Environmental Science Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5va00368g

Evaluating groundwater quality in an arsenic-contaminated aquifer in the Red River Delta using machine learning: a case study in Van Phuc, Hanoi, Vietnam

Thi Duyen Vu, [ab] Thanh Dam Nguyen, [c] Thi Kim Trang Pham, [d] Michael Berg [e] and Hung Viet Pham *[a]

Groundwater quality in rapidly urbanising megacities such as Hanoi is increasingly threatened by over-extraction and widespread contamination. Despite heavy reliance on groundwater as the primary water source, its quality has rarely been assessed comprehensively and objectively. This study proposes a machine learning – based approach for evaluating groundwater quality in Van Phuc, where aquifers are affected by intensive exploitation and arsenic pollution. The Extreme Gradient Boosting (XGBoost) algorithm was employed to rank and select parameters according to their importance to overall water quality. Among the eleven input indicators, eight parameters, including As, total hardness, Mn, Na, Cl^-, NH_4^+, Fe, and F^-, showed substantial contributions, with As identified as the most influential variable, whereas pH, SO_4^{2-}, and total dissolved solids (TDS) contributed negligibly. Four aggregation functions were employed to compute the overall groundwater quality index (GWQI), and the National Sanitation Foundation (NSF) model yielded the most consistent and reliable classification for the study area. Application of the developed framework indicated that only one sample exhibited good water quality, while the remainder fell into fair (41.4%), marginal (20.7%), and poor (34.5%) categories. Spatial water quality patterns were closely aligned with hydrogeochemical zonation: poor-to-marginal conditions predominated within the Holocene aquifer, improving with depth across the redox transition zone, and generally achieving better quality in the Pleistocene aquifer. The proposed approach provides a transparent and transferable tool for groundwater quality assessment in stressed urban aquifers, reducing subjectivity while enhancing interpretability and supporting evidence-based water management.

Received 17th October 2025
Accepted 6th December 2025

DOI: 10.1039/d5va00368g

rsc.li/esadvances

Environmental significance

Groundwater in rapidly urbanising regions such as Hanoi is increasingly threatened by overextraction and arsenic contamination, posing long-term risks to public health and sustainable water management. Conventional groundwater quality indices often rely on subjective expert judgment and fail to capture nonlinear interactions among hydrochemical parameters. This study introduces a transparent and transferable machine learning – based framework using the XGBoost algorithm to objectively evaluate groundwater quality in arsenic-affected aquifers. The model identifies key parameters controlling contamination and delineates spatial patterns consistent with hydrogeochemical processes. The findings provide an interpretable, data-driven tool that enhances groundwater quality assessment, supports evidence-based management, and contributes to the broader goal of safeguarding urban water resources under environmental stress.

1. Introduction

Groundwater is a critical freshwater source for domestic, industrial, and agricultural activities, especially in rapidly urbanising regions. Globally, about 50% of the urban population depends on groundwater as the primary water supply.[1] In Vietnam, millions of people across the Red River Delta rely on it for daily consumption and economic activities.[2] However, aquifer systems worldwide, including those in Vietnam, are facing significant depletion due to excessive pumping and rapid urban expansion.[3] Overextraction reduces aquifer storage,

[a]Key Laboratory of Analytical Technology for Environmental Quality and Food Safety Control, VNU University of Science, Vietnam National University, Hanoi, 334 Nguyen Trai, Thanh Xuan Ward, Hanoi 100000, Vietnam. E-mail: vietph@vnu.edu.vn
[b]Graduate University of Science and Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Nghia Do Ward, Hanoi 100000, Vietnam
[c]Faculty of Chemistry, VNU University of Science, Vietnam National University, Hanoi, 334 Nguyen Trai, Thanh Xuan Ward, Hanoi 100000, Vietnam
[d]Research Center for Environmental Technology and Sustainable Development, VNU University of Science, Vietnam National University, Hanoi, 334 Nguyen Trai, Thanh Xuan Ward, Hanoi 100000, Vietnam
[e]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water, 8600 Dübendorf, Switzerland

changes hydraulic gradients, and enhances the risk of surface water intrusion, which collectively alter redox conditions, promote contaminant mobilisation, and degrade groundwater quality.[3–5]

Hanoi, a megacity of nearly 8.7 million inhabitants in 2024, relies heavily on deep aquifers. In 2017, exploited groundwater from these aquifers reached 1 million m^3 day^{-1}, supplying roughly 70% of the city's domestic demand.[6,7] Over a century of overextraction has resulted in significant drawdown, created cones of depression, and altered regional flow directions.[6,8] Concurrently, widespread contamination by arsenic, iron, manganese, and ammonium has also been reported.[2,9–12] The combination of depletion and contamination highlights an urgent need for quantitative, interpretable, and scalable groundwater quality assessment tools to support sustainable urban water management.

Assessing water quality, especially in complex hydrogeochemical regions, is challenging due to monitoring data's volume, variability, and heterogeneity.[13] The water quality index (WQI), initially developed in the 1960s, remains widely used to integrate multiple chemical indicators into a single score representing overall water quality status. Its simplicity and transparency make it effective for both scientific interpretation and public communication.[1,13,14] However, traditional WQI models rely heavily on expert judgment and locally defined weighting schemes, which introduce subjectivity, overlook nonlinear interactions among parameters, affecting the objectivity and reliability of the assessment and limit generalizability.[1,13–15]

Recent advances in data science, particularly machine learning (ML), provide promising alternatives to overcome the inherent limitations of conventional water quality assessment methods. ML-based frameworks can learn nonlinear relationships directly from data, improving predictive accuracy, enhancing objectivity, and reducing uncertainty in water quality assessment.[1,14,16–19] Numerous studies demonstrate that ensemble algorithms such as Random Forest, Gradient Boosting, Support Vector Machines, and Neural Networks outperform traditional regression and index-based methods for handling heterogeneous hydrochemical datasets.[20] These models not only provide predictive accuracy but also yield interpretable feature-importance metrics that help identify key hydrochemical drivers influencing water quality. Advances in data-driven modelling have also shown that machine learning and multivariate statistical approaches can capture complex, nonlinear relationships between water quality indicators in river and lake systems.[21–23] Although ML models have been widely used for water quality assessment, they also have certain weaknesses. They are often trained on relatively small and site-specific datasets, which makes them vulnerable to overfitting and limits their robustness and interpretability.[24,25] In addition, data-driven WQI frameworks may inherit substantial uncertainty from subjective or site-dependent choices such as indicator selection, sub-index functions, weighting schemes and aggregation rules, so quantifying and communicating model uncertainty is essential if such models are to be used in water-management decisions.[26]

Among these approaches, Extreme Gradient Boosting (XGBoost) has emerged as a robust and interpretable ensemble algorithms for environmental applications. Many studies have demonstrated that XGBoost is a boosting algorithm that provides high accuracy, the ability to model complex nonlinear relationships, avoid overfitting through regularisation, and robustness to incomplete datasets.[14,19,27–29] These features make it effective for complex hydrochemical datasets that are nonlinear, sparse, and influenced by coupled physical-chemical processes. Moreover, XGBoost offers feature-importance rankings enabling transparent interpretation of how each water-quality parameter influences the model output, an essential advantage for risk communication and evidence-based groundwater management.[1,19,28,29]

In Vietnam, the application of ML for water quality assessment has expanded rapidly in recent years to assess and forecast water quality in both surface and groundwater systems. For example, Khoi *et al.*,[30] evaluated twelve ML algorithms for predicting the WQI of the La Buong River (Southeast Vietnam), where XGBoost outperformed others. Le *et al.*,[31] demonstrated the utility of ML classifiers in capturing spatiotemporal variability in the Song Quao-Ca Giang water system, helping to establish a methodological foundation for subsequent WQI-focused studies. Lap *et al.*,[32] combined feature selection and ML in the An Kim Hai irrigation network (Hai Phong) to maintain high predictive accuracy while reducing data requirements. Nguyen *et al.*,[33] applied Bayesian model averaging with gradient-boosting regressors in Red River Delta irrigation systems, achieving $R^2 \approx 0.96$ with minimal input parameters. Additionally, Nguyen *et al.*,[34] used ML to predict groundwater quality in Quang Tri's coastal aquifer, confirming the model's relevance for subsurface applications. These studies highlight the growing maturity of ML-based WQI modelling in Vietnam; however, few have focused on Hanoi's deep aquifers, where chronic overextraction and arsenic contamination converge, leaving a crucial research and management gap.

