Geochemical and Machine Learning Approaches to Groundwater Fluoride Prediction in Karaga District, Northern Ghana

Emmanuel Daanoba Sunkari[1, 2]*, Dickson Abdul-Wahab[3], Mélida Gutiérrez[4,] Prasun Chakrabarti[5] & Abayneh Ataro Ambushe[2]

[1]Mining Engineering, Faculty of Integrated and Advanced Technology, Sir Padampat Singhania University, Udaipur-313601, Rajasthan, India

[2]Department of Chemical Sciences, Faculty of Science, University of Johannesburg, P.O. Box 524, Auckland Park 2006, Johannesburg, South Africa

[3]Graduate School of Nuclear and Allied Sciences, University of Ghana, Legon, Accra, Ghana

[4]School of Earth, Environment and Sustainability, Missouri State University, Springfield, Missouri, USA

[5]Department of Computer Science and Engineering, Faculty of Computing and Informatics, Sir Padampat Singhania University, Udaipur-313601, Rajasthan, India

*Corresponding Author: emmanueldaanobasunkari@gmail.com

*ORCID: 0000-0002-0898-2286

Abstract

Fluoride contamination of groundwater affects over 200 million people globally, with Africa serving as a primary hotspot. The Karaga District in Ghana's Northern Region represents a critical fluoride hotspot, where 4 out of 10 children likely face exposure to concentrations exceeding 1.5 mg/L. Despite being identified as high-risk, the specific geochemical mechanisms controlling fluoride mobilization in the region's Voltaian Supergroup aquifers remain inadequately understood, limiting the development of targeted mitigation strategies. This study aimed to develop and validate an integrated framework combining geochemical modelling, compositional data analysis, and machine learning to predict fluoride concentrations and elucidate mobilization mechanisms in Karaga District’s groundwater. About 34 groundwater samples from the Karaga District

were collected and analyzed for hydrochemical parameters. The data was processed using PHREEQC for geochemical modelling and isometric log-ratio transformation for compositional analysis. Additionally, 6 supervised machine learning algorithms were trained on 152 archived samples from neighbouring districts and subsequently validated using the 34 newly collected groundwater samples. A mechanistic Mobility Index was developed using fluoride-independent components and entropy-based weighting. Fluoride concentrations ranged from 0.07 to 6.04 mg/L, with 17.6% exceeding WHO guidelines. Na-HCO_3 waters dominated (64.7%), but Na-Cl waters exhibited the highest fluoride (mean 3.75 mg/L), revealing that evaporite dissolution drives extreme contamination. Machine learning identified total dissolved solids and pH as primary predictors, demonstrating nonlinear fluoride behaviour. The Multilayer Perceptron model achieved R^2 of 0.668, while the Mobility Index demonstrated exceptional discrimination for WHO exceedance (AUROC 0.94), with robust spatial transferability across communities. This integrated approach provides a mechanistically grounded, field-deployable framework for fluoride risk assessment. The Mobility Index enables cost-effective community screening using only basic measurements, supporting targeted intervention strategies in fluoride-endemic regions globally.

Keywords: Groundwater Quality, Geochemical Modelling, PHREEQC, Compositional Data Analysis, SHAP Analysis

1 Introduction

Fluoride contamination of groundwater is a widespread global issue that affects over 200 million people in more than 100 countries, with Africa, Asia, and Latin America serving as primary hotspots for this environmental and public health challenge (Shaji et al., 2024). The severity and global distribution of this contamination is exemplified by regional statistics: in India, fluoride levels in groundwater can reach up to 48 mg/L, affecting over 90 million people with dental and skeletal fluorosis (Bera et al., 2021),

while in Mexico, 97% of tested groundwater samples exceeded fluoride limits, with concentrations reaching 8.8 mg/L (Padilla-Reyes et al., 2024).

The health implications of fluoride exposure demonstrate clear dose-response relationships that vary according to concentration and duration of exposure. Chronic exposure to fluoride above 1.5 mg/L is directly linked to dental and skeletal fluorosis, while prolonged exposure can cause arthritis, kidney, liver, and neurological problems (Dar & Kurella, 2024; Sunkari & Ambushe, 2024). Children represent a particularly vulnerable population, with studies in Thailand revealing a 54.3% prevalence of dental fluorosis among children exposed to groundwater containing ≥1.5 ppm fluoride (Rojanaworarit et al., 2021). Beyond skeletal effects, elevated fluoride levels are associated with hypertension and cardiovascular impairment in endemic regions (Varol & Varol, 2012), underscoring the critical importance of accurate prediction and risk assessment capabilities for fluoride-affected groundwater systems.

The Karaga District in Ghana's Northern Region exemplifies the severity of geogenic contamination in West African sedimentary aquifers and represents a critical fluoride hotspot that demands urgent scientific attention. Recent country-wide hazard modelling identified the Karaga District as one of the most severely affected areas in Ghana, where 4 out of 10 children are potentially exposed to fluoride concentrations exceeding 1.0 mg/L (Araya et al., 2022). The geological setting of the Karaga District is particularly problematic for fluoride contamination, as the weathering of fluoride-bearing rocks from the Voltaian Supergroup and dissolution of fluoride-rich minerals (fluorapatite, amphiboles, fluorite, biotite, and muscovite) create hydrogeochemical conditions conducive to elevated groundwater fluoride concentrations (Sunkari et al., 2022, 2025a). The Voltaian formations contain fluoride-bearing minerals within their mudstones and sandstones, with documented fluoride concentrations that frequently exceed safe drinking water limits in neighbouring districts (Apambire et al., 1997). The broader Northern Ghana region, including the Karaga District, falls within an estimated 920,000 people at risk from

fluoride contamination, with approximately 240,000 children (0-9 years) living in at-risk areas (Araya et al., 2022).

The mobilization of fluoride in groundwater systems is governed by complex, interconnected geochemical processes that operate across multiple spatial and temporal scales. Primary control mechanisms include the presence and dissolution of fluoride-bearing minerals, with fluorite (CaF_2), fluorapatite, biotite, muscovite, hornblende, and amphiboles serving as the principal geological sources (Ali et al., 2019; Aravinthasamy et al., 2020; Alam et al., 2024). Long-term water-rock interactions facilitate fluoride leaching from host minerals, with the weathering of granitic and basaltic aquifers being particularly effective in releasing fluoride through silicate dissolution processes (Chen et al., 2023; Alam et al., 2024).

The solubility and mobility of fluoride in groundwater systems are strongly influenced by hydrogeochemical parameters including pH, calcium concentration, and alkalinity. High pH conditions enhance fluoride solubility by promoting desorption from mineral surfaces and facilitating fluorite dissolution (Luo et al., 2018; Ali et al., 2019). Low calcium levels reduce fluorite saturation, enabling more fluoride to remain in solution rather than precipitating as calcium-fluoride minerals (Ali et al., 2019; Chen et al., 2023). Alkalinity, represented primarily by bicarbonate (HCO_3^-) concentrations, exhibits positive correlations with fluoride levels by facilitating mineral dissolution and competitive desorption from mineral surfaces (Aravinthasamy et al., 2020).

Competitive ion effects play a significant role in fluoride mobilization, particularly the presence of competing anions such as OH^- and HCO_3^-, which compete with fluoride for adsorption sites on mineral surfaces including goethite and gibbsite, thereby enhancing fluoride mobility in alkaline environments (Luo et al., 2018; Ali et al., 2019). Additionally, sodium-rich water types (Na-HCO_3) are frequently associated with higher fluoride concentrations, likely due to cation exchange processes that alter the ionic strength and competitive equilibria within the groundwater system (Alam et al., 2024).

Various methodological approaches have been employed to predict fluoride concentrations in groundwater, each offering distinct advantages while suffering from specific limitations that constrain their effectiveness in complex hydrogeological settings. Statistical methods including Random Forest (RF), Artificial Neural Networks (ANN), and Logistic Regression (LR) have demonstrated varying degrees of success, with RF achieving 89% accuracy, followed by ANN (85%) and LR (76%) in Chinese groundwater systems (Nafouanti et al., 2021). Kriging interpolation methods have proven effective for spatial prediction of fluoride concentrations based on point measurements in Pakistani aquifers (Ahmad et al., 2023), though these geostatistical approaches assume stationarity and may oversimplify regional variability in complex geological settings.

Thermodynamic modelling approaches utilizing software such as PHREEQC have provided mechanistic insights into geochemical processes controlling fluoride enrichment, with simulations revealing that fluoride mobilization is often driven by fluorite dissolution and other fluoride-bearing mineral interactions, supported by specific pH and calcite precipitation conditions (Liu & Chen, 2024). However, thermodynamic models require detailed chemical input parameters and accurate mineral assemblage data that are often unavailable under field conditions, limiting their applicability in data-sparse environments.

Machine learning applications have gained considerable popularity in hydrogeochemistry, with advanced algorithms including Extreme Learning Machine (ELM), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and XGBoost demonstrating strong predictive capabilities. In Punjab, India, ELM achieved impressive performance with R^2 of 0.95 and RMSE of 0.33 (Kerketta et al., 2024), while in Turkey, XGBoost and Convolutional Neural Networks emerged as top performers for fluoride prediction using diverse water quality parameters (Demir Yetiş et al., 2024). In Pakistan, Random Forest modelling successfully mapped high-risk fluoride zones and estimated that approximately 13 million

people were exposed to concentrations exceeding 1.5 mg/L (Ling et al., 2022).

Compositional data analysis (CoDA) methods, particularly Principal Component Analysis (PCA) applied to hydrogeochemical datasets, have been employed to identify underlying processes influencing fluoride levels, revealing relationships between fluoride and parameters such as Na^+, HCO_3^-, and total dissolved solids (Narsimha Adimalla, 2020; Liu et al., 2021; Sunkari et al., 2025b). However, most fluoride modelling studies apply PCA heuristically without proper compositional data transformations, potentially introducing spurious correlations and misrepresenting relationships among chemical constituents. The application of rigorous CoDA frameworks, including isometric log-ratio (ilr) transformation, remains a notable gap in fluoride prediction literature.

Despite the diversity of approaches applied to fluoride prediction, several fundamental knowledge gaps and methodological limitations persist that constrain the development of robust, transferable prediction frameworks. Single method approaches frequently fail to capture the nonlinear, multivariate nature of fluoride mobilization and transport processes. Geostatistical methods such as kriging assume spatial stationarity and may oversimplify regional variability in heterogeneous geological settings (Nafouanti et al., 2021). Thermodynamic models, while mechanistically robust, rely heavily on accurate input parameters including saturation indices and detailed mineral presence data that are rarely available in field applications (Liu & Chen, 2024).

This study advances fluoride risk assessment beyond ML-only or geochemistry-only workflows by (i) coupling PHREEQC-based thermodynamic interpretation with rigorously transformed compositional features (ilr/CoDA) to avoid spurious correlation artefacts; (ii) enforcing leakage-aware, fluoride-blind feature design for supervised prediction and validating models on an independently collected Karaga dataset; and (iii) introducing a mechanistically interpretable Mobility Index (MI) that translates geochemical controls into a screening-oriented risk signal,

supported by calibration and discrimination testing. Related integrative groundwater-quality and entropy-based risk studies have demonstrated the value of combining indexing, multivariate structure, and health-risk framing (Kumar & Singh, 2024b, 2024a, 2025); however, they typically do not integrate CoDA-consistent geochemical representation with externally validated fluoride prediction and interpretability in a single, reproducible workflow.

A particularly significant limitation in current fluoride research is the underutilization of model interpretability methods. Machine learning studies often focus exclusively on predictive accuracy while neglecting interpretability tools such as SHAP (SHapley Additive exPlanations) or comprehensive feature importance analysis, which are crucial for understanding key variables influencing fluoride levels (Demir Yetiş et al., 2024). This limitation severely constrains the ability to translate complex models into actionable public health insights or evidence-based policy decisions, particularly in fluoride-endemic areas where mechanistic understanding is essential for effective intervention strategies.

Studies increasingly advocate for the integration of machine learning approaches with geochemical knowledge to improve both predictive power and mechanistic interpretability. Combining ion-specific geochemical inputs with ensemble machine learning models has demonstrated superior fluoride risk mapping capabilities in Pakistan (Ling et al., 2022). Such hybrid approaches leverage the explanatory strength of mechanistic models with the pattern recognition capabilities of artificial intelligence, particularly valuable in settings where detailed chemical datasets are limited (Kerketta et al., 2024).

Compositional data analysis methods remain significantly underutilized in fluoride research, despite their ability to address the closure problem inherent in hydrochemical data where concentrations sum to a constrained total. Most studies apply PCA without appropriate compositional data transformations, potentially creating spurious correlations and misrepresenting relationships among chemical elements (Narsimha

Adimalla, 2020; Liu et al., 2021). Proper application of CoDA frameworks, including isometric log-ratio transformation, represents a critical methodological gap that limits the robust interpretation of multivariate hydrogeochemical relationships in fluoride systems.

The complexity of fluoride behaviour in sedimentary aquifer systems, combined with the critical public health implications in regions such as the Karaga District, necessitates the development of integrated methodological frameworks that combine mechanistic understanding with advanced predictive capabilities. Despite being identified as a high-risk fluoride area through national modelling studies, detailed mechanistic understanding of fluoride mobilization processes in the Karaga District's Voltaian Supergroup aquifers remains limited, with existing studies focusing primarily on regional hazard mapping rather than local geochemical controls and process-based prediction.

This study aims to address these critical knowledge gaps by developing and validating an integrated approach that combines geochemical modelling, compositional data analysis, and machine learning techniques for predicting fluoride concentrations and understanding mobilization mechanisms in the Karaga District groundwater system. The specific objectives are to: (1) characterize the hydrogeochemical controls on fluoride mobilization in Voltaian Supergroup aquifers through comprehensive water chemistry analysis and thermodynamic modelling; (2) develop and validate machine learning models for predicting fluoride concentrations and WHO guideline exceedance using geochemically-informed feature sets; (3) apply rigorous compositional data analysis techniques, including isometric log-ratio transformation and sequential binary partitioning, to identify fundamental geochemical processes controlling fluoride behaviour; (4) integrate SHAP analysis and other interpretability methods to translate complex model predictions into mechanistic insights and actionable risk assessment tools; and (5) establish critical thresholds and develop a composite fluoride mobility index for early warning and targeted intervention strategies.

The integrated methodology is designed to be portable to similar sedimentary aquifer contexts; however, broader transferability beyond Northern Ghana should be evaluated using additional seasons and independent regions. The findings will advance both scientific understanding of fluoride geochemistry in sedimentary systems and practical capabilities for risk assessment, targeted mitigation strategies, and evidence-based decision-making in regions where geogenic fluoride contamination threatens water security and public health.

2 Materials and Methods

2.1 Study Area

2.1.1 Geographic Setting and Climate

The Karaga District is located in the northeast of Ghana's Northern Region, covering an area of 3,119.3 km^2 with a population of 114,225 as of 2021 (Ghana Statistical Service (GSS), 2021). The district capital, Karaga, is situated 24 km from Gushegu and 94 km from Tamale, the regional capital. Karaga District lies between latitudes 9°30' South to North and longitudes 0° to 45' West, bordering West and East Mamprusi to the north, Savelugu/Nanton to the west, and Gushegu to the south and east (Fig 1a). The district experiences a tropical continental climate, with a rainy season from May to October and mostly dry conditions for the rest of the year. Annual rainfall ranges between 900 and 1000 mm, with the heaviest precipitation occurring in August and September. Temperatures remain high throughout the year, reaching a maximum of 36 degrees Celsius or higher in March and April, while the lowest temperatures are observed between November and February. The district's vegetation is characterized by typical Guinea Savannah, consisting of tall grasses interspersed with drought-resistant trees such as shea and dawadawa, which serve as a source of income for the local population (Ghana Statistical Service (GSS), 2021).

2.1.2 Geology and Hydrogeology

The Karaga District is predominantly underlain by rocks of the Voltaian Supergroup, which covers most of northern Ghana and unconformably overlies the lower Proterozoic Birimian Supergroup, associated granitoids, and the lower to middle Protozoic Tarkwaian Supergroup (Ghana Geological Survey (GGS), 2009). The Voltaian Supergroup, dating from the late Proterozoic to early Paleozoic era, is divided into Upper, Lower, and Middle formations, dominated by sandstones with minor mudstone, arkoses, feldspars, shales, graywackes, siltstones, evaporites, limestones, and conglomerates. The sediments were primarily sourced from a glacial period followed by prolonged marine invasions (Achcampong & Hess, 1998; Anani, 1999). The supergroup is further classified into three lithostratigraphic groups: Oti/Pendjari, Kwahu/Bombouaka, and Tamale/Obosum beds (in chronological order) (Abu et al., 2019). The study area is predominantly composed of rocks from the Oti/Pendjari group and the Tamale/Obosum beds (Fig 1b), primarily consisting of mudstone and sandstones (Ghana Geological Survey (GGS), 2009). Figure 1b presents the geological map of the district, including the mapped lithostratigraphic units of the Voltaian Supergroup and major structural lineaments/fault traces, together with towns and sampling locations. The dominant surface units are mudstone–siltstone packages with interbedded arkosic/lithic sandstones (Bimbila Formation) and undifferentiated mudstone–siltstone–sandstone units (Obosum Group), with localized sandstone-dominated units (Bunya and Panabako formations) and minor tuffaceous/laminated intervals (Darebe Member of the Kodjari Formation).

The hydrogeology of the area is mainly controlled by secondary porosity (Menyeh & Sarpong Asare, 2013). Due to the absence of primary porosity in the lithologies, groundwater occurrence is mostly attributed to secondary porosity caused by jointing, shearing, fracturing, and weathering. The success rate of drilling boreholes in the Oti-Pendjari Group is about 56%, with yields ranging from 0.41 to 9 m3/h and a mean yield of approximately 6.2 m^3/h (Dapaah-Siakwan & Gyau-Boakye, 2000).

The recharge rate of the Voltaian varies between 2.07×10^{-5} m/day and 2.85×10^{-4} m/day, contributing about 0.3% to 4.1% of the region's annual precipitation (Sunkari et al., 2018, 2024).

