[CrossRef\]](#)
- Legendre, P. Spatial autocorrelation: Trouble or new paradigm? *Ecology* **1993**, *74*, 1659–1673. [\[CrossRef\]](#)
- Tesoriero, A.J.; Voss, F.D. Predicting the probability of elevated nitrate concentrations in the Puget Sound Basin: Implications for aquifer susceptibility and vulnerability. *Groundwater* **1997**, *35*, 1029–1039. [\[CrossRef\]](#)
- Cambardella, C.A.; Moorman, T.B.; Novak, J.; Parkin, T.; Karlen, D.; Turco, R.; Konopka, A. Field-Scale variability of soil properties in central Iowa soils. *Soil Sci. Soc. Am. J.* **1994**, *58*, 1501–1511. [\[CrossRef\]](#)

16. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
17. 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; pp. 785–794.
18. Knoll, L.; Breuer, L.; Bach, M. Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. *Sci. Total Environ.* **2019**, *668*, 1317–1327. [[CrossRef](#)] [[PubMed](#)]
19. Ransom, K.M.; Nolan, B.T.; Stackelberg, P.; 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. [[CrossRef](#)] [[PubMed](#)]
20. Naghibi, S.A.; Pourghasemi, H.R.; Dixon, B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ. Monit. Assess.* **2016**, *188*, 44. [[CrossRef](#)] [[PubMed](#)]
21. Rodríguez-Galiano, V.; Mendes, M.P.; García-Soldado, M.J.; Chica-Olmo, M.; Ribeiro, L. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Sci. Total Environ.* **2014**, *476*, 189–206. [[CrossRef](#)] [[PubMed](#)]
22. Nolan, B.T.; Fienen, M.N.; Lorenz, D.L. A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. *J. Hydrol.* **2015**, *531*, 902–911. [[CrossRef](#)]
23. Anning, D.W.; Paul, A.P.; McKinney, T.S.; Huntington, J.M.; Bexfield, L.M.; Thiros, S.A. *Predicted Nitrate and Arsenic Concentrations in Basin-Fill Aquifers of the Southwestern United States*; Scientific Investigations Report 2012-5065; US Geological Survey: Reston, VA, USA, 2012.
24. Liang, Y.; Zhang, X.; Sun, Y.; Yao, L.; Gan, L.; Wu, J.; Chen, S.; Li, J.; Wang, J. Enhanced groundwater vulnerability assessment to nitrate contamination in Chongqing, Southwest China: Integrating novel explainable machine learning algorithms with DRASTIC-LU. *Hydrol. Res.* **2024**, *55*, 663–682. [[CrossRef](#)]
25. Hengl, T.; Heuvelink, G.B.; Rossiter, D.G. About regression-kriging: From equations to case studies. *Comput. Geosci.* **2007**, *33*, 1301–1315. [[CrossRef](#)]
26. Podgorski, J.; Berg, M. Global analysis and prediction of fluoride in groundwater. *Nat. Commun.* **2022**, *13*, 4232. [[CrossRef](#)] [[PubMed](#)]
27. Skala, J.; Žižala, D.; Minařík, R. Machine learning for predictive mapping of exceedance probabilities for potentially toxic elements in Czech farmland. *J. Environ. Manag.* **2025**, *380*, 125035. [[CrossRef](#)] [[PubMed](#)]
28. Standardization Administration of China. *GB/T 14848-2017*; Standard for Groundwater Quality. Standards Press of China: Beijing, China, 2017.
29. 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. [[CrossRef](#)] [[PubMed](#)]
30. Gómez-Escalonilla, V.; Montero-González, E.; Díaz-Alcaide, S.; Martín-Loeches, M.; del Rosario, M.R.; Martínez-Santos, P. A machine learning approach to site groundwater contamination monitoring wells. *Appl. Water Sci.* **2024**, *14*, 250. [[CrossRef](#)]
31. *HJ 164-2020*; Technical Specifications for Environmental Monitoring of Groundwater. China Environment Publishing Group: Beijing, China, 2020.
32. *HJ 535-2009*; Water Quality—Determination of Ammonia Nitrogen—Nessler’s Reagent Spectrophotometry. Ministry of Environmental Protection of the People’s Republic of China: Beijing, China, 2009.
33. Moran, P.A. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23. [[CrossRef](#)]
34. Dale, M.R.; Fortin, M.-J. *Spatial Analysis: A Guide for Ecologists*; Cambridge University Press: Cambridge, UK, 2014.
35. Anselin, L. Local indicators of spatial association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115. [[CrossRef](#)]
36. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
37. Han, H.; Suh, J. Spatial prediction of soil contaminants using a hybrid random forest–ordinary kriging model. *Appl. Sci.* **2024**, *14*, 1666. [[CrossRef](#)]
38. Valerio, A.; Chen, Y.-C.; Liu, C.-Y.; Chen, Y.-Y.; Lin, C.-Y. A Hybrid Regression–Kriging–Machine Learning Framework for Imputing Missing TROPOMI NO₂ Data over Taiwan. *Remote Sens.* **2025**, *17*, 2084. [[CrossRef](#)]
39. 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. [[CrossRef](#)]
40. Zhou, S.; Xiong, H.; Song, X.; He, Z.; Peng, Q.; Guo, X.; Xiong, R.; Yang, S.; Ma, C.; Wang, Y. Spatial prediction of groundwater nitrate by explainable machine learning models and remote sensing indicators under conditions of imbalanced datasets. *Geocarto Int.* **2025**, *40*, 2557953. [[CrossRef](#)]
41. Chen, S.; Wu, W.; Hu, K.; Li, W. The effects of land use change and irrigation water resource on nitrate contamination in shallow groundwater at county scale. *Ecol. Complex.* **2010**, *7*, 131–138. [[CrossRef](#)]
42. Sahour, H.; Gholami, V.; Vazifedan, M. A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. *J. Hydrol.* **2020**, *591*, 125321. [[CrossRef](#)]

43. Zhang, T.; Wu, J.; Chu, H.; Liu, J.; Wang, G. Interpretable machine learning based quantification of the impact of water quality indicators on groundwater under multiple pollution sources. *Water* **2025**, *17*, 905. [[CrossRef](#)]
44. Chen, K.; Liu, Q.; Yang, T.; Ju, Q.; Zhu, M. Risk assessment of nitrate groundwater contamination using GIS-based machine learning methods: A case study in the northern Anhui plain, China. *J. Contam. Hydrol.* **2024**, *261*, 104300. [[CrossRef](#)] [[PubMed](#)]
45. Koldasbayeva, D.; Tregubova, P.; Gasanov, M.; Zaytsev, A.; Petrovskaia, A.; Burnaev, E. Challenges in data-driven geospatial modeling for environmental research and practice. *Nat. Commun.* **2024**, *15*, 10700. [[CrossRef](#)] [[PubMed](#)]
46. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [[CrossRef](#)]
47. Ma, Q.; Ge, W.; Arab, A.; Butscher, C.; Taherdangkoo, R. Explainable machine learning with ensemble-based uncertainty quantification for groundwater quality indices in the Ganfu Plain, China. *Environ. Earth Sci.* **2026**, *85*, 196. [[CrossRef](#)]

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.