Despite growing global adoption, most ML-WQI studies still focus on surface waters or require extensive datasets that are rarely available for groundwater systems. Few have targeted groundwater in Southeast Asian megacities, where aquifers face compounded threats from urbanisation and geogenic contamination. Addressing this gap is crucial to advancing data-driven groundwater management and understanding the nonlinear interplay between hydrochemical parameters in such vulnerable aquifer systems.

In this study, we develop and apply an ML-based WQI framework to evaluate groundwater quality in Van Phuc, Thanh Tri District, Hanoi – a site representative of aquifers affected by massive abstraction and widely contaminated groundwater. The XGBoost algorithm is employed for its robustness, interpretability, and ability to model nonlinear hydrochemical interactions.[14,16,19,27–29] Specifically, we aim to (1) identify key hydrochemical indicators influencing groundwater quality through feature-importance analysis; (2) compute ML-based WQI to quantify groundwater (GWQI) in Van Phuc; and (3) explore spatial variation in groundwater quality in Van Phuc.

We hypothesise that ML-based feature selection will provide a more objective, transferable framework than conventional WQI approaches. The proposed method integrates environmental chemistry insight with data-driven modelling, offering a scalable, interpretable, and evidence-based tool to support sustainable groundwater management in Hanoi and comparable urban environments globally.

2. Materials and methods

2.1. Study site

The study site, Van Phuc village, is located on a meander of the Red River, about 15 km southeast of Hanoi and outside the river's dike system (Fig. 1). The area is underlain by complex quaternary sedimentary formations (50–90 m thick), comprising two aquifers: a reducing Holocene aquifer with peat-rich clay and sand layers, and a less reducing Pleistocene aquifer characterised by yellow-brown sand and gravel.[10,35–37]

Over the past century, intensive groundwater abstraction from deep aquifers in Hanoi has led to substantial drawdown, reversing flow direction in the Van Phuc village. Groundwater now flows from southeast to northwest toward Hanoi at a rate of 40 m year^{-1} and has been sustained over the last 50–60 years.[8,10,37] As a result, the Red River acts as the primary recharge source for both aquifers within a 5 km corridor, raising concern that contaminated water from the Holocene aquifers may migrate downward into the underlying, previously uncontaminated Pleistocene aquifer.[8,10] The geological and hydrological conditions and geographical location make this area ideal for studying groundwater quality and biogeochemical characteristics. Our research group has been investigating biogeochemical characteristics in the study area for an extended period. However, no studies have used machine learning to assess groundwater quality and derive the WQI.

At the scale of the entire Red River Delta, large-scale surveys and probabilistic mapping by Winkel *et al.*,[2] show that a substantial fraction of domestic tubewells abstract groundwater with As concentrations above the World Health Organisation (WHO) guideline value of 10 μg L^{-1}, with the highest probabilities occurring along the banks of the Red River and in parts of the southwestern delta. Van Phuc is situated within one of these arsenic-affected zones and is well known for its elevated arsenic concentration in groundwater.[2,10]

2.2. Sample collection and analysis

Groundwater from 29 tubewells with depths ranging from 20 to 54 meters below ground level (mgbl) was collected in April 2019, following procedures similar to those described by Stopelli *et al.*[37] Briefly, groundwater was collected after purging for at least 10 minutes to remove stagnant water, and the field multi-

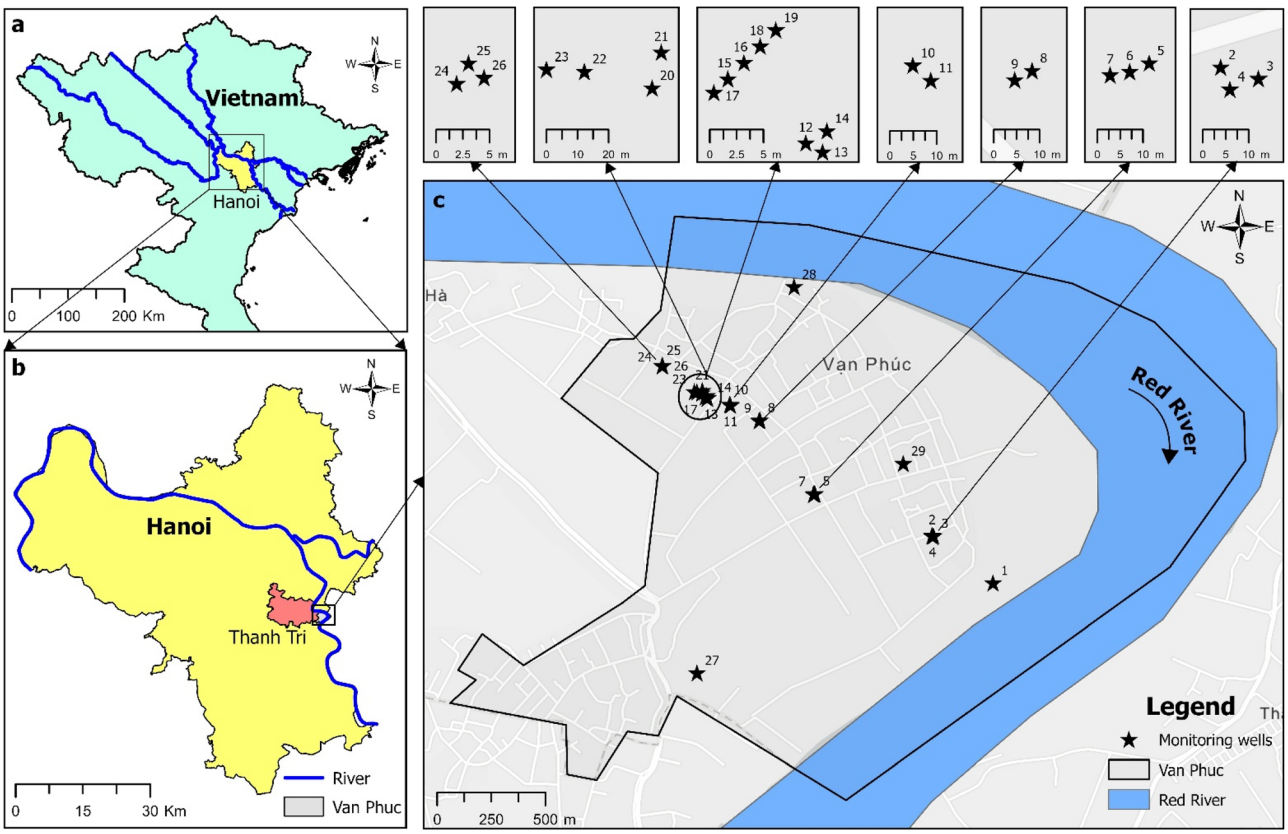

Fig. 1 (a) Regional map of the Hanoi area and the main rivers (blue lines) in northern Vietnam. (b) Location of Van Phuc village in Thanh Tri District relative to central Hanoi and the Red River. (c) Monitoring wells ($n = 29$) in Van Phuc, where groundwater samples were collected for this study.

meters HQ40d (Hach, USA) had stabilised for pH, dissolved oxygen (DO), electrical conductivity (EC), and redox potential (ORP). Pumped and filtered (through a 0.45 μm cellulose acetate membrane) groundwater was then collected into pre-washed polyethene bottles. Samples designated for cations, ammonium, and phosphate analysis were acidified to pH < 2 with concentrated HNO_3, while those for anions were left unacidified. All the samples were then stored with ice packs in the field and at 4 °C in the laboratory for later analysis.

The chemical composition of groundwater was analysed at the VNU Key Laboratory of Analytical Technology for Environmental Quality and Food Safety Control (KLATEFOS), VNU University of Science. Major cations (Na, K, Ca and Mg), as well as total iron (Fe) and manganese (Mn), were quantified using atomic absorption spectroscopy (AA-6800, Shimadzu, Japan). Total arsenic (As) concentration was determined by hydride generation atomic absorption spectrometry (HG-AAS) on the same instrument. Internal reference materials (ARS, certified by Eawag, Switzerland) were analysed concurrently for quality control. Ammonium (NH_4^+) and phosphate (PO_4^{3-}) ions were measured using a UV-Vis spectrophotometer (UV-1800, Shimadzu, Japan). Anions (F^-, Cl^-, SO_4^{2-}, NO_2^- and NO_3^-) were determined by ion-exchange chromatography (HIC-20A super, Shimadzu, Japan), employing the Shimadzu PIA reference material to ensure analytical reliability. All measurements were performed under calibration conditions with R^2 values above 0.995. The relative standard deviation (RSD) of triplicate analyses remained below 5%, and recovery rates for reference materials consistently ranged between 90% and 110%, confirming the precision and accuracy of the analysis.