Within the Voltaian sedimentary sequence, spatial variability in mudstone–sandstone proportions and the presence of carbonate-bearing and evaporite-influenced horizons can plausibly generate the observed contrast between Na–HCO_3 waters (dominant, lower mean fluoride) and Na–Cl waters (minority, highest fluoride). In particular, mudstone-rich intervals and clay-bearing units promote cation exchange and calcium depletion (natural softening), while saline end-member signatures (Na–Cl) are consistent with evaporite dissolution or saline mixing that further suppresses Ca^{2+} availability, thereby enhancing fluorite undersaturation and fluoride persistence in solution.

2.1.3 Sampling

A total of 34 groundwater samples were collected from active public boreholes drilled and maintained by World Vision International in the Karaga District, Northern Ghana (Fig. 1a). The samples were collected using 0.5 L polyethylene bottles in October 2022. The sample bottles were thoroughly precleaned with deionized water, 10% nitric acid, and distilled water to ensure they were free from contamination (Sunkari & Abu, 2019). Before sampling, the boreholes were pumped for about 5 minutes to purge the aquifers and prevent cross-contamination. The groundwater samples were filtered using hand-held syringes with filter heads of 0.45 µm cellulose filter membrane. Physicochemical parameters such as pH, temperature, electrical conductivity (EC), and total dissolved solids (TDS) were monitored using water quality probes. The 0.5 L polyethylene bottles were tightly sealed, and to avoid chemical alterations, the samples were kept in an ice chest at 4 °C. The samples were then transported to the Ghana Atomic Energy Commission's (GAEC) Laboratory for ion analysis. These 34 Karaga samples were reserved exclusively as an external validation set; model development used an independent archive of 152 groundwater samples from Bawku West, Garu, Gusheigu, Kintampo South,

Saboba, Savelugu, Talensi, West Gonja, and Zabzugu (details in Section 2.7)

2.2 Laboratory Analysis

DIONEX ICS-90 ion chromatography system (Dionex Corporation, California, USA) was used to determine the concentrations of Cl^-, SO_4^{2-}, F^-, and NO_3^-. The concentrations of HCO_3^- were determined through in situ titration using 1,690,001 HACH® digital titrator (Hach Company, Loveland, Colorado, USA). However, concentrations of K^+ and Na^+ were measured using a Flame photometer (FP910-4), while Ca^{2+} and Mg^{2+} concentrations were measured using an AA2OFS Fast Sequential atomic absorption spectrometer. For accuracy and precision, BCR 398 and 399 certified reference materials (CRMs) were used with continuous internal validation, mostly by repeated analyses of standards and samples. This produced a precision of 5% for all the ions analysed.

2.3 Hydrogeochemical Characterization

Hydrogeochemical characterization was performed to understand the chemical composition and water-rock interactions governing groundwater quality in the Karaga District. This analysis established baseline water quality parameters and identified potential factors controlling fluoride mobility.

2.3.1 Correlation Analysis

Spearman's rank correlation analysis was performed to identify significant relationships between fluoride and other hydrogeochemical parameters. Spearman's correlation was selected over Pearson's correlation due to its robustness against non-normal distributions and nonlinear relationships commonly encountered in hydrogeochemical datasets. The correlation coefficient (ρ) for each parameter pair was calculated as:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \quad (1)$$

where d_i is the difference between the ranks of corresponding variables and n is the number of observations. A correlation matrix was generated to visualize the strength and direction of these relationships, with statistical significance assessed at $p < 0.05$.

2.3.2 Water Type Classification

Hydrochemical facies were determined to classify groundwater samples based on their dominant ions. The classification followed the isometric log-ratio (ILR) coordinates as proposed by Shelton et al. (2018) instead of traditional Piper diagram to address the compositional nature of the data. For each water sample, the dominance of each ion was evaluated using the following criterion:

$$D_i = \frac{C_i}{\sum_{j=1}^{n} C_j} > 0.5 \quad (2)$$

where D_i is the dominance indicator for ion i, C_i is the concentration of ion i in meq/L, and $\sum_{j=1}^{n} C_j$ is the sum of concentrations of all ions in the respective cation or anion group. A threshold of 50% was used to identify dominance. For samples where no single ion exceeded this threshold, a mixed type was assigned. Equation and details of ILR are presented in supplementary methods (Equation 3 to 6).

2.3.3 Geochemical Parameters

Several geochemical ratios and indices were calculated to better understand the processes controlling fluoride mobility: (i) Ca^{2+}/F^- ratio: Indicator of potential fluorite (CaF_2) dissolution or precipitation control, (ii) HCO_3^-/F^- ratio: Measure of potential competitive effects between bicarbonate and fluoride, (iii) Na^+/Ca^{2+} ratio: Indicator of cation exchange processes, (iv) Mg^{2+}/Ca^{2+} ratio: Measure of differential dissolution of carbonate minerals, (v) HCO_3^-/Ca^{2+} ratio: Indicator of calcite

precipitation potential, (vi) $F^-/\sum$Anions: Proportion of fluoride relative to total anion content (vii) pH/F^- index: Combined effect of pH and fluoride normalized by calcium, (viii) TDS/F^- ratio: Relationship between total mineralization and fluoride, (ix) Weathering index: $(Na^+ + K^+)/(Ca^{2+} + Mg^{2+} + Na^+ + K^+)$, indicating silicate versus carbonate weathering dominance. These derived parameters provided additional insights into the geochemical processes controlling fluoride mobilization beyond what could be observed from raw concentration data alone.

2.4 Compositional Data Analysis

Major-ion hydrochemical data are compositional; therefore, concentrations were converted from mg L^{-1} to meq L^{-1} and closed to unit sum prior to multivariate analysis to avoid closure-induced artefacts. A sequential binary partition (SBP) was constructed to generate geochemically interpretable balances: the first split was forced to separate cations from anions (Table 1), while the remaining partitions were guided by Ward clustering on clr-transformed compositions. The resulting isometric log-ratio (ilr) coordinates were used for (i) water-type/facies classification and (ii) subsequent multivariate analyses, including PCA and correlation assessment. Full mathematical definitions and derivations for closure, clr, SBP construction, ilr mapping, and PCA are provided in Supplementary Methods (Equations 7 to 14)

2.5 Geochemical Modelling

Geochemical modelling was employed to simulate the chemical speciation, mineral saturation states, and thermodynamic controls on fluoride mobility in groundwater (Sunkari et al., 2023). This approach provided mechanistic insights into the processes governing fluoride release and sequestration that could not be obtained through direct measurement alone.

2.5.1 PHREEQC Implementation

Geochemical modelling was performed using PHREEQC version 3.8.6 (Parkhurst and Appelo, 2021), a widely accepted equilibrium speciation and reaction-path model developed by the U.S. Geological Survey. The

program solves a set of nonlinear equations describing chemical equilibria to determine the distribution of aqueous species, saturation indices, and activities in solution. For this study, the Lawrence Livermore National Laboratory (LLNL) thermodynamic database was selected due to its comprehensive coverage of fluoride minerals and aqueous complexes. The PHREEQC models were implemented programmatically using the phreeqpy interface (Müller et al., 2011), which allows direct integration with Python for batch processing and analysis of multiple samples. Each groundwater sample was modelled as a separate SOLUTION block.

2.5.2 Speciation Calculations

The chemical speciation of fluoride and relevant cations was calculated for each sample to determine the distribution between free ions and complexed species. The modelling considered all possible aqueous complexes involving fluoride, with particular attention to calcium-fluoride complexes (CaF^+) which play a crucial role in fluoride mobility. For each sample, the following species concentrations were extracted from the PHREEQC output: Free fluoride ion (F^-), Calcium-fluoride complex (CaF^+) and Free calcium ion (Ca^{2+}). The percentage distribution of fluoride between free and complexed forms was calculated as:

$$F_{free}(\%) = \frac{m_{F^-}}{m_{F^-} + m_{CaF^+}} \times 100 \quad (3)$$

$$CaF(\%) = \frac{m_{CaF^+}}{m_{F^-} + m_{CaF^+}} \times 100 \quad (4)$$

where m_{F^-} and m_{CaF^+} represent the molal concentrations of free fluoride and calcium-fluoride complex, respectively.

2.5.3 Activity Calculations

The activity of dissolved species, which represents their effective concentration in solution, was calculated using the extended Debye-Hückel equation:

$$\log \gamma_i = -Az_i^2 \frac{\sqrt{I}}{1 + Ba_i\sqrt{I}} + bI \quad (5)$$

where γ_i is the activity coefficient of species i, z_i is its charge, I is the ionic strength of the solution, a_i is the effective ion size parameter, and A, B, and b are temperature-dependent constants (Kontogeorgis et al., 2018). The ionic strength of each sample was calculated as:

$$I = \frac{1}{2}\sum_i m_i z_i^2 \quad (6)$$

where m_i is the molality of ion i.

The activities of key species were then calculated as:

$$a_i = \gamma_i \times m_i \quad (7)$$

with particular focus on the activities of fluoride (a_{F^-}) and calcium ($a_{Ca^{2+}}$) due to their importance in controlling fluorite solubility.

2.5.4 Saturation Indices

The saturation state of groundwater with respect to various minerals was evaluated using saturation indices (SI). The SI is defined as the logarithm of the ratio between the ion activity product (IAP) and the solubility product constant (K_{sp}):

$$SI = \log_{10}\frac{IAP}{K_{sp}} \quad (8)$$

An SI value of zero indicates equilibrium, positive values indicate supersaturation (tendency for precipitation), and negative values indicate undersaturation (tendency for dissolution).

Saturation indices were calculated for the following minerals relevant to fluoride geochemistry, thus Fluorite (CaF_2), Calcite ($CaCO_3$), Dolomite ($CaMg(CO_3)_2$),Gypsum ($CaSO_4{\cdot}2H_2O$) and Other minerals such as Sylvite (KCl), Halite (NaCl), Magnesite ($MgCO_3$), Anhydrite ($CaSO_4$), Mirabilite ($Na_2SO_4{\cdot}10H_2O$), Thenardite (Na_2SO_4), Nitrocalcite ($Ca(NO_3)_2{\cdot}4H_2O$), and Nitratine ($NaNO_3$). Special attention was given to fluorite, as it is the principal mineral controlling fluoride solubility in most groundwater systems. The fluorite equilibrium distance was calculated as the absolute value of the fluorite saturation index (Hem, 1985; Parkhurst & Appelo, 1999): Fluorite Equilibrium Distance = $|SI_{fluorite}|$. This parameter

quantifies how far a sample is from equilibrium with fluorite, regardless of whether it is under- or supersaturated.

2.5.5 Thermodynamic Controls on Fluoride Solubility

The solubility of fluorite can be expressed by the equilibrium reaction: $CaF_2 \rightleftharpoons Ca^{2+} + 2F^-$

with the corresponding solubility product: $K_{sp} = a_{Ca^{2+}} \times (a_{F^-})^2$. Rearranging to express fluoride activity:

$$a_F = \sqrt{\frac{K_{sp}}{a_{Ca^{2+}}}} \quad (9)$$

Which illustrates the inverse relationship between calcium and fluoride activities at equilibrium with fluorite. For each sample, the theoretical fluoride activity at equilibrium with fluorite was calculated and compared with the actual fluoride activity to assess the degree of thermodynamic control.

2.5.6 Derived Geochemical Parameters from PHREEQC

Several additional parameters were derived from the PHREEQC modelling results to aid in the interpretation of fluoride geochemistry: PHREEQC pH the pH calculated by PHREEQC after achieving charge balance, Ionic strength: the total ionic strength of the solution, Free calcium percentage: the proportion of calcium present as free Ca^{2+} ions, Free fluoride percentage: the proportion of fluoride present as free F^- ions and calcium-fluoride complex percentage: the proportion of fluoride complexed with calcium. These derived parameters, along with the saturation indices and activities, were incorporated into the subsequent machine learning and statistical analyses to improve model performance and interpretability.

2.6 Fluoride Mobility Index Development

A mechanistic Mobility Index (MI) was developed to rank the propensity of groundwater to mobilise F^- prior to observing fluoride itself. Components were deliberately constructed to be F-independent and then combined

with entropy-based weights estimated on the ML (development) subset and applied unchanged to the SA (external) subset. The workflow comprised: component construction, robust quantile scaling, Na/Cl gating of the weathering signal, entropy-weight computation on ML, composite MI formation, and a multi-layer validation suite (construct, directional and criterion validity, plus spatial cross-validation). The approach was adopted to preserve mechanistic interpretability while remaining agnostic to the unknown F^- outcome during construction, thereby mitigating circularity.

2.6.1 Theoretical Basis for the Mobility Index

Let concentrations (mg L^{-1}) of ($Na, HCO_3^-, Ca^{2+}, Mg^{2+}, Cl^-$) be converted to milliequivalents per litre via $meq, L^{-1} = \frac{c_{mg,L^{-1}}}{M}, |z|$, where (M) is molar mass and (z) is ionic valence. Closing the ($Na, HCO_3^-, Ca^{2+}, Mg^{2+}$) subcomposition to unity yields proportions ($s = (s_{Na}, s_{HCO_3}, s_{Ca}, s_{Mg})$) with ($\sum s_k = 1$). A single isometric log-ratio (ILR) coordinate was then defined to capture alkali–bicarbonate versus alkaline-earth balance:

$$z_W \equiv \frac{1}{2}(\ln s_{Na} + \ln s_{HCO_3} - \ln s_{Ca} - \ln s_{Mg}). \quad (10)$$

Higher (z_W) indicates silicate-weathering/alkali dominance relative to Ca–Mg control.

Mineral control was represented by fluorite undersaturation, $U_{raw} = \max(-SI_{fluorite}, 0)$,

ensuring only undersaturated conditions increase mobility. Acid–base control entered through measured pH, and medium chemistry through the activity (or activity-coefficient) term for fluoride, here denoted (a_F) (larger effective activity favouring mobility under otherwise similar conditions). A Na/Cl “gate” damped marine-like or evaporitic signatures that can inflate (z_W) without truly reflecting silicate weathering: when the equivalent-ratio ($\rho = Na/Cl \in [0.75, 0.95]$), the weathering component was multiplicatively shrunk by a factor ($s = 0.5$) (else unchanged). The ILR coordinate provides a compositionally coherent, dimensionless weathering signal;

undersaturation and activity embody thermodynamic drivers, and Na/Cl gating attenuates non-weathering salinity artefacts.

2.6.2 Entropy-Based Weighting Method

To combine Mobility Index components without unit-driven dominance, each component was mapped to a common [0,1] scale using a robust quantile transform fitted on the development (ML) subset and applied unchanged to the external (SA) subset. Component weights were then estimated only on the ML subset using Shannon entropy weighting and subsequently locked and reused for SA to prevent leakage. The entropy method yields data-adaptive, scale-free weights that emphasize components with higher information dispersion. Full formulae for probability normalization, entropy computation, and weight derivation are provided in Supplementary Methods S2.

2.6.3 Mobility Index Formulation

(a) Component construction and scaling

Four F-independent components were constructed:

$$W \equiv f_W(z_W,\rho) = \begin{cases} s \cdot \tilde{z}_W, & \rho \in [0.75,0.95], \\ \tilde{z}_W, & \text{otherwise}, \end{cases} \qquad (11)$$

$$U \equiv \widetilde{U_{raw}}, \quad P \equiv \widetilde{pH}, \quad G \equiv \widetilde{a_F},$$

where tildes denote the robust quantile map $\tilde{x} = \min\left\{\max\left(\frac{x-q_{0.05}}{q_{0.95}-q_{0.05}},0\right),1\right\}$, with $(q_{0.05},q_{0.95})$ fitted on ML for each component and applied unchanged to SA.

(b) Composite index

Let $X_{ij} \in [0,1]$ be the scaled value of component (j) for sample (i), and $w = (w_1,\ldots,w_p)$ the entropy weights from Section 2.6.2. The Mobility Index for sample (i) was defined as

$$MI_i = \frac{\sum_{j=1}^{p} w_j X_{ij}}{\sum_{j=1}^{p} w_j} \quad \text{with} \quad p = 4, \qquad (12)$$

$$j \in \{\text{Weathering}, \quad \text{Undersat}, \text{pH}, \Gamma_F\}.$$

A sensitivity analysis computed an equal-weights variant for benchmarking. The Robust scaling confines all components to a common [0,1] support, improving numerical stability and comparability; gated weathering avoids spurious salinity effects; the linear convex aggregation preserves interpretability as a transparent trade-off among mechanisms.

2.6.4 Validation of the Mobility Index

Validation was designed to probe internal coherence (construct), expected directional behaviour, and external screening performance (criterion), with spatial robustness checks.

(a) Construct validity

Spearman rank correlations were estimated between MI and (i) $ILR_{weather}$ and (ii) fluorite undersaturation, separately for ML and SA. For each pair, non-parametric 95% confidence intervals were obtained via bootstrap ($B = 10{,}000$) and permutation p-values were computed under label exchangeability:

$$\hat{r} = \rho_S(x,y), \; CI_{0.95} = [\widehat{r^*_{0.025}}, \widehat{r^*_{0.975}}], \qquad (13)$$

$$p = \frac{1 + \sum_{b=1}^{B} I\{|\rho_S(x,\pi_b(y))| \geq |\hat{r}|\}}{1 + B}.$$

This was chosen so that rank-based, resampling-driven inference avoids parametric assumptions and is robust to heavy-tailed hydrochemical distributions.

(b) Directional check (SA only)

A median quantile regression tested the expected monotone relation between MI and a calcium control proxy,

$$MI = \beta_0 + \beta_1 \log\left(1 + \frac{Ca}{F}\right) + \varepsilon, \qquad \tau = 0.5. \qquad (14)$$

A negative β_1 is hypothesised a priori under the fluorite-control mechanism (increasing Ca relative to F suppresses mobility). Only the sign and significance of β_1 were used for directional verification. The quantile

regression at $\tau = 0.5$ is insensitive to heteroscedasticity and focuses on the typical (median) response, aligning with a directional check rather than full calibration.

(c) Criterion validity for exceedance screening

Exceedance of the WHO guideline was encoded as $(Y = I\{F > 1.5,mg,L^{-1}\}$. A logistic calibration was fitted on ML using MI as the single predictor and evaluated on SA:

$$\Pr(Y = 1|MI) = \text{logit}^{-1}(\alpha + \beta\ MI), \tag{15}$$

reporting AUROC, Average Precision (PR-AUC), Brier score, calibration slope/intercept and a calibration curve (quantile binning). Two mechanistic baselines were compared by replacing MI with (i) undersaturation only and (ii) weathering only. A locked non-parametric isotonic mapping $p = \hat{f}(MI)$ was also learned on ML via cross-validated out-of-fold fits and then applied to SA; expected calibration error (ECE, quantile-binned) was computed as

$$\text{ECE} = \sum_{b=1}^{B} w_b\ |\hat{p}_b - \hat{o}_b|, \tag{16}$$

with w_b the empirical bin frequency, $\hat{p}_b$ the mean predicted probability, and $\hat{o}_b$ the observed event rate. As a sensitivity, a “recalibration-in-the-large” shift α^* was optionally obtained on SA and applied as $p_{adj} = \text{logit}^{-1}(\text{logit}(p) + \alpha^*)$. Using MI as a single screening covariate preserves parsimony and interpretability; isotonic calibration guarantees monotonicity without imposing a parametric link, appropriate for small-to-moderate samples.

(d) Spatial cross-validation on development data

To assess geographic robustness, ML-only screening performance was estimated under GroupKFold by community using MI as the feature in a logistic model, reporting fold-wise AUROC, PR-AUC and Brier score. Grouped folds prevent information leakage across communities and provide a conservative estimate of transportability.

2.7 Machine Learning Framework

A supervised framework comparing a Histogram-based Gradient Boosting Regressor (HistGradientBoosting), a Random Forest Regressor (RandomForest), Ridge and Lasso regularised linear models, an Extreme Gradient Boosting algorithm (XGBoost), and a Multilayer Perceptron Regressor (MLP) was implemented to predict groundwater fluoride concentrations and to isolate the dominant controlling factors. For the machine learning model development 152 archived samples (Bawku West, Garu, Gusheigu, Kintampo South, Saboba, Savelugu, Talensi, West Gonja, Zabzugu) was used and an a priori external set of 34 Karaga samples. The latter was never used for feature engineering, imputation, scaling, or model selection. Within the 152-sample corpus, (k)-fold cross-validation with $k = 5$ was adopted; all preprocessing was fit on each training fold and applied to its validation fold to prevent leakage. The external set was internally partitioned into a calibration subset (for screening-threshold selection) and a test subset for one-time evaluation. The statistical summary of the 152 archived dataset is presented in Table S2.

2.7.1 Feature Engineering

Only fluoride-blind predictors enumerated in the modelling dataset were used. The ML models were trained using a fluoride-blind predictor set comprising: field parameters (pH, electrical conductivity, temperature, and TDS/EC-derived salinity proxy where applicable); major cations (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}); major anions (HCO_3^-/alkalinity, Cl^-, SO_4^{2-}, NO_3^-); and selected geochemically informed, fluoride-blind PHREEQC-derived descriptors (e.g., ionic strength and saturation indices for calcite and magnesite, and salinity end-member indicators such as halite/sylvite where computed without using fluoride as an input feature). In addition, compositional balances (ilr coordinates) derived from the major-ion composition were included where specified in the CoDA workflow. Variables directly computed from fluoride (F) or that embed measured fluoride (e.g., fluoride activity or fluorite SI computed using measured F)

were excluded from ML predictors to prevent target leakage. Any column directly derived from fluoride or the analyte (F) was excluded to avoid target leakage. Continuous features were sanitised by replacing $\pm\infty$ with NaN, followed by median imputation and standardisation to zero mean and unit variance. Relevance screening employed mutual information with a select K best operator; the score for a feature (X_j) was

$$I(X_j;Y) = \iint p(x_j,y)\log\left(\frac{p(x_j,y)}{p(x_j),p(y)}\right)dx_j dy, \tag{17}$$

and the retained dimensionality k was tuned within cross-validation.

2.7.2 Target Variable Transformation

A transformed-target regressor was used with a monotone log-link to stabilise heteroscedasticity: $z = \log(1 + y)$, $\hat{y} = \exp(\text{clip}(\hat{z}, l, u)) - 1$. Here l ,u are log-scale bounds implied by plausible fluoride limits; clipping prevents explosive back-transforms while preserving order and scale.

2.7.3 Model Selection

The candidate set as mentioned in, each base learner f_θ, out-of-fold predictions within 5-fold CV yielded the primary objective

$$\theta^\star = \arg\max_\theta \; R^2_{CV}(\theta) \tag{18}$$

In parallel, a screening classifier estimated exceedance probability $p = \Pr(Y > 1.5 \mid X)$ using logistic regression with class weighting. The decision threshold τ was selected on the Karaga calibration subset to achieve a target sensitivity (default (0.95)), i.e. $\tau^\star = \arg\min_\tau |\,\text{sens}(\tau) - 0.95|$.

2.7.4 Modelling Pipeline Architecture

Let ($X \in R^{n\times p}$), ($y \in R^n$). The regression pipeline consisted of median imputation (I), standardisation (S), mutual-information selection ($M * k$), and a regressor (f_θ), wrapped by the target transform:

$$\hat{z} = f_\theta(,M_k(S(I(X)))),\ \hat{y} = \exp(\text{clip}(\hat{z},l,u)) - 1. \tag{19}$$

All operators ($I,S,M * k,f_\theta$) were fit within each training fold and applied to its validation fold; the selected model was refit on all 152 development samples and evaluated once on the Karaga test subset.

2.7.5 Hyperparameter Optimisation

Hyperparameter optimisation was performed using Optuna's Bayesian sampler within the inner cross-validation loop. Search spaces covered the selector width (k) and model-specific parameters (e.g., boosting learning rate/depth/iterations, forest size/depth, α for Ridge/Lasso, tree/regularisation terms for XGBoost, and hidden-layer code/α/learning rate for MLP). For each model, the best configuration identified in cross-validation was refit on the full 152-sample development set and evaluated once on the held-out Karaga test subset under strict externality (no feature fitting or calibration on SA). Full optimisation details (search spaces and objective definition) are provided in Supplementary Tables (Table S1).

2.7.6 Model Evaluation Strategy

Model performance was estimated using a nested cross-validation design to avoid optimistic bias during tuning. The outer loop used 5-fold stratified cross-validation repeated five times (stratification via quantile-binned fluoride concentrations), while the inner loop used 5-fold cross-validation for hyperparameter optimisation. Performance was summarised using R^2, RMSE, and MAE, and the final selected model was refit on all development samples prior to one-time evaluation on the Karaga test subset. Full nested-CV specification and metric equations are provided in Supplementary Methods (Equation 36 to 38).

2.7.7 WHO Guideline Exceedance Classification

Beyond regression performance, the models were evaluated on their ability to correctly classify samples as exceeding or not exceeding the WHO guideline value for fluoride (1.5 mg/L). This binary classification performance was assessed using: (i) Confusion Matrix: A tabulation of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), (ii) Classification Metrics: Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$, Precision: $\frac{TP}{TP+FP}$,

Recall (Sensitivity): $\frac{TP}{TP+FN}$, F1 Score: $2 \times \frac{Precision \times Recall}{Precision+Recall}$, and Specificity: $\frac{TN}{TN+FP}$, and (iii) Receiver Operating Characteristic (ROC) Curve Analysis: The area under the ROC curve (AUC) was calculated to assess the model's ability to discriminate between exceedance and non-exceedance cases.

2.7.8 Learning Curve Analysis

Learning curves were generated for the best-performing model to assess how model performance (R^2) varied with training set size. This analysis helped identify whether the model would benefit from additional data collection or was already capturing the maximum predictable variance:

$$\text{Learning Curve} = \{\left(n_i, R^2_{train}(n_i), R^2_{validation}(n_i)\right)\}_{i=1}^{k} \quad (20)$$

where n_i is the i-th training set size, $R^2_{train}(n_i)$ is the R^2 score on the training set of size n_i, and $R^2_{validation}(n_i)$ is the R^2 score on the validation set when trained with n_i samples.

2.8 Model Interpretability

Model interpretability techniques were applied to extract meaningful insights from the machine learning models, translating complex patterns into actionable hydrogeochemical knowledge. These methods enabled us to identify key drivers of fluoride occurrence, quantify their relative importance, and understand their marginal effects on fluoride concentrations.

2.8.1 Feature Importance Analysis

Feature importance analysis was conducted to identify the most influential predictors in the fluoride prediction best model. The best model for this study base on the model evaluation strategy was Random Forest; hence the built-in Feature Importance was use. This measure quantifies the total reduction in impurity (variance) attributed to each feature across all trees in the ensemble:

$$\text{Importance}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t \in T_m} \Delta i(s_t, j) \quad (21)$$

where M is the number of trees, T_m is the set of nodes in tree m, $\Delta i(s_t,j)$ is the impurity reduction at node s_t when splitting on feature j, defined as: $\Delta i(s_t,j) = i(s_t) - p_L \cdot i(s_{t_L}) - p_R \cdot i(s_{t_R})$, with $i(s_t)$ being the impurity at node s_t, p_L and p_R the proportion of samples going to the left and right child nodes, respectively, and $i(s_{t_L})$ and $i(s_{t_R})$ the impurities of the left and right child nodes.

2.8.2 SHAP Analysis

SHapley Additive exPlanations (SHAP) analysis was employed to provide a unified, theoretically-grounded approach to model interpretation based on cooperative game theory (Salih et al., 2024). SHAP values represent the contribution of each feature to the prediction for each individual sample, considering all possible feature combinations.

For a given sample, the SHAP value for feature j is defined as:

$$\phi_j = \sum_{S \subseteq F \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{j\}) - f_x(S)] \quad (22)$$

where F is the set of all features, S is a subset of features excluding j, $f_x(S)$ is the expected model output when only the features in subset S are known, and $f_x(S \cup \{j\})$ is the expected output when feature j is also known.

3 Results

3.1 Hydrogeochemical Characterization

3.1.1 Descriptive Statistics

The hydrogeochemical analysis of groundwater samples from the Karaga District revealed distinct patterns in major ion concentrations and physicochemical parameters. According to Table 2, the groundwater samples exhibited generally alkaline conditions with pH values ranging from 6.00 to 8.33 (mean = 7.57 ± 0.55), with only 5.9% of samples falling outside the WHO recommended range of 6.5-8.5. Electrical conductivity

varied substantially across the study area, ranging from 66.60 to 9,890 µS/cm (mean = 1,062 ± 1,632 µS/cm). Total dissolved solids (TDS) showed similar variability, spanning from 46.62 to 4,950 mg/L (mean = 618.1 ± 829.1 mg/L). The major cation chemistry was dominated by sodium, with concentrations ranging from 1.19 to 2,144 mg/L (mean = 183.6 ± 368.2 mg/L), and 17.6% of samples exceeding the WHO guideline of 200 mg/L. Calcium concentrations were generally low, varying from 0.001 to 125.6 mg/L (mean = 7.98 ± 22.52 mg/L). Magnesium levels were similarly modest, ranging from 0.001 to 46.80 mg/L (mean = 4.19 ± 8.15 mg/L). Potassium showed considerable variation (0.60-195.0 mg/L; mean = 25.27 ± 57.45 mg/L). Among the major anions, bicarbonate exhibited the highest concentrations, ranging from 19.70 to 520.0 mg/L (mean = 310.4 ± 143.5 mg/L). Chloride concentrations varied from 0.26 to 3,520 mg/L (mean = 168.1 ± 619.5 mg/L), with 8.8% surpassing WHO aesthetic guideline standards. Sulfate levels were generally low (0.001-26.70 mg/L; mean = 4.69 ± 5.40 mg/L), while nitrate concentrations ranged from 0.001 to 55.00 mg/L (mean = 7.84 ± 12.27 mg/L), with 2.9 % exceeding WHO heath-based guideline limits. In general, most parameter examine in groundwater sampled revealed right-skewed patterns, with numerous outliers indicating significant hydrogeochemical heterogeneity across the study area (Fig. S1). Fluoride concentrations showed substantial spatial variability, ranging from 0.07 to 6.04 mg/L with a mean of 1.34 ± 1.31 mg/L and median of 1.03 mg/L. Notably, 17.6% of groundwater samples exceeded the WHO drinking water guideline of 1.5 mg/L, indicating a significant public health concern. The distribution of fluoride concentrations (Fig. S2) was observed to be right-skewed, with most of the samples clustering between 0.2 and 1.2 mg/L, while several extreme outliers reached concentrations of 4-6 mg/L. The distribution exhibited a long tail extending toward higher concentrations, suggesting localized zones of severe fluoride enrichment likely associated with specific geochemical processes or hydrogeological conditions within the aquifer system. Spatial maps of selected hydrochemical parameters and sampling locations are provided in Fig. S3. Potassium is highest in the northwestern

sector (Fig. S3a), whereas nitrate and sodium show elevated values mainly in the southern to southwestern sector (Fig. S3b–c). Fluoride hotspots occur in the central, southern, and northwestern parts of the district, including Tong, Tamaligu, Nyong Kuma, Bagurugu Fulaniyili, Nyong Yapalsi, and Sung Dry Season Garden (0.07–6.04 mg/L). Elevated NO_3^- and K^+ may reflect agricultural inputs, whereas elevated F^- is plausibly driven by a combination of silicate weathering and mineral dissolution processes, including fluorite dissolution.

3.1.2 Water Types and Facies

The hydrochemical classification based on isometric log-ratio (ILR) transformation revealed distinct water type distributions across the Karaga District groundwater system. As presented in Fig. 2(a)., and Table S3, the dominant water type was Na-HCO_3, comprising 64.7% of all samples (22 out of 34), with fluoride concentrations ranging from 0.07 to 1.36 mg/L (mean = 0.83 ± 0.39 mg/L). This prevalence reflects the dominant influence of silicate weathering processes releasing sodium ions and carbonate buffering producing bicarbonate ions, characteristic of crystalline rock aquifers where silicate minerals dominate the water-rock interaction. K-HCO_3 waters represented the second most common facies (14.7% of samples), with notably higher fluoride concentrations ranging from 1.00 to 4.70 mg/L (mean = 2.54 ± 1.56 mg/L). This water type likely indicates areas influenced by fertilizer inputs or K-feldspar weathering processes. Na-Cl waters comprised 8.8% of samples and exhibited the highest fluoride concentrations, ranging from 1.40 to 6.04 mg/L (mean = 3.75 ± 2.32 mg/L), reflecting halite or evaporite dissolution and potential saline intrusion processes. The Na-HCO_3-Cl mixed type (5.9% of samples) showed intermediate characteristics with fluoride levels from 0.24 to 1.56 mg/L (mean = 0.90 ± 0.93 mg/L), indicating hydrochemical mixing between silicate weathering and evaporitic influences. Minor water types included Ca-Mg-HCO_3 (2.9% of samples) with the lowest fluoride

concentration (0.30 mg/L), representing classic carbonate-dominated waters, and Mg-Na-HCO_3-Cl (2.9% of samples) with moderate fluoride levels (1.33 mg/L), suggesting combined dolomite weathering, silicate alteration, and evaporite influences.

The Gibbs diagram (Gibbs, 1970) analysis (Fig. 2(b)) revealed that most groundwater samples plotted within the rock-weathering dominance field, characterised by intermediate TDS (100-1000 mg/L) and mid-range ionic fractions (0.2-0.8). However, a significant subset shifted into the evaporation-dominance zone, marked by high TDS (>1000 mg/L) and elevated Na^+ or Cl^- fractions (>0.8). Importantly, higher fluoride concentrations (warm colours in Fig. 2(b)) clustered predominantly in the evaporation domain, suggesting that evaporative concentration not only increases total dissolved solids but also enhances fluoride mobilization.

3.1.3 Correlation Analysis

Spearman rank correlation analysis revealed significant relationships between fluoride and various hydrogeochemical parameters, providing insights into the controlling mechanisms of fluoride mobilization. As shown in Fig. S4, fluoride exhibited strong positive correlations with several key parameters. The strongest correlations were observed with geochemically derived parameters: free fluoride molal concentration (r = 1.00), fluoride activity (r = 1.00), and saturation index of fluorite (r = 0.75), confirming the fundamental role of fluorite mineral equilibrium in controlling fluoride concentrations. Among the directly measured parameters, fluoride showed significant positive correlations with electrical conductivity (r = 0.56), ionic strength (r = 0.51), and total dissolved solids (r = 0.46), indicating that fluoride enrichment is associated with higher salinity conditions. The strong correlation with the saturation index of sylvite (r = 0.58) and halite (r = 0.21) further supports the influence of evaporite dissolution processes on fluoride mobilization. Nitrate exhibited a moderate positive correlation with fluoride (r = 0.34), suggesting that areas with agricultural influence

may experience enhanced fluoride release, possibly due to deeper flushing of fluoride-bearing minerals under intensive recharge conditions. Fluoride demonstrated notable negative correlations with several parameters that indicate competitive processes. The HCO_3^-/F^- ratio showed a strong negative correlation (r = -0.77), reflecting the competitive inhibition of fluoride by bicarbonate ions for adsorption sites on mineral surfaces. The Ca^{2+}/F^- ratio exhibited a moderate negative correlation (r = -0.13), consistent with calcium-fluoride complexation processes that reduce free fluoride availability. The fluorite equilibrium distance showed a strong negative correlation (r = -0.75), confirming that samples closer to fluorite equilibrium tend to have higher fluoride concentrations. Also, in Fig. S4 important inter-ionic relationships can be noticed that provide context for fluoride behaviour. The very strong positive correlation between Na^+ and Cl^- (r = 0.78) indicates halite or evaporite dissolution as a dominant salinity control. The strong negative correlation between Na^+ and K^+ (r = -0.49) suggests ion-exchange processes, while the positive correlation between K^+ and Mg^{2+} (r = 0.55) implies that fertilizer-impacted areas also exhibit elevated magnesium from parallel weathering processes. Collectively, these relationships indicate that fluoride mobilization takes place within a complex hydrogeochemical system, governed by processes such as mineral dissolution, ion exchange, competitive adsorption, and evaporative concentration.