2.3. GWQI model

2.3.1. Parameter selection. Among the 23 parameters analysed, 11 physico-chemical groundwater parameters, including pH, concentrations of total As, total Fe, Mn^{2+}, Na^+, NH_4^+, F^-, Cl^-, SO_4^{2-}, hardness (HN), and total dissolved solids (TDS), were selected as inputs to the ML models. These parameters were chosen because they are defined as parameters that need to be analysed and monitored when assessing groundwater quality according to the Vietnamese national technical regulation, as they are key hydrogeochemical factors in the aquifer systems of the Red River Delta and directly linked to human health risks (Table S1). From a hydrogeochemical perspective, As, Fe, Mn and NH_4^+ are directly linked to redox-driven mobilisation in the aquifer system;[2,10,12] whereas Na^+, Cl^- and SO_4^{2-} act as indicators of mixing with surface water or anthropogenic inputs.[38] Moreover, the hardness integrates Ca–Mg carbonate equilibria;[39] during that F^- and TDS capture salinity and overall mineralisation.[40] Additionally, parameters such as As, Fe, Mn and F^- are also recognised as health-relevant contaminants when present at elevated concentrations.[40–45] Since HN and TDS were absent from the original dataset, these values were calculated indirectly from the concentrations of Ca^{2+} and Mg^{2+} and EC[46] by using eqn (1) and (2).

$$\mathrm{HN} = M_{\mathrm{CaCO_3}} \times \left(\frac{C_{\mathrm{Ca^{2+}}}}{M_{\mathrm{Ca}}} + \frac{C_{\mathrm{Mg^{2+}}}}{M_{\mathrm{Mg}}} \right) \quad (1)$$

$$\mathrm{TDS} = 0.67 \times \mathrm{EC} \quad (2)$$

In which $M_{\mathrm{CaCO_3}}$, M_{Ca}, and M_{Mg} are molar masses of $CaCO_3$ (100.09 g mol^{-1}), Ca (40.08 g mol^{-1}), and Mg (24.31 g mol^{-1}) respectively; $C_{\mathrm{Ca^{2+}}}$ and $C_{\mathrm{Mg^{2+}}}$ are concentrations (mg L^{-1}) of Ca^{2+} and Mg^{2+} in groundwater.

2.3.2. GWQI computation. In this study, the groundwater quality index (GWQI) was calculated by adapting, with modifications, the approach introduced by Uddin *et al.*[14] This process encompassed three principal steps: (i) selection of indicators and computation of their weights; (ii) derivation of sub-index values for the indicators; and (iii) calculation of GWQI scores. Indicators were selected, and their weights were estimated using the development of a water-quality classification model utilising the XGBoost algorithm. The inputs to the model consisted of measured values of 11 above-mentioned physico-chemical groundwater parameters. The output predicted by the model was the pollution status (PS) for each sample. A PS value of 0 was assigned when all 11 parameters complied with the recommended guideline values, whereas a PS value of 1 was set when any parameter exceeded its accepted range.

Compared to previous publications, the most challenging aspect in developing the XGBoost model in this study is handling data imbalance. Among 29 Van Phuc groundwater samples, only one sample (VP25) was classified as PS = 0. This extreme imbalance between the PS = 0 and PS = 1 classes makes ML-based classification impractical. Additional PS = 0 samples from the study area could not be obtained, as the possibility of identifying new uncontaminated samples was low given the extensive contamination. Algorithmically generated techniques like SMOTE were also infeasible due to only a single minority class pattern. To mitigate this issue, seven PS = 0 groundwater samples from other Hanoi sites (same period sampling time) were added to the dataset to build the model. Consequently, the full dataset of groundwater quality used in this study includes 29 samples collected from monitoring wells in Van Phuc village in April 2019 and 7 additional tubewell samples from other sites in Hanoi that met all Vietnamese groundwater quality guidelines (Table S2). These additional samples were gathered during the same campaign, using identical sampling procedures, and were analysed with the same analytical methods as those collected from Van Phuc.

Consistent with the approaches of Uddin *et al.*,[14] and Wang *et al.*,[1] physicochemical parameters with vales below the method detection limits were assigned a value of 0 prior to model training. The complete dataset (n = 36, Table S2) was then normalised using min–max scaling to minimise the influence of differing measurement ranges (*e.g.*, high concentrations of TDS compared with trace-levels of As). The dataset was subsequently divided into training and testing datasets (80/20 split), with stratified 5-fold cross-validation applied under the constraint that each training and test set included at least two samples with PS = 0. Additionally, the single PS = 0 sample from the original Van Phuc dataset was consistently allocated to the test set to evaluate model discriminability directly. Three widely used ML algorithms, including decision tree (DT),

random forest (RF), and XGBoost, were compared to identify the most suitable approach for classifying groundwater quality in the study area. For initial screening purposes, each algorithm was executed with a simple hyperparameter optimisation (Table S3). Evaluation metrics included the area under the receiver-operating characteristic curve (AUC), logloss, accuracy, sensitivity, and specificity. As shown in Table S4, the DT model proved inadequate, with the lowest AUC (0.5) and the highest logloss (0.693). The RF algorithm performed robustly, achieving a lower prediction logloss (0.299) than XGBoost (0.351), indicating a strong probability calibration. However, for the specific purpose of identifying unpolluted samples without error, XGBoost was superior, achieving 100% accuracy, sensitivity, and specificity on the test set compared to slightly lower metrics for RF. Given the priority of avoiding false positives in an imbalanced groundwater quality dataset, combined with its high interpretability, XGBoost was finally selected as the optimal model. Therefore, XGBoost was selected as the most appropriate algorithm for computing the GWQI in the Van Phuc area, Hanoi.

Hyperparameter tuning for the XGBoost model was optimised *via* grid search (Table S5) across 24 200 combinations of learning rate (0.01–0.10), max_depth (1–10), subsample (0.50–0.95), colsample_bytree (0.50–0.95), and scale_pos_weight (0.25 and 1), min_child_weight (1), and gamma (0). The number of boosting rounds (n_{rounds}) was fixed at 1000, with early stopping implemented after 50 rounds. Unlike Uddin *et al.*,[14] model performance was evaluated using logloss and AUC, which are more suitable metrics for binary classification than root mean square error (RMSE).[47] Optimal hyperparameters were chosen to maximise AUC and minimise logloss on the test set, while ensuring correct classification of the original PS = 0 sample.

After hyperparameter optimisation, the model was retrained on the original 29-sample Van Phuc dataset, and feature importances for the 11 indicators were extracted. These indicators were ranked by importance, and their weights were calculated using the rank order centroid (ROC) method, as described by Uddin *et al*.[14] Sub-index values for indicators with non-zero weights were then derived using eqn (3)–(5) based on their condition as shown in Table 1.

$$s_{i,j} = 100 - 100 \times \frac{C_j}{\mathrm{STD}_{u,j} - \mathrm{STD}_{l,j}} \quad (3)$$

$$s_{i,j} = 100 \times \frac{C_j - \mathrm{STD}_{l,j}}{\mathrm{STD}_{u,j} - \mathrm{STD}_{l,j}} \quad (4)$$

$$s_{i,j} = 100 - 100 \times \frac{C_j - \mathrm{STD}_{l,j}}{\mathrm{STD}_{u,j} - \mathrm{STD}_{l,j}} \quad (5)$$

where $s_{i,j}$, C_j, $\mathrm{STD}_{l,j}$, and $\mathrm{STD}_{u,j}$ are the sub-index value, the measured concentration, the lower threshold value, and the upper threshold value of indicator j, respectively.