3.2 Compositional Data Analysis Findings

3.2.1 Sequential Binary Partition Results

The sequential binary partition (SBP) structure was constructed using a hierarchical clustering approach based on centred log-ratio (clr) transformed data, with one modification where the first partition was forced to separate cations from anions to reflect the fundamental electrochemical division in water chemistry. As detailed in Table 1, the optimal SBP structure revealed nine distinct binary balances (B1-B9) that

capture meaningful geochemical contrasts within the hydrochemical dataset. The first balance (B1) represents the overall electroneutrality contrast between total cations (Na^+, K^+, Ca^{2+}, Mg^{2+}) and total anions (Cl^-, SO_4^{2-}, HCO_3^-, NO_3^-, F^-), providing a fundamental charge balance indicator. Subsequent balances systematically partition specific ion groups: B2 contrasts fertilizer-derived potassium against nitrate contamination, B3 distinguishes dolomite/magnesium release from agricultural inputs (K^+, NO_3^-), while B4 separates silicate weathering sodium from gypsum/evaporite sulfate. Balance B5 contrasts calcite dissolution calcium against halite or saline intrusion chloride, and B6 represents carbonate buffering (HCO_3^-) opposed to hardness and salinity load (Ca^{2+}, Cl^-). The more complex partitions include B7, which contrasts evaporite Na-SO_4 sources against freshwater Ca-HCO_3 signatures, and B8, which isolates fluoride mobilization against the general salinity matrix. The final balance (B9) represents secondary silicate weathering plus agricultural inputs contrasted with background salinity.

3.2.2 ILR Transformation Results

The isometric log-ratio transformation of the nine binary partitions produced orthogonal coordinates that effectively captured the major sources of hydrochemical variability while addressing the compositional nature of the data. Table S4 presents the correlation patterns between ILR coordinates, revealing significant relationships that reflect underlying geochemical processes. The correlation matrix demonstrates both strong positive and negative associations between different balances, indicating coupled and competitive processes within the groundwater system. Key geochemical contrasts emerged from the ILR analysis. Balance ILR1 (overall charge balance) showed strong positive correlations with ILR4 ($r = 0.63$), ILR5 ($r = 0.54$), ILR7 ($r = 0.59$), and ILR9 ($r = 0.59$), indicating that cation-dominated waters tend to coincide with silicate weathering, calcite dissolution, and mixed evaporite/carbonate weathering signatures. Conversely, ILR1 exhibited negative correlations with ILR3 ($r = -0.47$) and ILR8 ($r = -0.41$), suggesting that overall cation dominance opposes dolomite weathering relative to agricultural inputs and fluoride

mobilization relative to bulk salinity. Fig. 3 illustrates the pairwise relationships among highly correlated ILR balances, revealing distinct hydrogeochemical end-members. The strong negative correlation between ILR3 (dolomite weathering vs agricultural inputs) and ILR5 (calcite vs halite, r = -0.74) and ILR9 (secondary weathering vs background, r = -0.75) demonstrates that geogenic carbonate weathering and agricultural/secondary weathering processes rarely co-dominate in the same samples. The positive correlation between ILR9 and ILR5 (r = 0.75) indicates that agricultural plus silicate weathering signals coincide with calcite dissolution, suggesting mixed carbonate-agricultural weathering end-members. The fluoride-specific balance (ILR8) showed important relationships with other processes. Its negative correlation with ILR1 (r = -0.41) suggests that fluoride release is strongest in waters with lower overall cation dominance, consistent with dilution control or decoupling from bulk weathering processes.

3.2.3 PCA Results

Principal component analysis of the ILR-transformed data was explored to identify the main sources of hydrogeochemical variability and their relationships to fluoride behaviour. The first four principal components captured 91.2% of the total variance in the system: PC1 explained 37.5%, PC2 explained 27.1%, PC3 explained 17.9%, and PC4 explained 8.7% of the variance. Fig. S5 presents the loadings of each ILR coordinate on the first four principal components, revealing distinct process-driven axes that characterize the groundwater system. PC1 represents a "Salinity & Weathering Intensity Axis" with strong positive loadings from ILR9 (0.62), ILR7 (0.38), ILR5 (0.27), ILR4 (0.21), and ILR1 (0.30), contrasted against negative loadings from ILR3 (-0.48) and ILR6 (-0.11). This axis captures a continuum from low-TDS, magnesium-carbonate-dominated waters (negative scores) to high-TDS, mixed evaporite/carbonate-weathering and agricultural waters (positive scores). PC2 functions as a "Fluoride

Mobilization Axis" with a dominant loading from ILR8 (0.88) and secondary negative loadings from ILR1 (-0.29) and ILR7 (-0.31), almost exclusively contrasting fluoride release against general salinity and evaporite signatures. PC3 operates as a "Carbonate Buffering vs Marine/Evaporite Axis" with strong positive loading from ILR6 (0.68) and strong negative loading from ILR7 (-0.54), differentiating CO_2-carbonate equilibrium control from marine/evaporite-dominated end-members. PC4 represents a "Carbonate/Evaporite vs Agricultural-Weathering Axis" with positive loadings from ILR7 (0.58) and ILR6 (0.47), contrasted against negative loadings from ILR9 (-0.50) and ILR8 (-0.50), forming a mixing axis between carbonate/evaporite-controlled waters and soil-weathering plus agricultural/fluoride-enriched water types. The biplot analysis (Fig. S6) reveals distinct sample clustering patterns related to these process axes. In the PC1 vs PC2 biplot, samples in the upper-right quadrant (high PC1, high PC2) combine high TDS/weathering with strong fluoride mobilization, while upper-left samples (low PC1, high PC2) represent low-salinity but fluoride-rich waters characteristic of geogenic fluoride release in dilute aquifers. The PC3 vs PC4 biplot further discriminates carbonate-buffered waters from evaporite-influenced and agricultural/weathering-impacted zones.

Conversely, correlation analysis between principal component scores and fluoride concentrations (Table S5) revealed weak and non-significant relationships across all principal components, with PC3 showing the strongest but still modest positive correlation (r = 0.210, $p > 0.05$), while PC1 (r = -0.133, $p > 0.05$), PC2 (r = -0.138, $p > 0.05$), and PC4 (r = 0.006, $p > 0.05$) demonstrated minimal associations with fluoride concentrations. These weak correlations suggest that fluoride behaviour is not strongly aligned with the major sources of hydrogeochemical variability captured by the principal components, indicating that fluoride mobilization operates through more complex, localized processes that are not adequately represented by the dominant regional-scale geochemical patterns identified through compositional data analysis.

3.3 Geochemical Modelling Results

3.3.1 Speciation Results

The PHREEQC geochemical modelling revealed distinct patterns in fluoride speciation and calcium-fluoride complexation across the groundwater samples. As presented in Table 2, free fluoride ions (F^-) dominated the speciation, accounting for 99.4% to 100% of total dissolved fluoride (mean = 99.9 ± 0.13%). The calcium-fluoride complex (CaF^+) represented only a minor fraction, ranging from 0% to 0.623% (mean = 0.05 ± 0.13%) of total fluoride speciation. This predominance of free fluoride indicates that complexation with calcium plays a relatively minor role in fluoride sequestration under the prevailing hydrogeochemical conditions in the Karaga District groundwater system. The activity coefficients and effective concentrations showed considerable variability across the study area. Fluoride activity ranged from 0.07 to 5.28 mol/L (mean = 1.21 ± 1.16 mol/L), while calcium activity varied dramatically from 0.001 to 47.3 mol/L (mean = 4.35 ± 9.74 mol/L). The ionic strength of the groundwater samples ranged from 0.001 to 0.10 mol/L (mean = 0.01 ± 0.02 mol/L), reflecting the variable salinity conditions across the aquifer system. These wide ranges in activities and ionic strength indicate significant spatial heterogeneity in the thermodynamic controls on fluoride behaviour. The low degree of calcium-fluoride complexation can be attributed to the generally low calcium concentrations observed in most samples, where calcium levels ranged from 0.001 to 125.6 mg/L with a mean of only 7.98 mg/L. Under these conditions, insufficient calcium is available to form significant amounts of CaF^+ complexes, allowing fluoride to remain predominantly as free F^- ions.

3.3.2 Saturation Indices

The saturation index calculations provided crucial insights into the thermodynamic controls on fluoride mobility and mineral equilibria within the groundwater system. Table 3 shows that fluorite saturation indices

ranged from -8.48 to -1.09 (mean = -3.37 ± 1.51), indicating that all groundwater samples were undersaturated with respect to fluorite (CaF_2). This widespread undersaturation suggests that fluorite dissolution is thermodynamically favoured throughout the study area, providing a potential source of fluoride to the groundwater. The relationship between fluorite saturation indices and fluoride concentration showed a strong positive correlation (r = 0.75, Fig. S4), confirming that samples closer to fluorite equilibrium tend to have higher fluoride concentrations. Other mineral saturation indices revealed the broader geochemical context influencing fluoride behaviour. Calcite saturation indices ranged from -6.98 to 0.21 (mean = -1.4 ± 1.5), with most samples being undersaturated, promoting calcium carbonate dissolution that could compete with fluoride for calcium ions. Dolomite showed similar patterns with saturation indices from -12.6 to 2.01 (mean = -1.29 ± 2.74). Gypsum was consistently undersaturated across all samples (-8.42 to -3.12; mean = -5.17 ± 1.3), indicating limited sulfate mineral control on the system chemistry. The evaporite minerals showed significant undersaturation, with sylvite saturation indices ranging from -9.92 to -5.91 (mean = -8.31 ± 0.72) and halite from -11.1 to -3.83 (mean = -7.65 ± 1.67). Notably, the sylvite saturation index exhibited a strong positive correlation with fluoride (r = 0.58, Fig. S4), suggesting that potassium-bearing mineral dissolution processes may be coupled with fluoride mobilization. The consistent undersaturation with respect to fluorite, combined with variable degrees of undersaturation for other minerals, indicates that the groundwater system is actively dissolving multiple mineral phases, with fluorite dissolution being a primary control on fluoride concentrations while other mineral equilibria modulate the overall ionic strength and competitive ion effects.

3.3.3 Fluoride Mobility Index

The mechanistic Mobility Index (MI) was developed to rank groundwater fluoride mobilization propensity using four fluoride-independent

components as outline in section 2.6. The entropy-based weighting assigned the highest importance to the fluoride activity component (44.20%), followed by fluorite undersaturation (23.90%), weathering ILR coordinate (16.00%), and pH (15.90%), as shown in Table S6. Components were scaled using robust quantile transformation and combined through linear convex aggregation to preserve mechanistic interpretability while avoiding target leakage. Construct validity was confirmed through significant correlations between MI and theoretical components. The weathering coordinate showed positive associations on both ML (ρ = 0.404, p = 0.0001) and study area (SA) datasets (ρ = 0.46, p = 0.007), consistent with silicate weathering enhancing fluoride mobility. Directional validation through median quantile regression demonstrated the expected negative relationship between MI and calcium-to-fluoride ratio (β_1 = -0.052, p = 0.000601), supporting fluorite control mechanisms. Criterion validity for WHO exceedance screening yielded exceptional performance with AUROC of 0.976 on the SA dataset, substantially outperforming individual components (undersaturation only: 0.792; weathering only: 0.607). Spatial cross-validation by community maintained robust discrimination (AUROC = 0.828 ± 0.069), suggesting that the model retains robust performance within the development dataset under community-grouped cross-validation. However, transfer to the SA dataset revealed calibration mismatch requiring intercept adjustment (α^* = -1.601) to correct baseline risk while preserving discrimination, as illustrated in Fig. 4. The MI effectively identified waters trending toward Na-HCO_3 facies with fluorite undersaturation as having elevated fluoride mobility potential.

3.4 Machine Learning Model Performance

3.4.1 Model Comparison

The supervised machine learning framework compared six algorithms using nested cross-validation on 152 development samples with external validation on 34 Karaga samples. Performance evaluation employed

coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE) metrics across 5-fold cross-validation. Models were trained and evaluated using the predictor set listed in Methods (section 2.7.1), ensuring the external Karaga dataset was not used in any stage of feature engineering, scaling, or model selection. The best model was the Multilayer Perceptron with R^2 of 0.668±0.189 and the lowest MAE of 0.654±0.141. Followed by the non-linear ensemble methods, Histogram-based Gradient Boosting (HGB) achieving R^2 of 0.648±0.208, MAE of 0.747±0.396, and RMSE of 1.436±0.968. Random Forest followed closely with R^2 of 0.586±0.156, while XGBoost showed moderate performance (R^2 = 0.350±0.090). Linear models performed poorly, with Ridge regression yielding negative R^2 values (-6.064±12.422) indicating substantial overfitting. Hyperparameter optimization using Optuna's Bayesian sampler with 150 trials demonstrated convergence by 40-50 trials, confirming adequate model tuning as illustrated in Fig. 5c.

3.4.2 WHO Guideline Exceedance Prediction

Binary classification performance for WHO guideline exceedance (1.5 mg/L) revealed excellent discrimination capability with area under the receiver operating characteristic curve (AUC) of approximately 0.94 on the external test set, as shown in Fig. 5e. The confusion matrix analysis on 34 external Karaga samples indicated 76.5% accuracy with high specificity (82.1%) and negative predictive value (88.5%), effectively ruling out exceedance cases. However, sensitivity was limited at 50.0%, missing half of true exceedances (3/6 cases), indicating a conservative decision threshold with prevalence of 17.6%. Classification metrics showed precision of 37.5%, F1-score of 42.9%, balanced accuracy of 66.1%, and Matthews correlation coefficient of 0.289. The results suggested that for operational screening where false negatives are costly, sensitivity could be enhanced by lowering the probability threshold or implementing class-weighted training strategies while monitoring the trade-off in false positives.

3.4.3 Feature Importance Results

Feature importance analysis using the Random Forest model revealed that total dissolved solids (TDS) dominated predictions with an importance score of 0.328, followed by ionic strength (0.143), pH (0.140), and saturation indices for calcite (0.139) and magnesite (0.135), as documented in Table 5. Secondary contributors included halite saturation index (0.113), electrical conductivity (0.094), and sylvite saturation index (0.076), while individual ion concentrations showed relatively minor importance. Chloride contributed 0.071 and bicarbonate only 0.035 to model predictions. This ranking demonstrated that bulk salinity parameters and carbonate equilibrium indicators were the primary controls on fluoride concentrations, with specific ion concentrations playing secondary roles. The dominance of TDS and ionic strength reflected the model's emphasis on overall water mineralization as the key driver of fluoride mobility, while carbonate saturation indices captured competitive effects on calcium availability that influence fluorite solubility.

3.4.4 SHAP Analysis Results

SHAP analysis on the best model (MLP) provided mechanistic insights into feature contributions, revealing that TDS and ionic strength were the dominant global predictors with mean absolute SHAP values indicating primary influence on model predictions, as shown in Fig. 6a. The beeswarm plot demonstrated that high TDS, ionic strength, and pH values consistently pushed fluoride predictions upward, while low values depressed them. Calcite and magnesite saturation indices showed positive SHAP contributions when approaching saturation, consistent with carbonate precipitation reducing free calcium and favouring fluoride persistence. SHAP dependence plots revealed non-linear relationships, with calcite saturation index showing negative contributions when undersaturated (SI < -2.5) and positive effects near saturation, while pH exhibited a monotonic positive effect on fluoride predictions (Fig. 7). Ionic strength and TDS displayed threshold effects, with minimal influence at low concentrations but strong positive impacts at elevated levels,

particularly when coupled with high pH. Waterfall plots for representative samples illustrated how individual geochemical features combined to drive predictions, with concentrated, alkaline waters (SA-KAR28) showing elevated fluoride potential while dilute, acidic conditions (SA-KAR3) resulted in suppressed predictions, as demonstrated in Fig. 8.

4 Discussion

4.1 Hydrogeochemical Controls on Fluoride Enrichment

4.1.1 Weathering and Water-Rock Interactions

The hydrogeochemical analysis indicates that fluoride enrichment in the groundwater of the Karaga District is primarily governed by water–rock interactions, especially silicate weathering processes within the Voltaian Supergroup formations. As outlined earlier (Section ..., Fig. 1), the study area is underlain mainly by mudstone–sandstone rich units belonging to the Oti/Pendjari Group and the Tamale/Obosum beds. The predominance of the Na–HCO_3 water type (64.7%) reflects silicate weathering (Yadav et al., 2023) and natural groundwater softening through cation exchange, which lowers Ca^{2+} concentrations. This reduction in Ca^{2+} activity promotes fluorite undersaturation, thereby enhancing fluoride mobilisation. The highest fluoride concentrations occur in Na–Cl waters, suggesting a saline end-member influence likely linked to evaporite dissolution, further suppresses calcium availability and increases ionic strength, intensifying fluoride persistence. Overall, these hydrochemical patterns point to lithological heterogeneity (variations in mudstone–sandstone ratios, evaporitic layers, limestones, and conglomerates) as a major control on the geochemical pathways that drive extreme fluoride levels in the area (Sunkari et al., 2024).

Evidence for residence time effects on fluoride mobilization is reflected in the relationship between salinity indicators and fluoride enrichment. The Gibbs diagram analysis (Fig. 2b) revealed that higher fluoride concentrations clustered predominantly in the evaporation domain,

suggesting that evaporative concentration not only increases total dissolved solids but also enhances fluoride mobilization (Hossain et al., 2016). This pattern indicates that longer residence times, allowing for both progressive mineral dissolution and evaporative concentration, are crucial for achieving elevated fluoride concentrations. The spatial heterogeneity observed in fluoride distributions, with concentrations ranging from 0.07 to 6.04 mg/L with a mean of 1.34 ± 1.31 mg/L (Table 2), suggests variable flow paths and residence times across the aquifer system (Shaji et al., 2024). Comparative analysis of fluoride concentrations in other regions highlights the severity of the issue in the Karaga District and the role of geological setting in controlling fluoride mobilization. In the Nubian Sandstone Aquifer System of North Africa, which shares similarities with the Voltaian Supergroup in terms of sedimentary rock dominance fluoride concentrations range from 0.3 to 2.5 mg/L, with higher levels attributed to groundwater circulation in deeper, confined portions of the aquifer (Mosaad et al., 2022). The Karaga District's fluoride levels (0.07 to 6.04 mg/L) extend to even higher concentrations, potentially reflecting more intensive weathering or localized mineral enrichment. In the volcanic-hosted Main Ethiopian Rift aquifer system, fluoride concentrations up to 68 mg/L have been reported, with the highest levels associated with rhyolitic and basaltic lava flow dissolution (Rango et al., 2012). While the Karaga District's fluoride levels are lower than these extreme values, they still exceed the concentrations found in many other sedimentary aquifers worldwide, such as the Rio Cuarto sedimentary aquifer in Argentina, where fluoride values ranged from 0.12 to 0.6 mg/L, attributed to Ca–HCO_3-type groundwater with high calcium content that suppresses fluoride mobility through CaF_2 precipitation (Blarasin et al., 2018).

4.1.2 Geochemical Mechanisms of Fluoride Mobilization

4.1.3 Water Type Influence

The relationship between hydrochemical facies and fluoride enrichment reveals distinct evolutionary pathways for fluoride mobilization across

different water types. The Na-HCO_3 water type, while dominant, showed relatively modest fluoride concentrations with fluoride concentrations (Table S3), suggesting that basic silicate weathering processes alone are insufficient to generate severe fluoride contamination. In contrast, Na-Cl waters comprised 8.8% of samples and exhibited the highest fluoride concentrations, ranging from 1.40 to 6.04 mg/L (mean = 3.75 ± 2.32 mg/L), likely driven by evaporite dissolution and cation exchange that reduces calcium and enhances fluoride mobility (Rena et al., 2022; Saini et al., 2023).. This pattern demonstrates that the evolution toward more saline water types facilitates enhanced fluoride mobilization.

Ion exchange processes significantly impact fluoride behaviour, as evidenced by the compositional data analysis results. The strong negative correlation between Na^+ and K^+ (r = -0.49) suggests ion-exchange processes, while the positive correlation between K^+ and Mg^{2+} (r = 0.55) implies that fertilizer-impacted areas also exhibit elevated magnesium from parallel weathering processes (Fig. 3). The intermediate K-HCO_3 water type shows elevated fluoride concentrations with fluoride concentrations ranging from 1.00 to 4.70 mg/L (mean = 2.54 ± 1.56 mg/L), likely indicating areas influenced by fertilizer inputs or K-feldspar weathering processes (Tyagi & Sarma, 2021). This suggests that ion exchange processes involving potassium release may create geochemical conditions favourable for enhanced fluoride mobilization.

The evolution of water chemistry along flow paths is clearly demonstrated by the principal component analysis results. PC1 represents a "Salinity & Weathering Intensity Axis" capturing a continuum from low-TDS, magnesium-carbonate-dominated waters (negative scores) to high-TDS, mixed evaporite/carbonate-weathering and agricultural waters (positive scores). The biplot analysis reveals that samples in the upper-right quadrant (high PC1, high PC2) combine high TDS/weathering with strong fluoride mobilization, while upper-left samples (low PC1, high PC2) represent low-salinity but fluoride-rich waters characteristic of geogenic fluoride release in dilute aquifers. This spatial organization suggests that

different flow path evolutionary stages create distinct hydrogeochemical environments, with both fresh geogenic waters and evolved saline waters capable of supporting elevated fluoride concentrations through different mechanistic pathways.

The mixed water types provide evidence for hydrogeochemical mixing processes that influence fluoride behaviour. The Na-HCO_3-Cl mixed type (5.9% of samples) showed intermediate characteristics with fluoride levels from 0.24 to 1.56 mg/L (mean = 0.90 ± 0.93 mg/L) (Table S3), indicating hydrochemical mixing between silicate weathering and evaporitic influences. This intermediate behaviour suggests that the transition between different water types may create transient geochemical conditions that either enhance or suppress fluoride mobilization, depending on the specific mixing ratios and competitive ion effects operating within the system.

4.2 Evaluation of Machine Learning Approaches

4.2.1 Comparative Performance of Models

The multilayer perceptron achieved the highest cross-validated performance (R^2 = 0.668 ± 0.189; MAE = 0.654 ± 0.141), indicating moderate explanatory power given the limited sample size and the heteroscedastic, tail-heavy fluoride distribution. Rather than relying solely on predictive accuracy, the principal contribution of this framework is that it constrains modelling to fluoride-blind, geochemically interpretable predictors, quantifies mechanistic controls via SHAP/feature importance, and converts these controls into an operational screening tool (the Mobility Index) with strong discrimination for WHO exceedance. This finding aligns with several studies conducted in India (De et al., 2023) Pakistan (Rashid et al., 2020) and China (Tian et al., 2025), which have similarly demonstrated the nonlinear nature of factors influencing fluoride levels in groundwater. Also, non-linear ensemble methods, specifically histogram-based gradient boosting and random forest, provided strong secondary performance, with histogram-based gradient boosting

achieving R^2 of 0.648±0.208 and random forest yielding R^2 of 0.586±0.156. In contrast, linear regression models performed poorly, with ridge and lasso yielding negative R^2 values that indicate substantial overfitting and failure to generalize across validation folds. XGBoost demonstrated moderate performance with R^2 of 0.350±0.090. This stark performance hierarchy reveals a fundamental characteristic of fluoride behaviour in hydrogeochemical systems: the relationship between fluoride concentrations and predictive variables is inherently nonlinear and cannot be adequately captured by linear parameterizations. Model comparison results are reported under the nested cross-validation and Optuna tuning protocol described in Methods (Section 2.7), with full optimisation specifications provided in Supplementary Methods S-ML1. Optuna convergence behaviour is illustrated in Fig. 5c, and the resulting best-parameter configurations are summarised in Table 4. Feature importance analysis revealed that total dissolved solids dominated predictions with an importance score of 0.328, reflecting salinity control on fluoride mobility. pH (0.140) and saturation indices for calcite (0.139) and magnesite (0.135) captured the thermodynamic controls governing fluoride speciation and mineral dissolution. Individual ion concentrations, including chloride (0.071) and bicarbonate (0.035), showed secondary importance. This ranking validates our mechanistic understanding that bulk water properties and carbonate equilibrium are more influential than specific ion concentrations, emphasizing that fluoride mobilization is regulated by overall mineralization and pH-driven changes in calcium availability. The learning curve analysis indicated a high-variance regime in which validation R^2 improves markedly with training set size but remains consistently below the training R^2. This gap suggests that gains in predictive accuracy would likely emerge from collecting additional data and employing tail-aware or heteroscedastic modelling strategies to better capture extreme fluoride levels. Residuals showed increasing variance with fitted values and a slight negative bias at the upper range, reflecting underprediction of rare, high-fluoride samples. These patterns indicate that while the model generalizes reasonably well across the typical

fluoride concentration range, its ability to predict extreme values remains limited by both data scarcity and inherent model limitations in handling the heteroscedastic noise structure in high-fluoride regimes.

4.2.2 Integration of Geochemical and Machine Learning Insights

Single-method approaches often fail to capture the nonlinear and multivariate nature of fluoride mobilisation in heterogeneous hydrogeochemical systems. Thermodynamic models are mechanistically grounded but depend on reliable mineralogical constraints and input chemistry, while purely statistical or geostatistical approaches may oversimplify spatial heterogeneity. In this study, we integrate PHREEQC-derived thermodynamic descriptors and compositional balances (CoDA) with machine-learning pattern recognition to connect predictive signals to chemically interpretable processes.

Interpretability analyses show that the model relies primarily on bulk mineralisation and carbonate-equilibrium indicators rather than any single dissolved ion. SHAP patterns indicate that higher total salinity (TDS) and ionic strength, together with higher pH and carbonate saturation behaviour (e.g., SI calcite and SI magnesite), consistently push fluoride predictions upward, consistent with geochemical controls on calcium activity and fluorite solubility. Dependence plots further suggest nonlinear threshold behaviour, with carbonate-equilibrium indicators switching influence as waters evolve from clearly undersaturated toward near-saturation conditions, consistent with carbonate precipitation reducing free Ca^{2+} and favouring fluoride persistence (Ayub et al., 2024). Sample-level predictions through waterfall plots demonstrated mechanistic coherence: concentrated, alkaline waters produced elevated predictions through elevated Mg^{2+} and positive magnesite saturation, while dilute, acidic waters remained suppressed due to low calcite saturation and low TDS (Paikaray & Mahajan, 2023; Lone et al., 2024).

These model-derived patterns align with hydrochemical evolution observed in the Karaga dataset. The dominance of Na–HCO_3 waters

(64.7%) reflects silicate weathering and base-exchange processes typical of Voltaian Supergroup settings, whereas the Na–Cl facies (8.8%) exhibits the highest fluoride concentrations (1.40–6.04 mg/L), indicating that evolution toward more saline waters through saline mixing and/or evaporite-related inputs coupled with cation exchange corresponds to conditions that favour enhanced fluoride mobilisation. The intermediate K–HCO_3 facies also shows elevated fluoride (mean 2.54 ± 1.56 mg/L), consistent with ion-exchange–driven geochemical shifts that can support fluoride liberation. Overall, the convergence between facies evolution and ML interpretability indicates that the model is encoding chemically coherent processes rather than spurious correlations.

4.2.3 Critical Thresholds and Nonlinear Relationships

Fluorite undersaturation emerges as the critical mineral control governing fluoride mobilization, with a strong positive correlation of r = 0.75 between fluorite saturation indices and measured fluoride concentrations. All groundwater samples in the study area were undersaturated with respect to fluorite (SI ranging from -8.48 to -1.09, mean = -3.37 ± 1.51), confirming that fluorite dissolution is thermodynamically favoured throughout the Karaga District and provides the dominant fluoride source. This widespread undersaturation demonstrates that only waters positioned near the fluorite equilibrium boundary maximize fluoride concentrations while remaining in the dissolution regime. Calcite saturation showed threshold behaviour, with SHAP dependence analysis revealing negative contributions when clearly undersaturated (SI < -2.5) and positive effects approaching saturation (SI ≥ 0), mechanistically linked to $CaCO_3$ precipitation reducing free calcium and favouring fluoride persistence through reduced fluorite formation. pH exhibited a monotonic positive relationship across the full measured range, consistent with surface deprotonation and alkaline desorption mechanisms that mobilize fluoride from mineral surfaces. Salinity variables displayed critical synergistic thresholds: TDS showed minimal influence below ~1000 mg/L but sharp positive effects at elevated concentrations, particularly when

coupled with pH > 7.5. Ionic strength demonstrated similar behaviour with threshold effects near 0.05 mol/L, with tail effects revealing suppression at extreme levels coupled with low pH due to charge shielding mechanisms (Paikaray & Mahajan, 2023; Ayub et al., 2024). Water-type evolution toward Na-HCO_3 facies comprising 64.7% of samples with mean fluoride 0.83 mg/L created high-risk hydrogeochemical conditions, while the highest fluoride concentrations (1.40–6.04 mg/L) occurred in Na-Cl waters comprising only 8.8% of samples, demonstrating compositional transitions marking critical thresholds for fluoride mobilization.

4.3 Public Health and Management Implications

4.3.1 Risk Assessment Framework

The binary classification model provides a quantitative foundation for risk stratification, achieving excellent discrimination (AUC = 0.94) but with diagnostic trade-offs relevant for operational deployment. On external Karaga samples, the classifier demonstrated 76.5% accuracy with high specificity (82.1%) and negative predictive value (88.5%), effectively ruling out exceedance cases. However, sensitivity was limited at 50.0%, missing half of true exceedances, indicating a conservative decision threshold with prevalence of 17.6%. This performance profile suits the model for first-pass screening to identify wells unlikely to exceed the WHO standard of 1.5 mg/L, with sensitivity enhancement possible through lowering the probability threshold or implementing class-weighted training strategies while monitoring false positive trade-offs. Geochemical characterization reveals actionable risk indicators for targeted intervention. High-risk communities exhibit Na-HCO_3 or Na-Cl water types with high TDS (>1000 mg/L), pH >7.5, and low calcium concentrations (<20 mg/L), particularly in locations like Tong, Tamaligu, Nyong Kuma, and Bagurugu Fulaniyili where fluoride concentrations ranged from 0.07 to 6.04 mg/L. Conversely, communities with Ca-Mg-HCO_3 water types, low TDS (<300 mg/L), and circumneutral pH represent lower-risk

hydrogeochemical signatures. The Mobility Index framework demonstrates robust geographic transferability for expanded deployment. Spatial cross-validation by community-maintained discrimination (AUROC = 0.828 ± 0.069), with the standard deviation indicating robust transferability. Transfer to external areas required intercept adjustment ($\alpha^* = -1.601$) to correct baseline risk while preserving discrimination, enabling operationalization in new communities through minimal local parameterization without compromising mechanistic validity.

4.3.2 Mitigation Strategies

Na-HCO_3-type and Na-Cl waters present distinct mitigation challenges requiring targeted interventions. The dominant Na-HCO_3 water type (64.7% of samples) with mean fluoride of 0.83 ± 0.39 mg/L represents alkaline, low-calcium conditions promoting fluoride mobility through calcium sequestration and alkaline desorption mechanisms. Na-Cl waters, comprising 8.8% of samples with the highest mean fluoride of 3.75 ± 2.32 mg/L, reflect halite or evaporite dissolution and potential saline intrusion that reduces calcium availability. For Na-HCO_3-type waters, a multi-pronged approach is warranted: managed aquifer abstraction prioritizing Ca-Mg-HCO_3 boreholes where available; water blending combining high-fluoride Na-HCO_3 waters with fresher, higher-calcium sources to simultaneously reduce salinity and promote fluorite precipitation; and pH neutralization using calcium hydroxide to raise pH controllably while introducing Ca^{2+} ions that reduce fluoride solubility (Narsimha et al., 2018). Point-of-use defluoridation performance can be sensitive to co-occurring salinity and competing ions; therefore, technology selection should explicitly account for the observed ionic strength/TDS range in Karaga groundwaters. Where higher salinity is present, pilot testing is recommended to confirm media performance under local water chemistry, and composite or modified sorbents may offer improved robustness depending on competing-ion loads. Agricultural activities intensify fluoride mobilization through complex mechanisms. Nitrate shows moderate

positive correlation with fluoride (r = 0.34), suggesting deep flushing of fluoride-bearing minerals under intensive recharge from fertilizer use. The K^+-Mg^{2+} correlation (r = 0.55) indicates fertilizer-impacted areas exhibit concurrent geochemical changes. Community-level interventions should prioritize: promoting organic or slow-release fertilizers in recharge zones; reducing abstraction intensity where agricultural demand drives deep drawdown; and constructing wetlands to attenuate agricultural runoff before recharge.

4.3.3 Early Warning System Development

The Mobility Index provides a cost-effective, field-operationalized early warning framework for community-level fluoride surveillance. Input parameters require only basic field instrumentation: electrical conductivity measured with portable probes, total derived from EC, pH with inexpensive meters, and routine major ion analysis available in regional laboratories.

In this study, MI computation is intended to be feasible once a standard major-ion dataset is available (field + routine laboratory chemistry), with implementation supported by a fixed-weight calculator so that end-users do not need to reproduce the full modelling workflow (Fig 9).

These field-measurable inputs replace computationally intensive geochemical modelling (PHREEQC), which served a calibration role but need not accompany operational deployment. The framework requires only that entropy-weighted components computed during development remain locked during external application, preserving mechanistic interpretability while enabling practical implementation. Karaga District's tropical continental climate with rainy season from May to October and dry season November to April creates seasonal fluoride variability amenable to temporal surveillance. Quarterly MI calculations capture seasonal transitions: dry season concentration effects potentially elevate fluoride through reduced recharge, while wet season dilution effects lower concentrations if recharge brings fluoride-poor waters. Threshold-based

decision triggers categorize risk simply: wells with MI <0.33 indicate very low risk; 0.33–0.67 intermediate risk; >0.67 high risk (corresponding to >80% probability of exceedance based on logistic calibration AUROC = 0.976). Institutional implementation requires minimal training: District Health Directorate oversees monitoring and coordination, Regional Water Authority links results to infrastructure maintenance, and Community Water Committees conduct quarterly measurements and alert authorities. Color-coded community maps using MI categories enable spatially explicit identification of high-risk clusters prompting targeted intervention. Intercept adjustment ($\alpha^* = -1.601$) accommodates local prevalence shifts without recalibrating component weights, eliminating community-specific recalibration requirements. This framework complements rather than replaces direct fluoride analysis by serving as a cost-effective pre-screening tool, reducing analytical burden while prioritizing wells warranting immediate testing and treatment activation.

4.4 Limitations and Future Directions

4.4.1 Limitations

While this study demonstrates strong internal validation and external validation within a single district, several limitations constrain broader spatial transferability. First, the Karaga external validation dataset comprises only 34 samples collected in a single season (October 2022), which may not capture seasonal variation in groundwater chemistry and fluoride concentrations. Multi-seasonal sampling would strengthen confidence in year-round model applicability. Second, validation was performed solely within the Karaga District, a specific geological setting (Voltaian Supergroup, semi-arid climate). Extension to other regions, particularly those with different geological formations (e.g., granitic aquifers, coastal aquifers) or climatic regimes (e.g., tropical, temperate), requires additional multi-region validation studies. The model parameters and feature relationships identified here may not apply directly to

fundamentally different hydrogeochemical systems. Third, the archived regional dataset (n=152) used for model training, while valuable, represents a static snapshot and may not account for temporal trends in groundwater chemistry. Fourth, the Mobility Index has been validated specifically for predicting WHO guideline exceedance (1.5 mg/L threshold) and may require recalibration for other decision thresholds or health-based standards used in different jurisdictions. Despite these limitations, the methodological framework integrating CoDA, PHREEQC modeling, and interpretable ML is generalizable and can be adapted to other regions through similar multi-stage validation protocols.