Finally, the indicator weights and sub-indices were aggregated to produce an overall GWQI score. Four functions, including the National Sanitation Foundation (NSF) index, weighted quadratic mean (WQM), Scottish Research Development Department (SRDD) index, and Wet Java (WJ) index, were applied for comparison, using eqn (6)–(9) where relevant, with NSF and WQM proposed by Uddin *et al.*,[14] as performing best in their context.

$$\mathrm{NFS} = \sum_{j=1}^{n} s_{i,j} \times w_j \quad (6)$$

$$\mathrm{WQM} = \sqrt{\sum_{j=1}^{n} w_j \left(s_{i,j}\right)^2} \quad (7)$$

$$\mathrm{SRDD} = \frac{1}{100}\left(\sum_{j=1}^{n} s_{i,j} \times w_j\right)^2 \quad (8)$$

$$\mathrm{WJ} = \prod_{j=1}^{n} \left(s_{i,j}\right)^{w_j} \quad (9)$$

where $s_{i,j}$; w_j; and n are the sub-index value of indicator j; the weight value of indicator j; and the number of indicators, respectively.

3. Results and discussion

3.1. Statistical analysis

A summary of the statistical analysis is presented in Table S1. Overall, the pH, TDS, and concentrations of Na, Cl^-, NO_2^-, NO_3^-, and SO_4^{2-} were consistently below Vietnamese groundwater quality limits, and nearly all samples also complied with thresholds for HN and F^-, with only two and one samples exceeding the permissible thresholds, respectively (Fig. 2).

In contrast, a significant proportion of the samples exhibited elevated levels of As, Fe, Mn, and NH_4^+, with exceedance rates of approximately 55%, 62%, 62%, and 79%, respectively, indicating widespread contamination (Fig. 2). These findings are consistent with earlier studies in Van Phuc[10,35,37] and with reports of similar contamination patterns elsewhere in the Red River Delta.[2,9,11,12]

Table 1 Piecewise function used to calculate sub-index values for groundwater quality indicators in Van Phuc

Indicator	Condition	Sub-index function
As, Fe, Mn^{2+}, Na^+, NH_4^+, F^-, Cl^-, SO_4^{2-}, TDS, HN		eqn (3)
pH	$5.8 \leq \mathrm{pH} < 6.7$	eqn (4)
	$6.7 \leq \mathrm{pH} \leq 7.6$	100
	$7.6 < \mathrm{pH} \leq 8.5$	eqn (5)

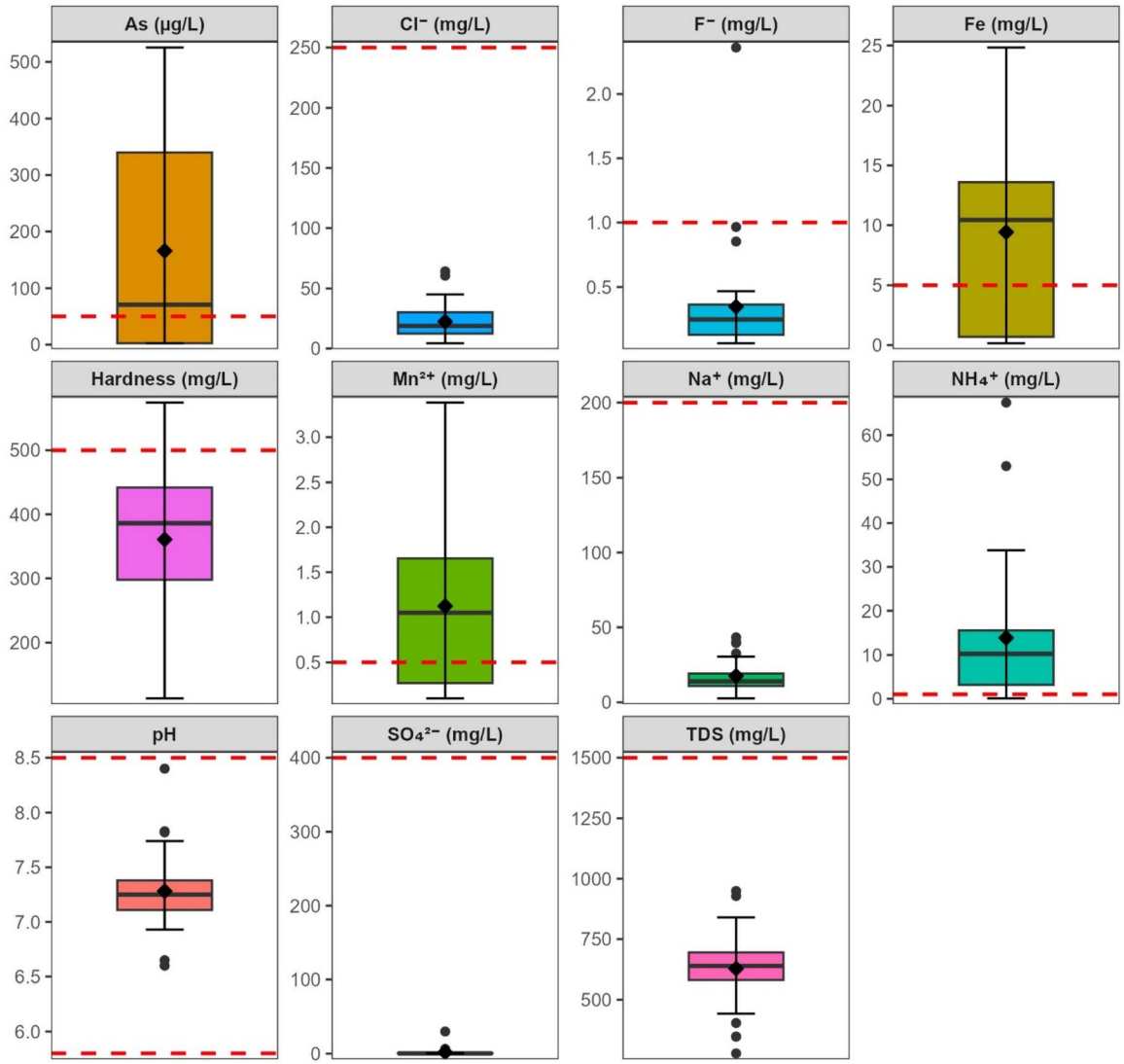

Fig. 2 Box plots illustrating the variability of 11 physico-chemical parameters measured in groundwater samples from 29 monitoring wells in Van Phuc. Boxes represent the interquartile range, the solid black line within each box is the median, the diamond symbol denotes the mean, and whiskers extend to the 5th and 95th percentiles, the black circle symbols beyond the whiskers denote outliers. The red dashed lines represent the Vietnamese guideline values for each parameter (two dashed lines in the pH plot indicate lower and upper guideline limits).

The widespread contamination observed in the study area, particularly the excessive As concentrations exceeding the World Health Organisation (WHO) and Vietnamese standard guideline value of 10 µg L^{-1} for drinking water, indicates that the groundwater is unsuitable for domestic use without prior treatment. Direct consumption of groundwater containing elevated levels of toxic metals such as As, Fe, and Mn poses serious health risks, including cancer and neurological disorders.[41–45]

3.2. XGBoost performance

The optimal hyperparameters were: $n_{\text{rounds}} = 376$, learning_rate = 0.05, max_depth = 5, subsample = 0.5, colsample_bytree = 0.5, scale_pos_weight = 1, gamma = 0, and min_child_weight = 1. Per-fold logloss and AUC values at these settings are detailed in Table S6. On average, AUC reached 1.0 for both training and testing sets, with logloss values of 0.204 and 0.257, respectively, demonstrating acceptable accuracy in predicting PS from the 11 parameters.

The model was retrained on the original 29-sample dataset to extract feature importances, which determined indicator ranks and weights. Of the 11 parameters, eight exhibited non-zero importance (Table 2), encompassing all parameters that breach guideline values in at least one sample (As, Mn, Fe, NH_4^+, F^-, and HN), whereas pH, SO_4^{2-}, and TDS contributed negligibly. Among the eight retained indicators, As was associated with the most frequent exceedances and emerged as the most influential variable. Notably, NH_4^+ displayed relatively low importance despite its prominence as a pollutant in the study area; in contrast, a comparable model by Wang *et al.*,[1] identified NH_4^+ as the primary predictor, highlighting potential dataset-

Table 2 Feature-importance ranking and ROC-derived weights of groundwater quality indicators based on the XGBoost model

Feature	Importance	Rank	ROC weight
As	0.3994	1	0.3397
HN	0.1649	2	0.2147
Mn^{2+}	0.1153	3	0.1522
Na^{+}	0.1052	4	0.1106
Cl^{-}	0.0840	5	0.0793
NH_4^{+}	0.0571	6	0.0543
Fe	0.0452	7	0.0335
F^{-}	0.0290	8	0.0156
pH	0	—	0
SO_4^{2-}	0	—	0
TDS	0	—	0

specific variations. Similarly, Uddin *et al.*,[14] reported dissolved oxygen (DOX, in summer) and molybdate reactive phosphorus (MRP, in winter) as key pollutants that were not influential in their seasonal models. The weights computed from these importance ranks are summarised in Table 2.