4.4.2 Uncertainties

Uncertainty arises from (i) sampling limitations (single-season sampling and limited exceedance cases in the external set), (ii) measurement and modelling uncertainty in derived thermodynamic descriptors (e.g., saturation indices and ionic strength from PHREEQC inputs), (iii) structural uncertainty because localized controls (e.g., clay–water interactions and micro-scale mineral heterogeneity) may not be captured by major-ion chemistry alone, and (iv) decision uncertainty because screening sensitivity depends on the selected exceedance threshold and probability cut-off. These uncertainties primarily affect confidence in the upper tail (rare high-fluoride events) and motivate expanded multi-season sampling, targeted mineralogical/trace-element constraints, and validation in additional regions

4.4.3 Future Research Directions

Advancing fluoride prediction in the Karaga District requires a strategic program integrating expanded data collection, process-based investigations, and methodological refinements. Currently, the external test set contains only six exceedance cases, creating unstable estimates and limiting sensitivity to 50%. Deliberately targeting the 26 communities with hazard quotient values exceeding 1.0 to increase exceedance samples from six to 20–30 cases would reduce confidence interval width and better

capture underprediction bias at high concentrations. Concurrent systematic sampling across all identified water types and distinct Voltaian Formation members would ensure adequate representation of the full compositional space and clarify geological controls independent of regional patterns. Temporal monitoring through bi-monthly sampling at sentinel wells over 24 months would quantify seasonal fluoride variability and enable development of state-space predictive models incorporating temporal dynamics. Companion investigations should apply rigorous compositional data analysis to trace elements including iron, aluminium, and silicon, which control fluoride behaviour through adsorption onto goethite, gibbsite, and clay minerals processes not captured by major-ion analysis alone. Laboratory experiments examining fluorite dissolution kinetics and fluoride adsorption on local Voltaian sediments would provide rate expressions necessary for reactive transport modelling. Isotope hydrology using $\delta^{2}H$ and $\delta^{18}O$, combined with stable fluorine isotope analysis ($\delta^{19}F$), would distinguish geogenic fluoride sources and link mineral origins to observed concentration patterns. Machine learning enhancements incorporating quantile regression and heteroscedastic neural networks would provide uncertainty quantification critical for public health decision-making, while SHAP interaction analysis would reveal mechanistic feature combinations driving extreme outcomes. These interconnected approaches would establish a robust, transferable framework for fluoride prediction applicable across Ghana and similar geological settings globally.

5 Conclusion

Fluoride contamination in the Karaga District poses a critical public health threat, with 17.6% of groundwater samples exceeding the WHO drinking water standard of 1.5 mg/L and affecting thousands of residents, particularly children vulnerable to dental and skeletal fluorosis. Although national hazard assessments identified the Karaga area as a high-risk zone, the specific geochemical mechanisms driving fluoride mobilization

in the region's Voltaian Supergroup aquifers have remained poorly understood. This study addressed that gap through an integrated framework combining geochemical modelling, compositional data analysis, and machine learning to uncover both the drivers and predictability of fluoride contamination at the local scale. Our findings revealed three pivotal insights. First, geochemical analysis showed that while Na-HCO_3 waters dominated (64.7% of samples), the most severe contamination occurred in Na-Cl waters, which reached concentrations as high as 6.04 mg/L, indicating that evaporite dissolution and cation exchange fundamentally control enhanced fluoride mobilization. Second, machine learning analysis exposed nonlinear relationships governing fluoride behaviour, with total dissolved solids and pH emerging as primary predictors not individual ion concentrations. Third, we developed the Mobility Index, a mechanistic tool that successfully identified high-risk waters while remaining independent of measured fluoride, achieving exceptional discrimination (AUROC of 0.94) for WHO guideline exceedance. The strength of our approach lies in integrating mechanistic understanding with predictive power. Single-method frameworks cannot capture fluoride's nonlinear behaviour; geochemical models alone require extensive lab data, while machine learning alone lacks interpretability. Our hybrid strategy addressed both limitations, encoding genuine hydrogeochemical relationships into a model that learns from data patterns. Operationally, the Mobility Index requires only field-measurable parameters electrical conductivity, pH, and routine ion analysis making it immediately deployable by district health authorities and community water committees for cost-effective screening and targeted intervention. Beyond Karaga, this framework is transferable to similar sedimentary aquifer settings globally, opening pathways for expanded validation and adaptation in other fluoride-endemic regions worldwide. This study is novel in three specific ways. First, it couples PHREEQC-based thermodynamic modelling with rigorously transformed compositional features (isometric log-ratio/CoDA) to avoid the spurious correlations that arise when standard statistics are applied to constrained hydrochemical

data, an approach that remains largely absent from fluoride prediction literature. Second, it enforces a leakage-aware, fluoride-blind feature design throughout model development and validates the resulting models on a fully independent, externally collected dataset from the Karaga District, providing a more rigorous test of generalisability than internal cross-validation alone can offer. Third, it introduces the Mobility Index, a mechanistically interpretable screening tool derived entirely from fluoride-independent components, which translates complex geochemical controls into a practical, field-deployable risk signal without requiring measured fluoride as an input.

Acknowledgement

The authors gratefully acknowledge all those who contributed to the fieldwork and other aspects of this study, whose efforts significantly enhanced its quality. The first author thanks the University of Johannesburg, South Africa for the continuous support as a Senior Research Associate at the Department of Chemical Sciences.

Author Contribution Statement

E.D.S.: Conceptualization, Methodology, Investigation, Data curation, Supervision, Writing- Original draft preparation, Validation, Writing- Reviewing and Editing. **D.A.:** Methodology, Investigation, Software, Data curation, Writing- Original draft preparation, Visualization, Writing- Reviewing and Editing. **M.G.:** Supervision, Writing- Reviewing and Editing. **P.C.:** Writing- Reviewing and Editing. **A.A.A:** Writing- Reviewing and Editing.

Statements & Declarations

Ethical Approval: not applicable

Sampling Permissions: Groundwater samples used in this study were collected from active public boreholes drilled and maintained by World Vision International in the Karaga District, Northern Ghana. These boreholes are publicly accessible and do not fall within restricted or privately owned land. Therefore, no special permissions from landowners were required for sample collection.

Consent to Participate: not applicable

Consent to Publish: not applicable

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests: Emmanuel Daanoba Sunkari is an Editorial Board Member of Scientific Reports. The author declares that they were not involved in the editorial handling or decision-making process for this manuscript. Therefore, the authors have no relevant financial or non-financial interests to disclose.

Data Availability: The authors declare that the data supporting the findings of this study are available within the paper.

References

Abu, M., Sunkari, E. D., & Şener, M. (2019). Untapped Economic Resource Potential of the Neoproterozoic to Early Paleozoic Volta Basin, Ghana: A Review. *Natural Resources Research*, *28*(4), 1429–1445. https://doi.org/10.1007/s11053-019-09478-5

Achcampong, S. Y., & Hess, J. W. (1998). Hydrogeologic and hydrochemical framework of the shallow groundwater system in the southern Voltaian Sedimentary Basin, Ghana. *Hydrogeology Journal*, *6*(4), 527–537. https://doi.org/10.1007/s100400050173

Ahmad, M., Mustafa, G., Ali, N., & Laiq, M. (2023). Statistical Prediction of Fluoride Concentration in Groundwater of District Multan, Pakistan, Using Kriging Methods. *Fluoride*, *56*(2), 156–168.

Alam, N., Husain, M. A., Singh, R., Jain, P. K., Eiche, E., Neidhardt, H., Marks, M., Kumar, M., & Biswas, A. (2024). Geochemistry of fluoride mobilization in the hard-rock aquifers of central India: Implication for fluoride-safe drinking water supply. *Applied Geochemistry*, *171*, 106106. https://doi.org/10.1016/j.apgeochem.2024.106106

Ali, W., Aslam, M. W., Junaid, M., Ali, K., Guo, Y., Rasool, A., & Zhang, H. (2019). Elucidating various geochemical mechanisms drive fluoride contamination in unconfined aquifers along the major rivers in Sindh and Punjab, Pakistan. *Environmental Pollution*, *249*, 535–549. https://doi.org/10.1016/j.envpol.2019.03.043

Anani, C. (1999). Sandstone petrology and provenance of the Neoproterozoic Voltaian group in the southeastern Voltaian Basin, Ghana. *Sedimentary Geology*, *128*(1–2), 83–98. https://doi.org/10.1016/S0037-0738(99)00063-9

Apambire, W. B., Boyle, D. R., & Michel, F. A. (1997). Geochemistry, genesis, and health implications of fluoriferous groundwaters in the upper regions of Ghana. *Environmental Geology*, *33*(1), 13–24. https://doi.org/10.1007/s002540050221

Aravinthasamy, P., Karunanidhi, D., Subramani, T., Srinivasamoorthy, K., & Anand, B. (2020). Geochemical evaluation of fluoride contamination in groundwater from Shanmuganadhi River basin, South India: implication on human health. *Environmental Geochemistry and Health*, *42*(7), 1937–1963. https://doi.org/10.1007/s10653-019-00452-x

Araya, D., Podgorski, J., Kumi, M., Mainoo, P. A., & Berg, M. (2022). Fluoride contamination of groundwater resources in Ghana: Country-wide hazard modeling and estimated population at risk. *Water Research*, *212*(September 2021), 118083. https://doi.org/10.1016/j.watres.2022.118083

Ayub, M., Javed, H., Rashid, A., Khan, W. H., Javed, A., Sardar, T., Shah, G. M., Ahmad, A., Rinklebe, J., & Ahmad, P. (2024). Hydrogeochemical properties, source provenance, distribution, and health risk of high fluoride groundwater: Geochemical control, and source apportionment. *Environmental Pollution*, *362*. https://doi.org/10.1016/J.ENVPOL.2024.125000

Bera, B., Bhattacharjee, S., Chamling, M., Ghosh, A., Sengupta, N., & Ghosh, S. (2021). Fluoride dynamics in precambrian hard rock terrain of North Singhbhum Craton and effect of fluorosis on human health and society. *Groundwater and Society: Applications of Geospatial Technology*, 319–348.

Blarasin, M., Matteoda, E., Cabrera, A., Lutri, V., Felizzia, J., & Author, C. (2018). Arsenic and Fluoride in Groundwater of the Sedimentary Aquifer in. *IOSR Journal of Environmental Science*, *12*(4), 71–77. https://doi.org/10.9790/2402-1204017177

Chen, K., Liu, Q., Yang, T., Ju, Q., & Yu, H. (2023). Geochemical characteristics, influencing factors and health risk assessment of groundwater fluoride in a drinking water source area in North Anhui Plain, Eastern China. *Stochastic Environmental Research and Risk Assessment*, *37*(10), 3879–3891. https://doi.org/10.1007/s00477-023-02485-2

Dapaah-Siakwan, S., & Gyau-Boakye, P. (2000). Hydrogeologic framework and borehole yields in Ghana. *Hydrogeology Journal*, *8*(4), 405–416. https://doi.org/10.1007/PL00010976

Dar, F. A., & Kurella, S. (2024). Utilization of organic waste from Chinar leaves as sustainable and eco-friendly adsorbent for fluoride removal.

Environmental Science and Pollution Research, 1–24. https://doi.org/10.1007/s11356-024-35147-z

De, A., Das, A., Joardar, M., Mridha, D., Majumdar, A., Das, J., & Roychowdhury, T. (2023). Investigating spatial distribution of fluoride in groundwater with respect to hydro-geochemical characteristics and associated probabilistic health risk in Baruipur block of West Bengal, India. *The Science of the Total Environment*, *886*. https://doi.org/10.1016/J.SCITOTENV.2023.163877

Demir Yetiş, A., İlhan, N., & Kara, H. (2024). Integrating deep learning and regression models for accurate prediction of groundwater fluoride contamination in old city in Bitlis province, Eastern Anatolia Region, Türkiye. *Environmental Science and Pollution Research*, *31*(34), 47201–47219. https://doi.org/10.1007/s11356-024-34194-w

Ghana Geological Survey (GGS). (2009). *Geological Map of Ghana – Scale 1:1 000 000.* Geological Survey Department (GSD).

Ghana Statistical Service (GSS). (2021). *Ghana 2021 Population and Housing Census General Report*. Ghana Statistical Service Accra, Ghana.

Gibbs, R. J. (1970). Mechanisms controlling world water chemistry. *Science*, *170*(3962), 1088–1090. https://doi.org/10.1126/science.170.3962.1088

Hem, J. D. (1985). Study and interpretation of the chemical characteristics of natural water. In *US Geological Survey Water-Supply Paper* (Vol. 2254). https://doi.org/10.3133/wsp2254

Hossain, S., Hosono, T., Yang, H., & Shimada, J. (2016). Geochemical Processes Controlling Fluoride Enrichment in Groundwater at the Western Part of Kumamoto Area, Japan. *Water, Air, and Soil Pollution*, *227*(10), 385. https://doi.org/10.1007/s11270-016-3089-3

Kerketta, A., Kapoor, H. S., & Sahoo, P. K. (2024). Groundwater fluoride prediction modeling using physicochemical parameters in Punjab,

India: a machine-learning approach. *Frontiers in Soil Science*, *4*. https://doi.org/10.3389/fsoil.2024.1407502

Kontogeorgis, G. M., Maribo-Mogensen, B., & Thomsen, K. (2018). The Debye-Hückel theory and its importance in modeling electrolyte solutions. *Fluid Phase Equilibria*, *462*, 130–152. https://doi.org/10.1016/j.fluid.2018.01.004

Kumar, A., & Singh, A. (2024a). Geospatial mapping and entropy-based analysis for groundwater evaluation with estimation of potential health risks due to nitrate and fluoride exposure. *Environmental Science and Pollution Research*, *31*(59), 66953–66976. https://doi.org/10.1007/s11356-024-35691-8

Kumar, A., & Singh, A. (2024b). Pollution source characterization and evaluation of groundwater quality utilizing an integrated approach of Water Quality Index, GIS and multivariate statistical analysis. *Water Supply*, *24*(10), 3517–3539. https://doi.org/10.2166/ws.2024.213

Kumar, A., & Singh, A. (2025). Entropy-based groundwater quality evaluation with multivariate analysis and Sobol sensitivity for non-carcinogenic health risks in mid-Gangetic plains, India. *Environmental Geochemistry and Health*, *47*(6), 186. https://doi.org/10.1007/s10653-025-02495-9

Ling, Y., Podgorski, J., Sadiq, M., Rasheed, H., Eqani, S. A. M. A. S., & Berg, M. (2022). Monitoring and prediction of high fluoride concentrations in groundwater in Pakistan. *Science of the Total Environment*, *839*, 156058. https://doi.org/10.1016/j.scitotenv.2022.156058

Liu, J., Peng, Y., Li, C., Gao, Z., & Chen, S. (2021). A characterization of groundwater fluoride, influencing factors and risk to human health in the southwest plain of Shandong Province, North China. *Ecotoxicology and Environmental Safety*, *207*, 111512. https://doi.org/10.1016/j.ecoenv.2020.111512

Liu, X., & Chen, K. (2024). Characterization, formation mechanism, and human health risk assessment of fluoride in shallow groundwater of Suzhou city, East China. *Water Supply*, *24*(9), 3196–3207. https://doi.org/10.2166/ws.2024.202

Liu, Y., Zhou, K., & Carranza, E. J. M. (2018). Compositional balance analysis for geochemical pattern recognition and anomaly mapping in the western Junggar region, China. *Geochemistry: Exploration, Environment, Analysis*, *18*(3), 263–276. https://doi.org/10.1144/geochem2017-050

Lone, S. A., Jeelani, G., & Mukherjee, A. (2024). Hydrogeochemical controls on contrasting co-occurrence of geogenic Arsenic (As) and Fluoride (F-) in complex aquifer system of Upper Indus Basin, (UIB) western Himalaya. *Environmental Research*, *260*. https://doi.org/10.1016/J.ENVRES.2024.119675

Luo, W., Gao, X., & Zhang, X. (2018). Geochemical processes controlling the groundwater chemistry and fluoride contamination in the yuncheng basin, China—an area with complex hydrogeochemical conditions. *PLoS ONE*, *13*(7), e0199082. https://doi.org/10.1371/journal.pone.0199082

Menyeh, A., & Sarpong Asare, V.-D. (2013). Geo-Electrical Investigation Of Groundwater Resources And Aquifer Characteristics In Some Small Communities In The Gushiegu And Karaga Districts Of Northern Ghana. *International Journal of Scientific & Technology Research*, *2*, 25–35. www.ijstr.org

Mosaad, S., Eissa, M., & Alezabawy, A. K. (2022). Geochemical modeling and geostatistical categorization of groundwater in Nubian Sandstone Aquifer, El Bahariya Oasis, Egypt. *Environmental Earth Sciences*, *81*(17), 421. https://doi.org/10.1007/s12665-022-10524-4

Müller, M., Parkhurst, D., & Charlton, S. (2011). Programming PHREEQC Calculations with C++ and Python A Comparative Study. *MODFLOW and More 2011: Integrated Hydrological Modeling*, *January*, 632–636.

http://docplayer.net/7809789-Programming-phreeqc-calculations-with-c-and-python-a-comparative-study.html

Nafouanti, M. B., Li, J., Mustapha, N. A., Uwamungu, P., & AL-Alimi, D. (2021). Prediction on the fluoride contamination in groundwater at the Datong Basin, Northern China: Comparison of random forest, logistic regression and artificial neural network. *Applied Geochemistry*, *132*, 105054. https://doi.org/10.1016/j.apgeochem.2021.105054

Narsimha, A., Venkatayogi, S., & Geeta, S. (2018). Hydrogeochemical data on groundwater quality with special emphasis on fluoride enrichment in Munneru river basin (MRB), Telangana State, South India. *Data in Brief*, *17*, 339–346. https://doi.org/10.1016/J.DIB.2018.01.059

Narsimha Adimalla. (2020). Assessment and Mechanism of Fluoride Enrichment in Groundwater from the Hard Rock Terrain: A Multivariate Statistical Approach. *Geochemistry International*, *58*(4), 456–471. https://doi.org/10.1134/S0016702920040060

Oh, J., Kim, K. H., Kim, H. R., Park, S., & Yun, S. T. (2024). Using isometric log-ratio in compositional data analysis for developing a groundwater pollution index. *Scientific Reports*, *14*(1), 12196. https://doi.org/10.1038/s41598-024-63178-6