Sub-index values (s_i) were calculated for the eight indicators, where $s_i = 0$ represents the poorest quality and $s_i = 100$ represents the highest quality. The four most frequently non-compliant parameters (As, Fe, Mn, NH_4^{+}) showed s_i ranges of 0–100, with mean values of 39.7 for As and 11.2 for NH_4^{+}. For F^{-} and HN, s_i ranged from 0–93.0 and 0–77.4, respectively, while the remaining two indicators had s_i values between 74.4 and 98.7.

Composite indices were computed using the NSF, WQM, SRDD, and WJ indices. While NSF, SRDD, and WJ are popular established approaches, WQM was recently introduced by Uddin *et al.*[14] The WJ index was unsuitable for the Van Phuc groundwater, resulting in zero scores for all samples except sample VP25. For the other indices, GWQI score ranges were NSF 16–83, WQM 36–86, and SRDD 3–69, with respective means of 44, 59, and 23 (Fig. 3). The maximum scores across all three indices were achieved for the single PS = 0 sample (83 for NSF, 86 for WQM, and 69 for SRDD). In contrast, the minima were associated with sample VP29, which exhibited the highest number of exceedances (5 out of 11): 16 for NSF, 36 for WQM, and 3 for SRDD.

Spearman's rank correlations revealed strong agreement among the three methods ($r = 0.972$–0.998, $p < 0.001$). However, index scores varied significantly according to Friedman's two-way analysis ($Q_{2,29} = 58.0$, $p < 0.001$): WQM generally produced higher scores than NSF ($Z_{\mathrm{NSF-WQM}} = -3.808$, $p < 0.001$), while SRDD yielded the lowest scores ($Z_{\mathrm{SRDD-NSF}} = 3.808$, $p < 0.001$; according to Wilcoxon signed-rank post-hoc with Bonferroni correction).

The most appropriate index among NSF, WQM, and SRDD was selected by evaluating the frequency of over- and under-estimates relative to compliance-based quality classes. Water-quality status was classified as: good (no exceedances; WQI 80–100), fair (1–2 exceedances; WQI 50–79), marginal (3 exceedances; WQI 30–49), and poor (≥4 exceedances; WQI 0–29).[14] WQM often overestimated (14 of 15), whereas SRDD

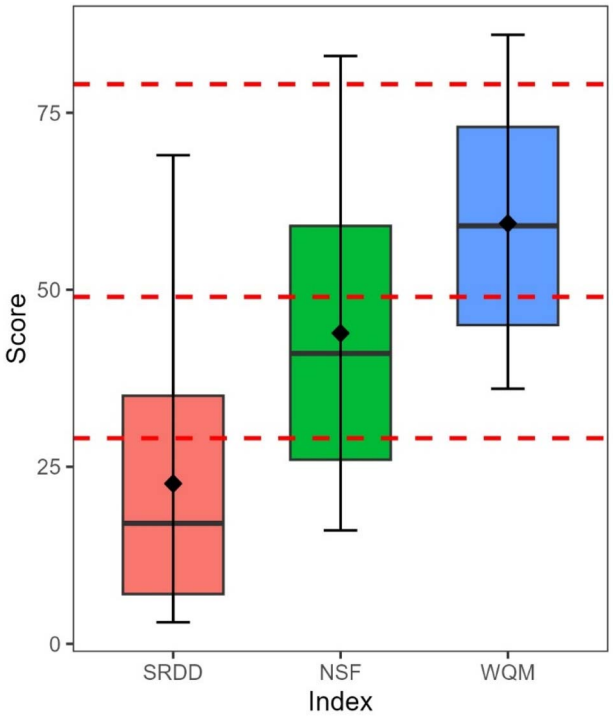

Fig. 3 Box plots summarise GWQI scores calculated using three different WQI aggregation models (NSF, SRDD and WJ) for the groundwater samples from Van Phuc. The diamond symbol indicates the mean value, and the solid black line in each box indicates the median GWQI values. The red dashed lines indicate the values used to classify water quality into four categories: good 80–100, fair 50–79, marginal 30–49 and poor 0–29. The diamond symbol indicates the mean value, and the solid black line in each box indicates the median GWQI values.

frequently underestimated (18 of 19). NSF generated the fewest misclassifications ($n = 5$; 2 overestimates, 3 underestimates) and was thus considered the most suitable for the study area. This selection contrasts with Uddin *et al.*,[14] where WQM was favoured, likely due to differences between coastal waters and groundwater systems. As elaborated earlier, several parameters exceeding guideline values were deemed non-influential in Uddin *et al.*[14]

Of the 29 samples in Van Phuc, only one was classified as good (GWQI = 83), twelve (41.4%) as fair, and the remaining 55.2% as marginal and poor. These results underscore substantial groundwater contamination in Van Phuc, emphasising the necessity for enhanced monitoring and management strategies to safeguard public health.

From a hydrogeochemical perspective, the feature-importance ranking obtained from XGBoost is consistent with the current understanding of arsenic mobilisation in the study area. The dominating role of As as the top-ranked feature is consistent with its high exceeding frequency and toxicity. Arsenic was released from sediment into groundwater through the reductive dissolution of Fe oxyhydroxides in the aquifer systems, supported by reducing bacteria, when the natural organic matter (NOM) is sufficiently available.[10,12,37] The second-

ranked indicator, hardness, acts as an integrated proxy for Ca–Mg carbonate equilibria and mixing between river-derived recharge and deeper Pleistocene groundwater, both of which control contaminant transport pathways.[39] High importance of Mn^{2+} reflects the progression of redox reactions towards more reducing conditions that may lead to the further mobilisation of As.[37] Na^+ and Cl^- likely capture mixing between young river-derived recharge and more mineralised, partly saline or anthropogenically impacted groundwater.[38,48] Although NH_4^+ is a widespread pollutant in the study area, its moderate importance suggests that it indirectly reflects reducing conditions, which is coherent with the release of As in the Holocene aquifer as mentioned above. The minor contributions of Fe and F^- further support the role of redox-controlled mobilisation of iron oxyhydroxides and water–sediment interactions in shaping groundwater quality in Van Phuc. The negligible contributions of pH, SO_4^{2-} and TDS in the model reflect their relatively limited variability within the dataset and their compliance with the guideline ranges.

Although a novel and reliable approach for assessing groundwater quality at the Van Phuc site in Hanoi has been developed, several limitations remain that warrant attention and may be addressable in future work to broaden applicability to other regions in Vietnam and globally. First, the XGBoost model was built on a relatively small and imbalanced dataset. While supplementing the dataset with additional samples from hydrogeochemical similar sites elsewhere in Hanoi to increase the minority class (PS = 0) partly mitigates this limitation, this practical approach is inconvenient and may prove challenging to implement when applied to datasets from other regions. It should be emphasised that Vietnam's national technical standards for groundwater are stringent, and obtaining uncontaminated samples, *i.e.*, samples meeting all specified technical criteria, is particularly challenging in heavily stressed aquifer systems such as Hanoi, where both geochemical processes and intensive anthropogenic exploitation impact groundwater. Among over 100 samples collected from other locations in Hanoi during the same sampling campaign at Van Phuc, only seven samples could be classified as PS = 0 for incorporation into the dataset. Although such compliant samples may be more readily available in less impacted regions beyond Hanoi, incorporating samples from geographically distant sites may introduce regional bias. Alternatively, data augmentation algorithms can be employed to generate minority-class samples artificially; however, this strategy requires sufficiently large original datasets to ensure the generation of realistic synthetic data. A more promising approach would involve shifting from binary classification (PS = 0 *versus* PS = 1) to multiclass classification, wherein PS = 0 represents samples meeting all national standards, PS = 1 represents samples with one or two parameter exceedances, PS = 2 represents samples with three exceedances, and PS = 3 represents samples with four or more exceedances. Multiclass classification would reduce the severity of dataset imbalance and would eliminate the need for augmented PS = 0 samples.