Padilla-Reyes, D. A., Dueñas-Moreno, J., Mahlknecht, J., Mora, A., Kumar, M., Ornelas-Soto, N., Mejía-Avendaño, S., Navarro-Gómez, C. J., & Bhattacharya, P. (2024). Arsenic and fluoride in groundwater triggering a high risk: Probabilistic results using Monte Carlo simulation and species sensitivity distribution. *Chemosphere*, *359*, 142305. https://doi.org/10.1016/j.chemosphere.2024.142305

Paikaray, S., & Mahajan, T. (2023). Hydrogeochemical processes, mobilization controls, soil-water-plant-rock fractionation and origin of fluoride around a hot spring affected tropical monsoonal belt of eastern Odisha, India. *Applied Geochemistry*, *148*. https://doi.org/10.1016/J.APGEOCHEM.2022.105521

Parkhurst, D.L., and Appelo, C. A. J. (2021). *PHREEQC Version 3* (3). https://doi.org/https://doi.org/10.3133/tm6A43

Parkhurst, D. L., & Appelo, C. A. J. (1999). User's Guide to PHREEQC (Version 2): A Computer Program for Speciation, Batch-Reaction, One-Dimensional Transport, and Inverse Geochemical Calculations. In *Water-Resources Investigations Report 99-4259* (Issue Water-Resources Investig. Rep. 99-4259). https://doi.org/10.3133/wri994259

Rango, T., Kravchenko, J., Atlaw, B., McCornick, P. G., Jeuland, M., Merola, B., & Vengosh, A. (2012). Groundwater quality and its health impact: An assessment of dental fluorosis in rural inhabitants of the Main Ethiopian Rift. *Environment International*, *43*(1), 37–47. https://doi.org/10.1016/j.envint.2012.03.002

Rashid, A., Farooqi, A., Gao, X., Zahir, S., Noor, S., & Khattak, J. A. (2020). Geochemical modeling, source apportionment, health risk exposure and control of higher fluoride in groundwater of sub-district Dargai, Pakistan. *Chemosphere*, *243*, 125409. https://doi.org/10.1016/j.chemosphere.2019.125409

Rena, V., Vishwakarma, C. A., Singh, P., Roy, N., Asthana, H., Kamal, V., Kumar, P., & Mukherjee, S. (2022). Hydrogeological investigation of fluoride ion in groundwater of Ruparail and Banganga basins, Bharatpur district, Rajasthan, India. *Environmental Earth Sciences*, *81*(17), 430. https://doi.org/10.1007/s12665-022-10520-8

Rojanaworarit, C., Claudio, L., Howteerakul, N., Siramahamongkol, A., Ngernthong, P., Kongtip, P., & Woskie, S. (2021). Hydrogeogenic fluoride in groundwater and dental fluorosis in Thai agrarian communities: a prevalence survey and case–control study. *BMC Oral Health*, *21*(1), 1–16. https://doi.org/10.1186/s12903-021-01902-8

Saini, A., Kanwar, P., Kumar, S., Tembhurne, S., & Roy, I. (2023). A study on the hydrogeochemical mechanisms controlling groundwater fluoride enrichment in Jaipur: a semi-arid terrain in India. *International Journal of Environmental Analytical Chemistry*, *103*(20),

8825–8845. https://doi.org/10.1080/03067319.2021.1998473

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Advanced Intelligent Systems*, *7*(1). https://doi.org/10.1002/aisy.202400304

Sauro Graziano, R., Gozzi, C., & Buccianti, A. (2020). Is Compositional Data Analysis (CoDA) a theory able to discover complex dynamics in aqueous geochemical systems? *Journal of Geochemical Exploration*, *211*, 106465. https://doi.org/10.1016/j.gexplo.2020.106465

Scealy, J. L., de Caritat, P., Grunsky, E. C., Tsagris, M. T., & Welsh, A. H. (2015). Robust principal component analysis for power transformed compositional data. *Journal of the American Statistical Association*, *110*(509), 136–148. https://doi.org/10.1080/01621459.2014.990563

Shaji, E., Sarath, K. V., Santosh, M., Krishnaprasad, P. K., Arya, B. K., & Babu, M. S. (2024). Fluoride contamination in groundwater: A global review of the status, processes, challenges, and remedial measures. *Geoscience Frontiers*, *15*(2), 101734. https://doi.org/10.1016/j.gsf.2023.101734

Shelton, J. L., Engle, M. A., Buccianti, A., & Blondes, M. S. (2018). The isometric log-ratio (ilr)-ion plot: A proposed alternative to the Piper diagram. *Journal of Geochemical Exploration*, *190*(February), 130–141. https://doi.org/10.1016/j.gexplo.2018.03.003

Sunkari, E. D., & Abu, M. (2019). Hydrochemistry with special reference to fluoride contamination in groundwater of the Bongo district, Upper East Region, Ghana. *Sustainable Water Resources Management*, *5*(4), 1803–1814. https://doi.org/10.1007/s40899-019-00335-0

Sunkari, E. D., Adams, S. J., Okyere, M. B., & Bhattacharya, P. (2022). Groundwater fluoride contamination in Ghana and the associated human health risks: Any sustainable mitigation measures to curtail the long term hazards? *Groundwater for Sustainable Development*, *16*,

100715. https://doi.org/10.1016/j.gsd.2021.100715

Sunkari, E. D., Hudu, A., Fosu, S., Gyimah, E., & Oppong, O. (2024). Hydrogeochemistry, sources, enrichment mechanism and human health risk assessment of groundwater fluoride in Saboba District in the Oti sub-basin of the Volta River Basin, northern Ghana. *Groundwater for Sustainable Development*, *25*(February), 101132. https://doi.org/10.1016/j.gsd.2024.101132

Sunkari, E. D., Zango, M. S., & Korboe, H. M. (2018). Comparative Analysis of Fluoride Concentrations in Groundwaters in Northern and Southern Ghana: Implications for the Contaminant Sources. *Earth Systems and Environment*, *2*(1), 103–117. https://doi.org/10.1007/s41748-018-0044-z

Tian, J., Wang, Z., Daskalopoulou, K., Zhang, M., Huo, Y., Cao, Y., Yang, J., Liu, W., Liu, J., & Xu, S. (2025). Hydrochemical characteristics, driving factors and health risk of fluoride in groundwater from the northwestern Ordos Basin, China. *Geoscience Frontiers*, *16*(5). https://doi.org/10.1016/J.GSF.2025.102123

Tyagi, S., & Sarma, K. (2021). Expounding major ions chemistry of groundwater with significant controlling factors in a suburban district of Uttar Pradesh, India. *Journal of Earth System Science*, *130*(3), 169. https://doi.org/10.1007/s12040-021-01629-8

Varol, E., & Varol, S. (2012). Does fluoride toxicity cause hypertension in patients with endemic fluorosis? *Biological Trace Element Research*, *150*(1–3), 1–2. https://doi.org/10.1007/s12011-012-9499-1

Yadav, A., Kumari, N., Kumar, R., Kumar, M., & Yadav, S. (2023). Fluoride distribution, contamination, toxicological effects and remedial measures: a review. *Sustainable Water Resources Management*, *9*(5), 150. https://doi.org/10.1007/s40899-023-00926-y

Zhou, X., Ma, Y., & Wu, W. (2023). Statistical depth for point process via the isometric log-ratio transformation. *Computational Statistics and*

Data Analysis, 187, 107813. https://doi.org/10.1016/j.csda.2023.107813

Tables and Figures

Tables

Table 1. Optimal SBP Structure and Hierarchical Clustering Patterns of Major Ions

Balance	Positive group (+)	Negative group (−)	Geochemical interpretation
B1	Na^+, K^+, Ca^{2+}, Mg^{2+}	Cl^-, SO_4^{2-}, HCO_3^-, NO_3^-, F^-	Overall electroneutrality: total cations vs. total anions
B2	K^+	NO_3^-	Fertiliser-derived K vs. nitrate contamination (agriculture)
B3	Mg^{2+}	K^+, NO_3^-	Dolomite / Mg release (rock weathering) opposed to agro-inputs
B4	Na^+	SO_4^{2-}	Silicate weathering / ion-exchange Na vs. gypsum / evaporite SO_4
B5	Ca^{2+}	Cl^-	Calcite dissolution vs. halite or saline intrusion
B6	HCO_3^-	Ca^{2+}, Cl^-	Carbonate buffering (alkalinity) vs. hardness & salinity load
B7	Na^+, SO_4^{2-}	Ca^{2+}, Cl^-, HCO_3^-	Marine/evaporite Na-SO_4 sources contrasted with freshwater Ca-HCO_3
B8	F^-	Na^+, Ca^{2+}, Cl^-, SO_4^{2-}, HCO_3^-	Fluoride mobilisation (geogenic) against the general salinity matrix
B9	K^+, Mg^{2+}, NO_3^-	Na^+, Ca^{2+}, Cl^-, SO_4^{2-}, HCO_3^-, F^-	Secondary silicate weathering + agricultural inputs vs. background salinity

Table 2. Summary statistics of physicochemical parameters in Karaga District groundwater

Parameters	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**	**WHO Standard**	**% Above / Outside**
pH	7.57	0.55	6.00	7.32	7.73	7.94	8.33	6.5 - 8.5	5.9
EC (*µS/cm*)	1062	1632	66.60	678.0	744.5	827.0	9890	-	-
TDS (*mg/L*)	618.1	829.1	46.62	369.5	427.0	552.3	4950	1000	8.8
Na^+ (mg/L)	183.6	368.2	1.190	50.08	111.6	164.0	2144	200	17.6
K^+ (mg/L)	25.27	57.45	0.600	1.400	2.000	3.975	195.0	-	-
Mg^{2+} (mg/L)	4.192	8.145	0.001	1.100	1.520	5.125	46.80	-	-
Ca^{2+} (mg/L)	7.983	22.52	0.001	0.600	1.450	2.850	125.6	-	-

Cl^- (mg/L)	168.1	619.5	0.260	5.340	16.34	36.31	3520	250	8.8
SO_4^{2-} (mg/L)	4.689	5.401	0.001	0.425	4.050	7.575	26.70	500	0.0
HCO_3^- (mg/L)	310.4	143.5	19.70	229.0	316.6	426.7	520.0	-	-
NO_3^- (mg/L)	7.840	12.27	0.001	0.908	2.960	10.40	55.00	50	2.9
F^- (mg/L)	1.343	1.306	0.070	0.805	1.025	1.330	6.040	1.5	17.6

Table 3. PHREEQC geochemical modelling results and fluoride speciation

Category	Parameter	Unit	min	max	mean	median	std
Mineral Saturation Indices							
	$SI_{fluorite}$	-	-8.48	-1.09	-3.37	-3.17	1.51
	$SI_{calcite}$	-	-6.98	0.21	-1.4	-1.05	1.5
	$SI_{dolomite}$	-	-12.6	2.01	-1.29	-0.74	2.74
	SI_{gypsum}	-	-8.42	-3.12	-5.17	-5	1.3
	$SI_{sylvite}$	-	-9.92	-5.91	-8.31	-8.42	0.72
	SI_{halite}	-	-11.1	-3.83	-7.65	-7.26	1.67
	$SI_{magnesite}$	-	-7.21	0.17	-1.52	-1.12	1.33
	$SI_{anhydrite}$	-	-8.6	-3.3	-5.35	-5.18	1.3
	$SI_{mirabilite}$	-	-15.7	-6.38	-9.14	-8.67	2.04
	$SI_{thenardite}$	-	-16.5	-7.16	-9.93	-9.47	2.04
Fluoride Speciation							
	F^- Free	%	99.4	100	99.9	99.9	0.13
	CaF^+	%	0	0.623	0.05	0.01	0.13
Activities							
	A_F	mol/L	0.07	5.28	1.21	0.94	1.16
	A_{Ca}	mol/L	0.001	47.3	4.35	0.98	9.74
	Ionic Strength	mol/L	0.001	0.10	0.01	0.01	0.02

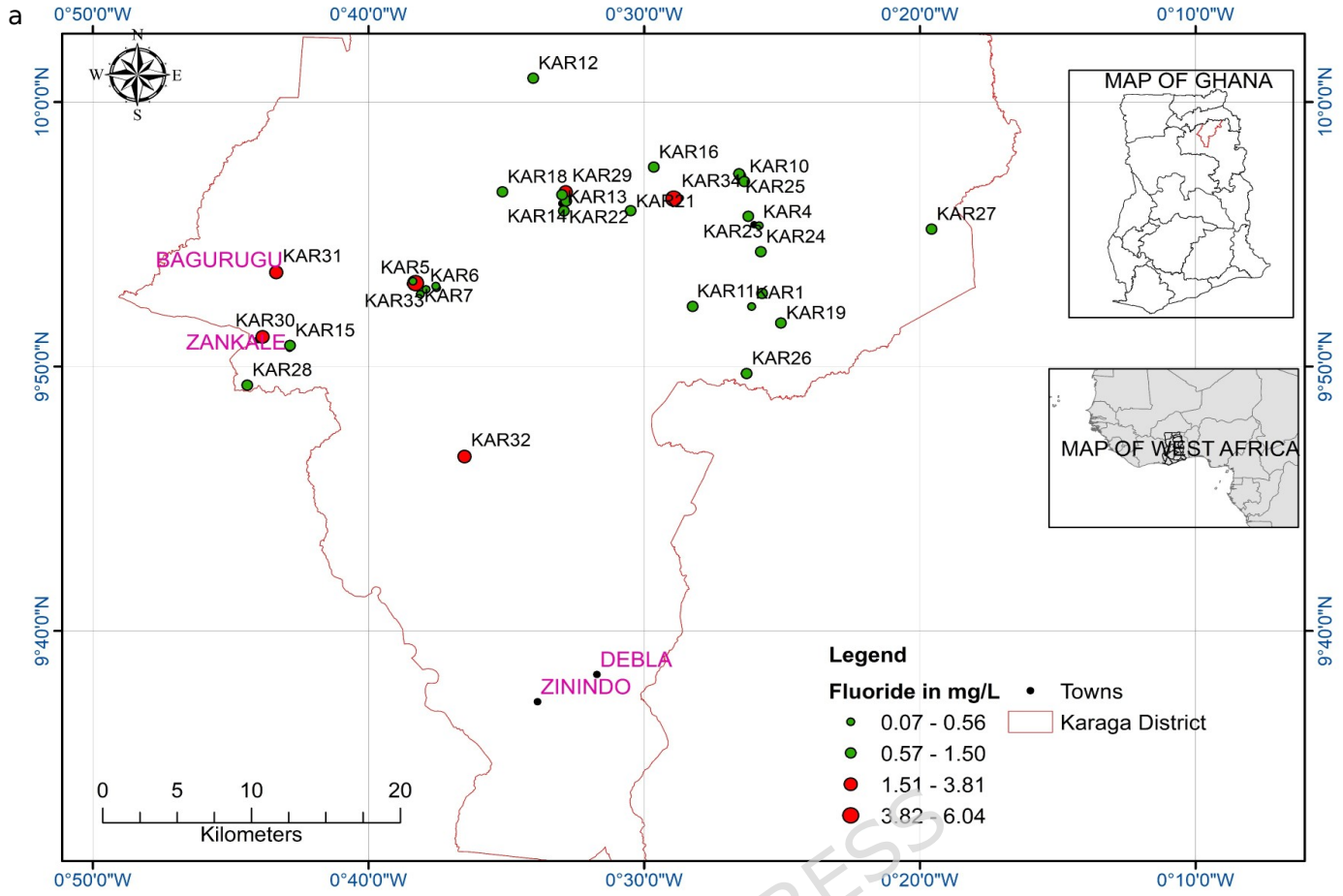
Table 4. Machine learning model performance comparison using nested cross-validation

Model	R^2	MAE	RMSE	Best Params
HGB	0.648±0.208	0.747±0.396	1.436±0.968	k: 11, learning_rate: 0.083, max_depth': 10, max_iter: 344, min_samples_leaf: 5, l2_regularization: 1.895×10^{-5}, loss': 'quantile', quantile: 0.572
RandomForest	0.586±0.156	0.762±0.324	1.534±0.873	k: 5, n_estimators: 460, max_depth: 25, min_samples_split: 3, min_samples_leaf: 1, max_features: sqrt
Ridge	-6.064±12.422	1.421±0.509	3.608±2.758	k: 6, alpha': 0.983

Lasso	-0.035±0.034	1.455±0.358	2.323±0.924	k: 12, alpha: 0.445
XGBoost	0.350±0.090	0.894±0.308	1.876±0.862	k: 25, n_estimators: 403, learning_rate: 0.013, max_depth: 8, min_child_weight: 1.132, subsample: 0.553, colsample_bytree: 0.891, gamma: 0.003, reg_alpha: 5.802×10^{-6}, reg_lambda: 0.107
MLP	0.668±0.189	0.654±0.141	1.196±0.357	k: 15, hidden_layer_code: 1, alpha': 0.064, learning_rate_init': 0.022

Table 5. Feature Importance

Rank	Feature	Importance
1	TDS (mg/L)	0.328
2	Ionic Strength (mol/L)	0.143
3	pH	0.14
4	$SI_{calcite}$	0.139
5	$SI_{magnesite}$	0.135
6	SI_{halite}	0.113
7	EC(µS/cm)	0.094
8	$SI_{sylvite}$	0.076
9	Cl^-(mg/L)	0.071
10	HCO_3^-(mg/L)	0.035