A second limitation is that this assessment is based on a single sampling campaign conducted in April 2019 and does not account for seasonal or inter-annual variability in groundwater quality. Temporal fluctuations may alter redox conditions and contaminant concentrations, potentially affecting the classification of indicators. Future studies could therefore incorporate time-series monitoring to quantify seasonal variability in the GWQI and to better characterise the robustness of the framework across different hydrogeochemical conditions.

3.3. Spatial distribution of groundwater quality

In the study area, most monitoring wells were deliberately installed along the dominant groundwater flow direction.[8,10,35,37] Based on the hydro(geo)chemical zonation proposed by Stopelli *et al.*,[37] these wells can be grouped into four zones: Zone B (wells VP1–VP4 and VP29) representing the Holocene aquifer near the river; Zone C (wells VP5–VP10) within the Holocene aquifer; Zone D (well VP12–VP21) corresponding to the redox transition zone (RTZ); and Zone E (well VP22–VP26) in the Pleistocene aquifer (Fig. 4). GWQI scores revealed apparent spatial variability consistent with this zonation, reflecting the complex hydro(geo)chemistry of the Holocene–Pleistocene system.

Zone B showed consistently poor groundwater quality due to the reductive dissolution of Fe-bearing minerals, which simultaneously mobilise arsenic.[37] Furthermore, recharge from surface water and the Red River water may introduce additional organic matter, enhancing the aquifer's reducing conditions and promoting further As release. Consequently, groundwater

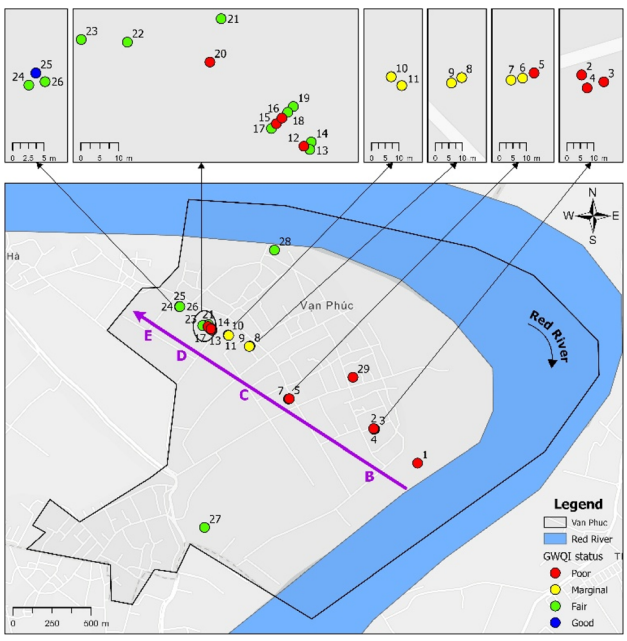

Fig. 4 Spatial distribution of groundwater quality along the Van Phuc transect based on NSF model. Each symbol represents a monitoring well positioned according to its location along the transect, colours indicate groundwater quality classes. The violet arrow represents the dominant groundwater flow direction, along which most monitoring wells are located. The letters B–E show the major hydro(geo)chemical zones as proposed by Stopelli *et al.*,[37] (2020): B = Holocene aquifer near the river; C = Holocene aquifers; D = redox transition zone (RTZ); and E = Pleistocene aquifers.

in Zone B remains of poor quality and effectively acts as a pollution corridor, facilitating the migration of arsenic from the river into deeper sections of the Holocene aquifer.

Zone C was the most severely impacted area[37] with strong Fe(III) reduction and methanogenic conditions, causing maximum of As, Fe, and NH_4^+ concentrations. Despite this, Mn remained low (0–0.27 mg L^{-1}) in most wells, suggesting secondary precipitation, which moderated water quality to marginal in most wells rather than poor as zone B.

Within Zone D, the Redox transition zone (RTZ), groundwater quality improves markedly with depth. Shallow wells (<30 m) exhibited poor WQI values, while deeper wells (>30 m) achieved fair quality. Along the flow path, As- and Fe-rich water from Zone C attenuated through coprecipitation with Fe oxyhydroxides, whereas Mn minerals were reduced, increasing dissolved Mn^{2+} concentrations.[37] The pronounced reduction in As and Fe in deeper RTZ wells significantly improved GWQI values. This finding underscores the RTZ's dual role as a hydro(geo) chemical boundary and a reactive "filter" attenuating contaminant transport and improving groundwater quality.

Groundwater quality in Zone E (Pleistocene aquifer) was generally better than in the Holocene system. Most wells displayed fair GWQI values; the only well achieving "good" status (at 38 m depth) was also located here. Stable Fe(III) oxyhydroxides in this zone[35,37] serve as an efficient arsenic sink, maintaining low As concentrations and stable water quality. Nonetheless, moderate dissolved Mn and NH_4^+ levels indicate ongoing redox processes, which may influence long-term As dynamics, particularly under intensive groundwater abstraction.

Overall, the spatial distribution of GWQI at Van Phuc reflects the coupled processes of arsenic mobilisation and attenuation within the Holocene–Pleistocene aquifer system. Zones B and C function as As source areas with poor to marginal GWQI, Zone D acts as a transition zone where contaminant attenuation enhances water quality, and Zone E represents a relatively stable aquifer with generally better GWQI. These results highlight the necessity of integrating groundwater quality assessment with stratigraphic architecture and hydro(geo)chemical processes. Furthermore, they emphasise the importance of groundwater treatment prior to use and the urgent need for sustainable management policies, as excessive pumping of deeper aquifers could exacerbate downward contaminant migration from the Holocene into the Pleistocene aquifers.

4. Conclusions

4.1. GWQI modelling

This study developed an ML-based WQI using an XGBoost algorithm to assess groundwater quality in an aquifer affected by over-exploitation and arsenic contamination. The results showed that among the 11 input parameters, eight parameters contributed significantly to groundwater quality in the study area, with As identified as the most influential indicator. The remaining three parameters, pH, SO_4^{2-} and TDS show negligible contributions. Among the four aggregation functions evaluated, the NSF model was the most suitable for the study area, as it produced the fewest misclassifications.

Applying the developed approach to 29 wells in Van Phuc, the computed overall GWQI indicated that only one sample was classified as "good". In contrast, the remaining samples were categorized as "fair" (41.4%), "marginal" (20.7%) and "poor" (34.5%). The spatial distribution of groundwater quality strongly correlated with hydrogeochemical zonation: poor to marginal water quality predominantly occurred within the Holocene aquifer, improving with depth across the redox transition zone, and generally exhibiting better quality within the Pleistocene aquifer.

These findings highlight that the proposed approach effectively reduces the subjectivity inherent in traditional methods while providing an interpretable and management-oriented framework for evaluating groundwater quality in stressed aquifer systems. Nevertheless, the present study is constrained by a limited and imbalanced dataset, as it relies on a single sampling campaign and includes a small number of good-quality samples from outside the study area to support model development. Future work should expand the spatial and temporal coverage, integrate additional redox-sensitive indicators, conduct formal uncertainty analyses and evaluate model transferability to other areas in Hanoi and comparable megacities to enhance robustness and generalizability.

Although this study focused on arsenic-contaminated aquifers, the proposed framework is generic and could be readily adapted to other contaminants of concern (*e.g.*, fluoride, nitrate, salinity or emerging pollutants) by redefining the pollution status and indicator set. With appropriate local guidelines, the same ML-based GWQI approach could also be extended to arsenic-affected deltas elsewhere in South and Southeast Asia, thereby enhancing comparability across regions because of the similarity of the hydrogeochemical properties of the huge area of the south Himalaya mountain geology and paleo-geology (*e.g.* the West Bengal, India and Ganges River Delta in Bangladesh and especially the inter-boundary regional Mekong Delta).

4.2. Public health considerations

Although groundwater contamination is widespread in the study area, notably, our field observations revealed that all households currently have access to municipal water for daily use. However, due to the region's fertile soil and agricultural dependence, untreated groundwater continues to be widely used to irrigate vegetables, crops, and ornamental plants. Such practices may lead to the accumulation of toxic elements, especially As, in edible crops, thereby posing potential health risks to consumers.[49–51] In addition, using groundwater with elevated ammonium concentrations may adversely affect crop growth and quality, since nitrogen in the water can act similarly to fertiliser-derived nitrogen.[52]

Therefore, appropriate treatment methods are essential to reduce pollutant concentrations before groundwater in the study area is used for any purpose. Among modern mitigation approaches, sand filtration systems represent a practical, low-

cost, and accessible technology capable of effectively removing contaminants, particularly heavy metals such as As, Fe, and Mn from groundwater.[53–57]

From the management perspective, the ML-based GWQI framework developed here could be used as a screening tool to prioritise wells for detailed monitoring and treatment, especially in the Holocene and shallow RTZ zones where poor and marginal water quality dominate. In parallel, controlling abstraction from the Pleistocene aquifer and promoting low-cost household and community-scale treatment (*e.g.*, sand filtration) would reduce exposure risks while safeguarding the long-term integrity of the Red River Delta groundwater resources.

Author contributions

Vu Thi Duyen: writing – original draft, writing – review & editing, visualisation, project administration, funding acquisition, investigation, formal analysis, conceptualisation, data curation. Nguyen Thanh Dam: writing – review & editing, methodology, visualisation, formal analysis, software, conceptualisation, data curation. Pham Thi Kim Trang: writing – review & editing, supervision, validation, data curation. Michael Berg: writing – review & editing, supervision, validation, resources. Pham Hung Viet: writing – review & editing, funding acquisition, supervision, resources.

Conflicts of interest

There are no conflicts to declare.

Data availability

All data supporting this study (including the full groundwater dataset, machine learning hyperparameter grids, and modelling results) have been included as part of the supplementary information (SI). Supplementary information is available. See DOI: https://doi.org/10.1039/d5va00368g.

Acknowledgements

This research has been done under the research project QG.23.16 of Vietnam National University, Hanoi. The authors thank Lang The Anh for his help during the sampling campaign.

References

1 X. Wang, Y. Tian and C. Liu, Assessment of groundwater quality in a highly urbanized coastal city using water quality index model and bayesian model averaging, *Front. Environ. Sci.*, 2023, **11**, 1086300.
2 L. H. E. Winkel, P. T. K. Trang, V. M. Lan, C. Stengel, M. Amini, N. T. Ha, P. H. Viet and M. Berg, Arsenic pollution of groundwater in Vietnam exacerbated by deep aquifer exploitation for more than a century, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 1246–1251.
3 S. Jasechko, H. Seybold, D. Perrone, Y. Fan, M. Shamsudduha, R. G. Taylor, O. Fallatah and J. W. Kirchner, Rapid groundwater decline and some cases of recovery in aquifers globally, *Nature*, 2024, **625**, 715–721.
4 F. Karandish, S. Liu and I. de Graaf, Global groundwater sustainability: A critical review of strategies and future pathways, *J. Hydrol.*, 2025, **657**, 133060.
5 C. N. Peters, C. Kimsal, R. S. Frederiks, A. Paldor, R. McQuiggan and H. A. Michael, Groundwater pumping causes salinization of coastal streams due to baseflow depletion: Analytical framework and application to Savannah River, GA, *J. Hydrol.*, 2022, **604**, 127238.
6 P. H. Giao, V. T. Hue, N. D. Han, N. T. H. Anh and N. N. Minh, Land subsidence prediction for a new urban mass rapid transit line in Hanoi, *Undergr. Space*, 2020, **5**, 93–104.
7 L. Wright-Contreras, H. March and S. Schramm, Fragmented landscapes of water supply in suburban Hanoi, *Habitat Int.*, 2017, **61**, 64–74.
8 A. van Geen, B. C. Bostick, P. Thi Kim Trang, V. M. Lan, N.-N. Mai, P. D. Manh, P. H. Viet, K. Radloff, Z. Aziz, J. L. Mey, M. O. Stahl, C. F. Harvey, P. Oates, B. Weinman, C. Stengel, F. Frei, R. Kipfer and M. Berg, Retardation of arsenic transport through a Pleistocene aquifer, *Nature*, 2013, **501**, 204–207.
9 T. Agusa, P. T. Trang, V. M. Lan, D. H. Anh, S. Tanabe, P. H. Viet and M. Berg, Human exposure to arsenic from drinking water in Vietnam, *Sci. Total Environ.*, 2014, **488–489**, 562–569.
10 M. Berg, P. T. K. Trang, C. Stengel, J. Buschmann, P. H. Viet, N. Van Dan, W. Giger and D. Stüben, Hydrological and sedimentary controls leading to arsenic contamination of groundwater in the Hanoi area, Vietnam: The impact of iron-arsenic ratios, peat, river bank deposits, and excessive groundwater abstraction, *Chem. Geol.*, 2008, **249**, 91–112.
11 T. Le Luu, Remarks on the current quality of groundwater in Vietnam, *Environ. Sci. Pollut. Res.*, 2019, **26**, 1163–1169.
12 D. Postma, F. Larsen, N. T. Minh Hue, M. T. Duc, P. H. Viet, P. Q. Nhan and S. Jessen, Arsenic in groundwater of the Red River floodplain, Vietnam: Controlling geochemical processes and reactive transport modeling, *Geochim. Cosmochim. Acta*, 2007, **71**, 5054–5071.
13 M. G. Uddin, S. Nash and A. I. Olbert, A review of water quality index models and their use for assessing surface water quality, *Ecol. Indic.*, 2021, **122**, 107218.
14 M. G. Uddin, S. Nash, A. Rahman and A. I. Olbert, A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment, *Water Res.*, 2022, **219**, 118532.
15 H. Zheng, S. Hou, J. Liu, Y. Xiong and Y. Wang, Advanced Machine Learning and Water Quality Index (WQI) Assessment: Evaluating Groundwater Quality at the Yopurga Landfill, *Water*, 2024, **16**, 1666.
16 A. Elmotawakkil, N. Enneya, S. K. Bhagat, M. M. Ouda and V. Kumar, Advanced machine learning models for robust prediction of water quality index and classification, *J. Hydroinform.*, 2025, **27**, 299–319.

17 A. M. Sajib, M. T. M. Diganta, A. Rahman, T. Dabrowski, A. I. Olbert and M. G. Uddin, Developing a novel tool for assessing the groundwater incorporating water quality index and machine learning approach, *Groundw. Sustain. Dev.*, 2023, **23**, 101049.
18 L. M. Sidek, H. A. Mohiyaden, M. Marufuzzaman, N. S. M. Noh, S. Heddam, M. Ehteram, O. Kisi and S. S. Sammen, Developing an ensembled machine learning model for predicting water quality index in Johor River Basin, *Environ. Sci. Eur.*, 2024, **36**, 67.
19 G. S. Solangi, Z. Ali, M. Bilal, M. Junaid, S. Panhwar, H. A. Keerio, I. H. Sohu, S. G. Shahani and N. Zaman, Machine learning, Water Quality Index, and GIS-based analysis of groundwater quality, *Water Pract. Technol.*, 2024, **19**, 384–400.
20 R. Haggerty, J. Sun, H. Yu and Y. Li, Application of machine learning in groundwater quality modeling - A comprehensive review, *Water Res.*, 2023, **233**, 119745.
21 M. K. Gupta, R. Kumar, M. K. Banerjee, N. K. Gupta, T. Alam, S. M. Eldin and M. Y. A. Khan, Assessment of Chambal River Water Quality Parameters: A MATLAB Simulation Analysis, *Water*, 2022, **14**, 4040.
22 M. Y. A. Khan, K. M. Gani and G. J. Chakrapani, Spatial and temporal variations of physicochemical and heavy metal pollution in Ramganga River—a tributary of River Ganges, India, *Environ. Earth Sci.*, 2017, **76**, 231.
23 Y. Li, M. Y. A. Khan, Y. Jiang, F. Tian, W. Liao, S. Fu and C. He, CART and PSO+KNN algorithms to estimate the impact of water level change on water quality in Poyang Lake, China, *Arabian J. Geosci.*, 2019, **12**, 287.
24 T. Xu and F. Liang, Machine learning for hydrologic sciences: An introductory overview, *WIREs Water*, 2021, **8**, e1533.
25 M. Zhu, J. Wang, X. Yang, Y. Zhang, L. Zhang, H. Ren, B. Wu and L. Ye, A review of the application of machine learning in water quality evaluation, *Eco-Environ. Health*, 2022, **1**, 107–116.
26 M. G. Uddin, S. Nash, A. Rahman and A. I. Olbert, A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches, *Water Res.*, 2023, **229**, 119422.
27 T. Chen and C. Guestrin, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 2016.
28 J. Liu, Q. Chu, W. Yuan, D. Zhang and W. Yue, WQI Improvement Based on XG-BOOST Algorithm and Exploration of Optimal Indicator Set, *Sustainability*, 2024, **16**, 10991.
29 B. Rammohan, P. Partheeban, R. Ranganathan and S. Balaraman, Groundwater Quality Prediction and Analysis Using Machine Learning Models and Geospatial Technology, *Sustainability*, 2024, **16**, 9848.
30 D. Khoi, N. Quan, D. Linh, P. Nhi and N. Thuy, Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam, *Water*, 2022, **14**, 12.
31 V. Le, N. Quan, H. Loc, N. Duyen, T. Dung, H. Nguyen and Q. Do, A Multidisciplinary Approach for Evaluating Spatial and Temporal Variations in Water Quality, *Water*, 2019, **11**, 16.
32 B. Lap, T. Phan, H. Du Nguyen, L. Quang, P. Hang, N. Phi, V. Hoang, P. Linh and B. Hang, Predicting Water Quality Index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system, *Ecol. Inform.*, 2023, **74**, 16.
33 D. Nguyen, H. Ha, N. Trinh and M. Nguyen, Application of artificial intelligence for forecasting surface quality index of irrigation systems in the Red River Delta, Vietnam, *Environ. Syst. Res.*, 2023, **12**, 17.
34 H. Nguyen, D. Tran, N. Hoang and T. Nguyen, Predictive modelling physico-chemical properties groundwater in coastal plain area of Vinh Linh and Gio Linh districts of Quang Tri Province, Vietnam, *Water Pract. Technol.*, 2022, **17**, 2100–2112.
35 E. Eiche, T. Neumann, M. Berg, B. Weinman, A. van Geen, S. Norra, Z. Berner, P. T. K. Trang, P. H. Viet and D. Stüben, Geochemical processes underlying a sharp contrast in groundwater arsenic concentrations in a village on the Red River delta, Vietnam, *Appl. Geochem.*, 2008, **23**, 3143–3154.
36 S. Fendorf, H. A. Michael and A. van Geen, Spatial and Temporal Variations of Groundwater Arsenic in South and Southeast Asia, *Science*, 2010, **328**, 1123–1127.
37 E. Stopelli, V. T. Duyen, T. T. Mai, P. T. K. Trang, P. H. Viet, A. Lightfoot, R. Kipfer, M. Schneider, E. Eiche, A. Kontny, T. Neumann, M. Glodowska, M. Patzner, A. Kappler, S. Kleindienst, B. Rathi, O. Cirpka, B. Bostick, H. Prommer, L. H. E. Winkel and M. Berg, Spatial and temporal evolution of groundwater arsenic contamination in the Red River delta, Vietnam: Interplay of mobilisation and retardation processes, *Sci. Total Environ.*, 2020, **717**, 137143.
38 H. Vu, B. Merkel and O. Wiche, Major ions, trace elements and evidence of groundwater contamination in Hanoi, Vietnam, *Environ. Earth Sci.*, 2022, **81**, 305.
39 S. Wen, M. Wen, S. Liang, G. Pang, J. Fan, M. Dong, Y. Wang, J. Zhang and Y. Ye, Spatial Distribution and Mechanisms of Groundwater Hardness in the Plain Area of Tangshan City, China, *Water*, 2024, **16**, 3627.
40 X. Wang, R. N. N. Weerasinghe, C. Su, M. Wang and J. Jiang, Origin and Enrichment Mechanisms of Salinity and Fluoride in Sedimentary Aquifers of Datong Basin, Northern China, *Int. J. Environ. Res. Publ. Health*, 2023, **20**, 1832.
41 N. Abu Bakar, W. N. Wan Ibrahim and S. M. Mohd Faudzi, Arsenic contamination in rice and drinking water: An insight on human cognitive function, *J. Hazard. Mater. Adv.*, 2025, **17**, 100543.
42 G. C. Ghosh, M. J. H. Khan, T. K. Chakraborty, S. Zaman, A. Kabir and H. Tanaka, Human health risk assessment of elevated and variable iron and manganese intake with arsenic-safe groundwater in Jashore, Bangladesh, *Sci. Rep.*, 2020, **10**, 5206.
43 M. I. Rushdi, R. Basak, P. Das, T. Ahamed and S. Bhattacharjee, Assessing the health risks associated with elevated manganese and iron in groundwater in

Sreemangal and Moulvibazar Sadar, Bangladesh, *J. Hazard. Mater. Adv.*, 2023, **10**, 100287.
44 S. Shankar, U. Shanker and Shikha, Arsenic contamination of groundwater: a review of sources, prevalence, health risks, and strategies for mitigation, *Sci. World J.*, 2014, **2014**, 304524.
45 D. van Halem, S. A. Bakker, G. L. Amy and J. C. van Dijk, Arsenic in drinking water: a worldwide water quality concern for water supply companies, *Drink. Water Eng. Sci.*, 2009, **2**, 29–34.
46 A. F. Rusydi, Correlation between conductivity and total dissolved solid in various type of water: A review, *IOP Conf. Ser. Earth Environ. Sci.*, 2018, **118**, 012019.
47 P. B. Le and Z. T. Nguyen, ROC Curves, Loss Functions, and Distorted Probabilities in Binary Classification, *Mathematics*, 2022, **10**, 1410.
48 N. Van Lam, H. Van Hoan and D. Duc Nhan, Investigation into Groundwater Resources in Southern Part of the Red River's Delta Plain, Vietnam by the Use of Isotopic Techniques, *Water*, 2019, **11**, 2120.
49 M. F. Alam, K. G. Villholth and J. Podgorski, Human arsenic exposure risk *via* crop consumption and global trade from groundwater-irrigated areas, *Environ. Res. Lett.*, 2021, **16**, 124013.
50 B. Das, M. K. Pandit, K. Ray, K. Bhattacharyya, A. Pari and P. Sidhya, Impact of irrigation and organic matter amendments on arsenic accumulation in selected vegetables, *Plant Soil Environ.*, 2016, **62**, 266–273.
51 S. Sandil, M. Óvári, P. Dobosy, V. Vetési, A. Endrédi, A. Takács, A. Füzy and G. Záray, Effect of arsenic-contaminated irrigation water on growth and elemental composition of tomato and cabbage cultivated in three different soils, and related health risk assessment, *Environ. Res.*, 2021, **197**, 111098.
52 R. S. Ayers and D. W. Westcot, *Water Quality for Agriculture*, Food and Agriculture Organization of the United Nations, Rome, 1985.
53 M. Berg, S. Luzi, P. T. K. Trang, P. H. Viet, W. Giger and D. Stüben, Arsenic Removal from Groundwater by Household Sand Filters: Comparative Field Study, Model Calculations, and Health Benefits, *Environ. Sci. Technol.*, 2006, **40**, 5567–5573.
54 A. S. Boersma, S. Haukelidsaeter, L. Kirwan, A. Corbetta, L. Vos, W. K. Lenstra, F. Schoonenberg, K. Borger, P. W. J. J. van der Wielen, M. A. H. J. van Kessel, C. P. Slomp and S. Lücker, Influence of filter backwashing on iron, manganese, and ammonium removal in dual-media rapid sand filters used for drinking water production, *Water Res.*, 2025, **270**, 122809.
55 C. A. Coles and D. Rohail, Effect of aeration, iron and arsenic concentrations, and groundwater matrix on arsenic removal using laboratory sand filtration, *Environ. Geochem. Health*, 2020, **42**, 4051–4064.
56 B. Danka and I. Ján, Removal of iron and manganese from water using filtration by natural materials, *Pol. J. Environ. Stud.*, 2010, **19**, 1117–1122.
57 M. Watson, J. Nikić, A. Tubić, M. K. Isakovski, M. Šolić, B. Dalmacija and J. Agbaba, Repurposing spent filter sand from iron and manganese removal systems as an adsorbent for treating arsenic contaminated drinking water, *J. Environ. Manage.*, 2022, **302**, 114115.