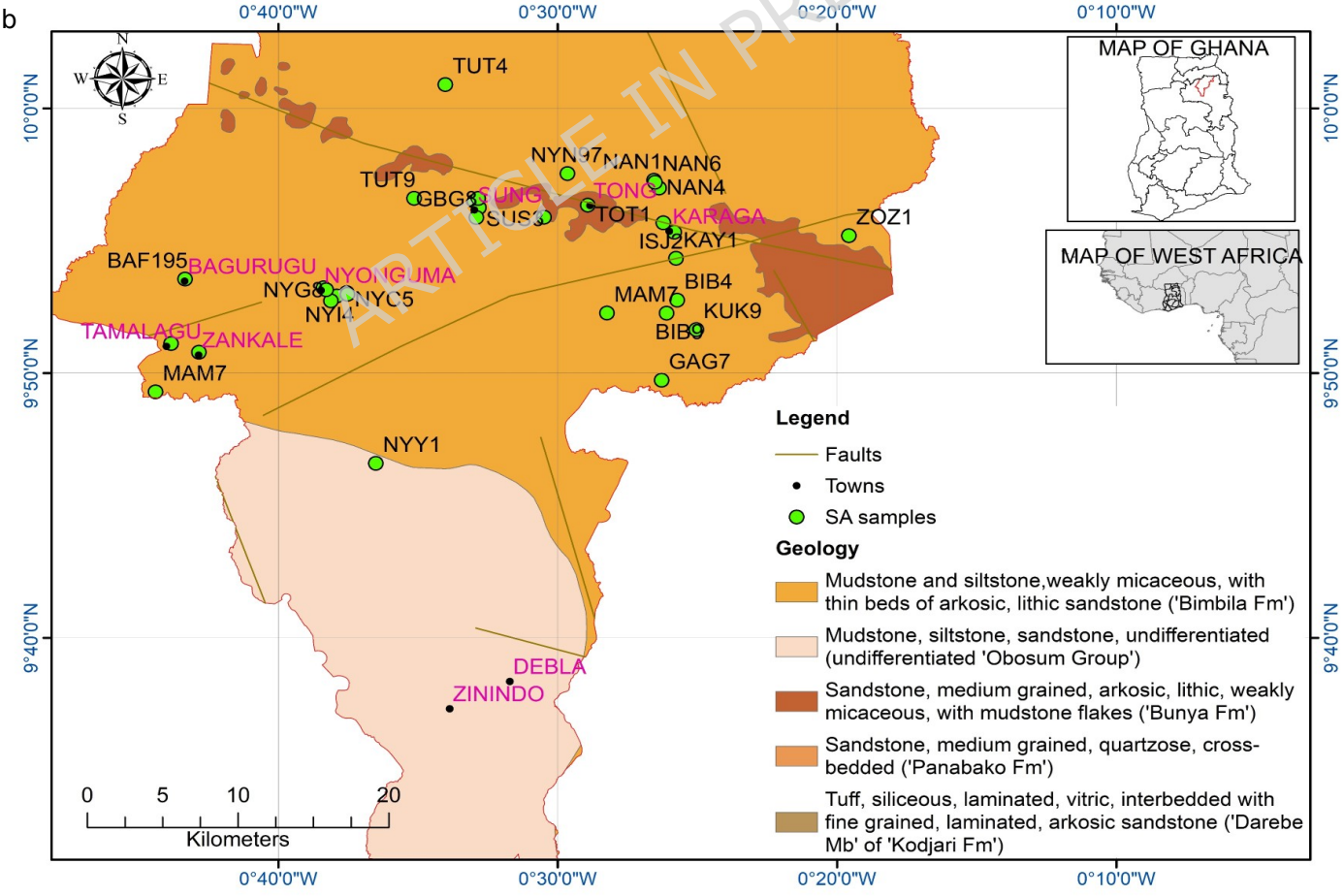


Fig. 1. (a) Study area map with sampling points and WHO-exceedance symbols, (b) geology

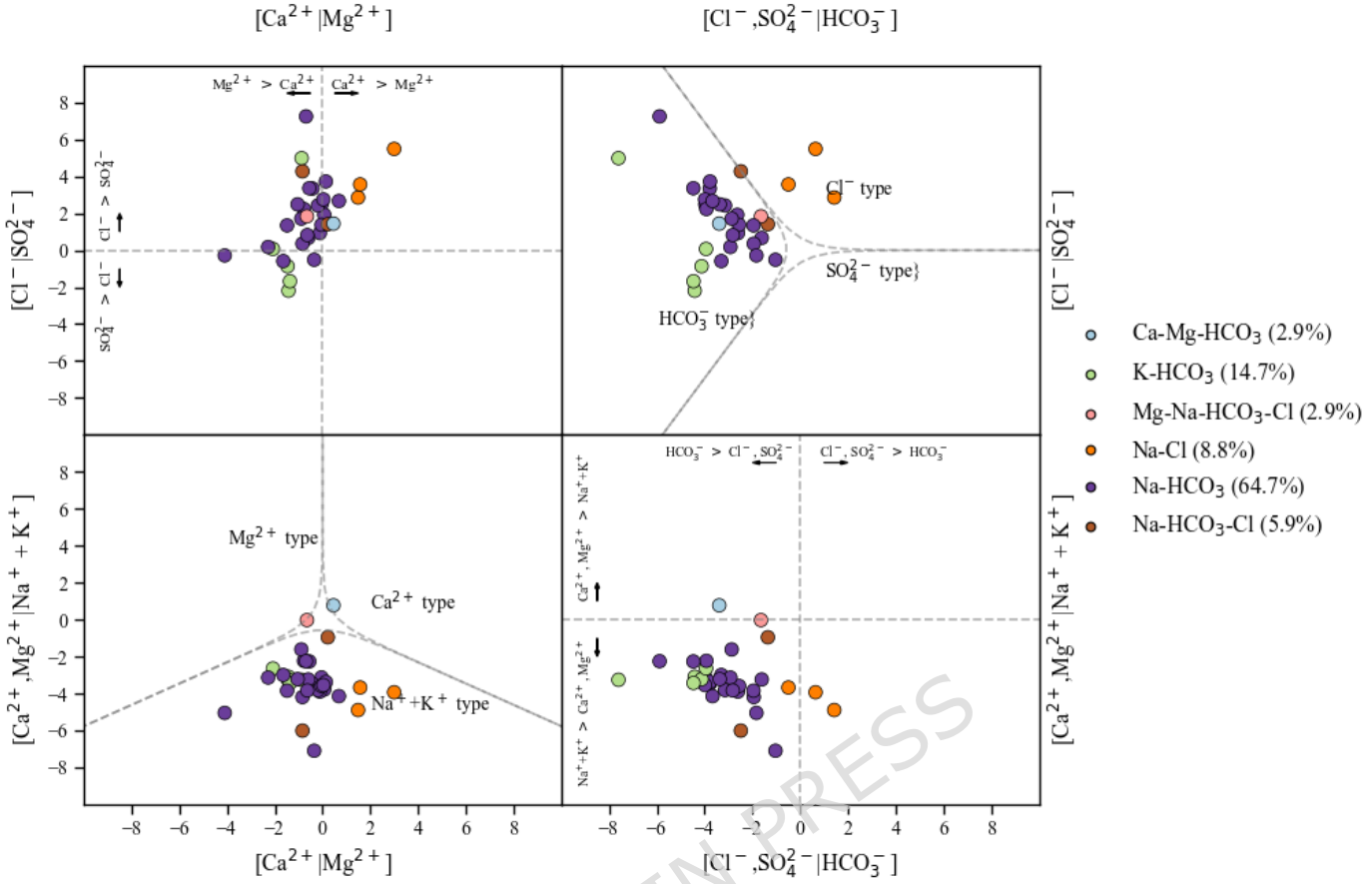


Fig. 2a. Isomeric water-type discrimination plot

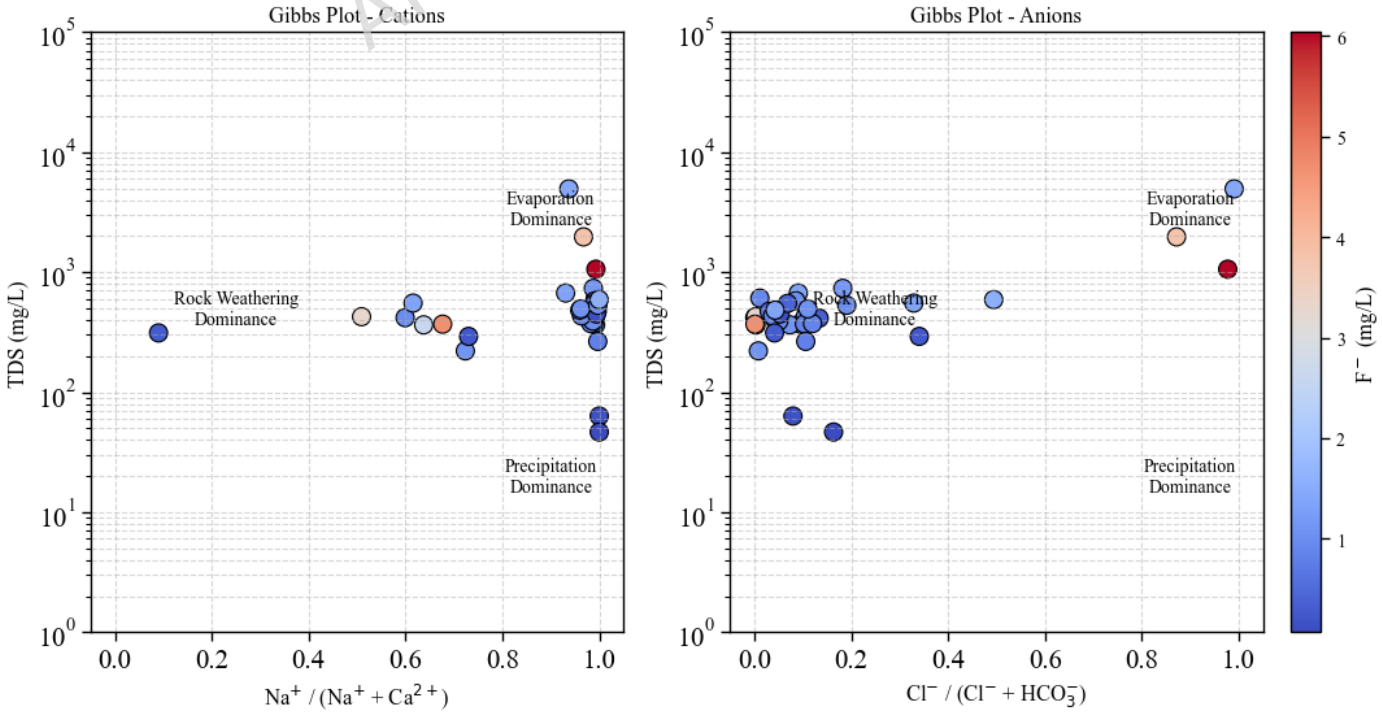


Fig. 2b. Dual Gibbs diagrams (cation & anion fields)

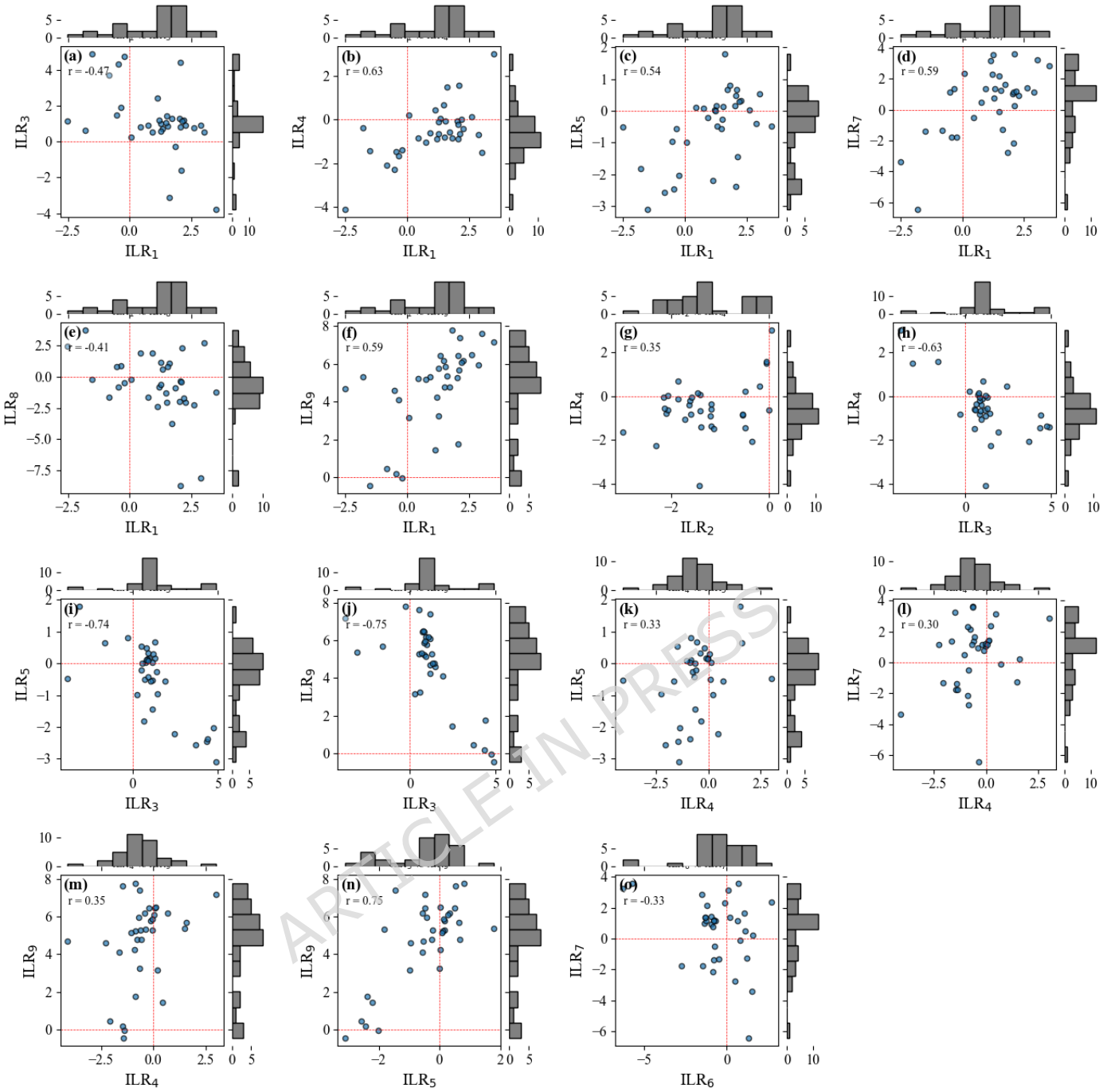


Fig 3. Pairwise relationships among high correlated isometric log-ratio (ILR) balances of groundwater major-ion compositions.

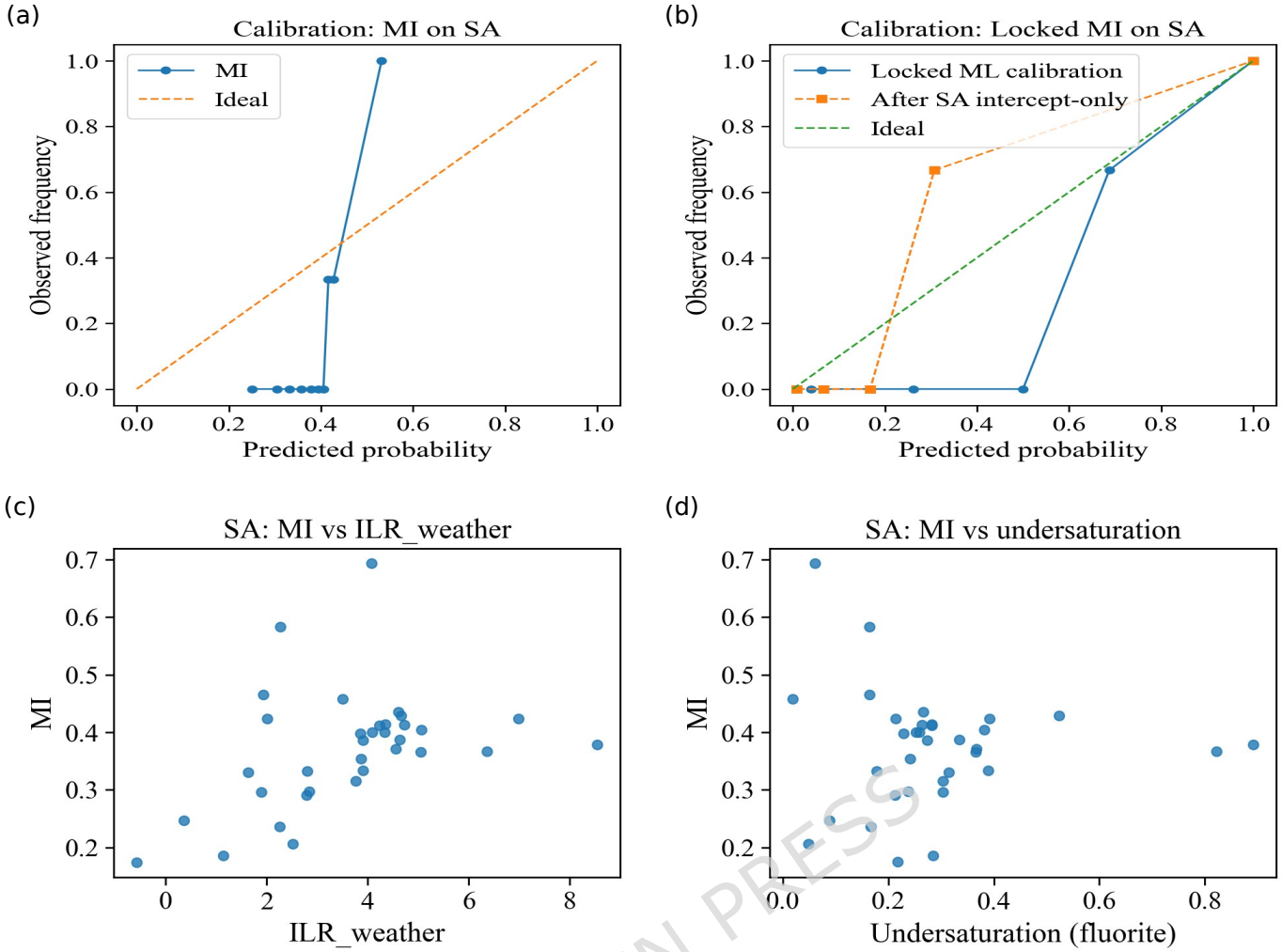


Fig. 4. Calibration and construct validity of the fluoride-independent Mobility Index (MI)

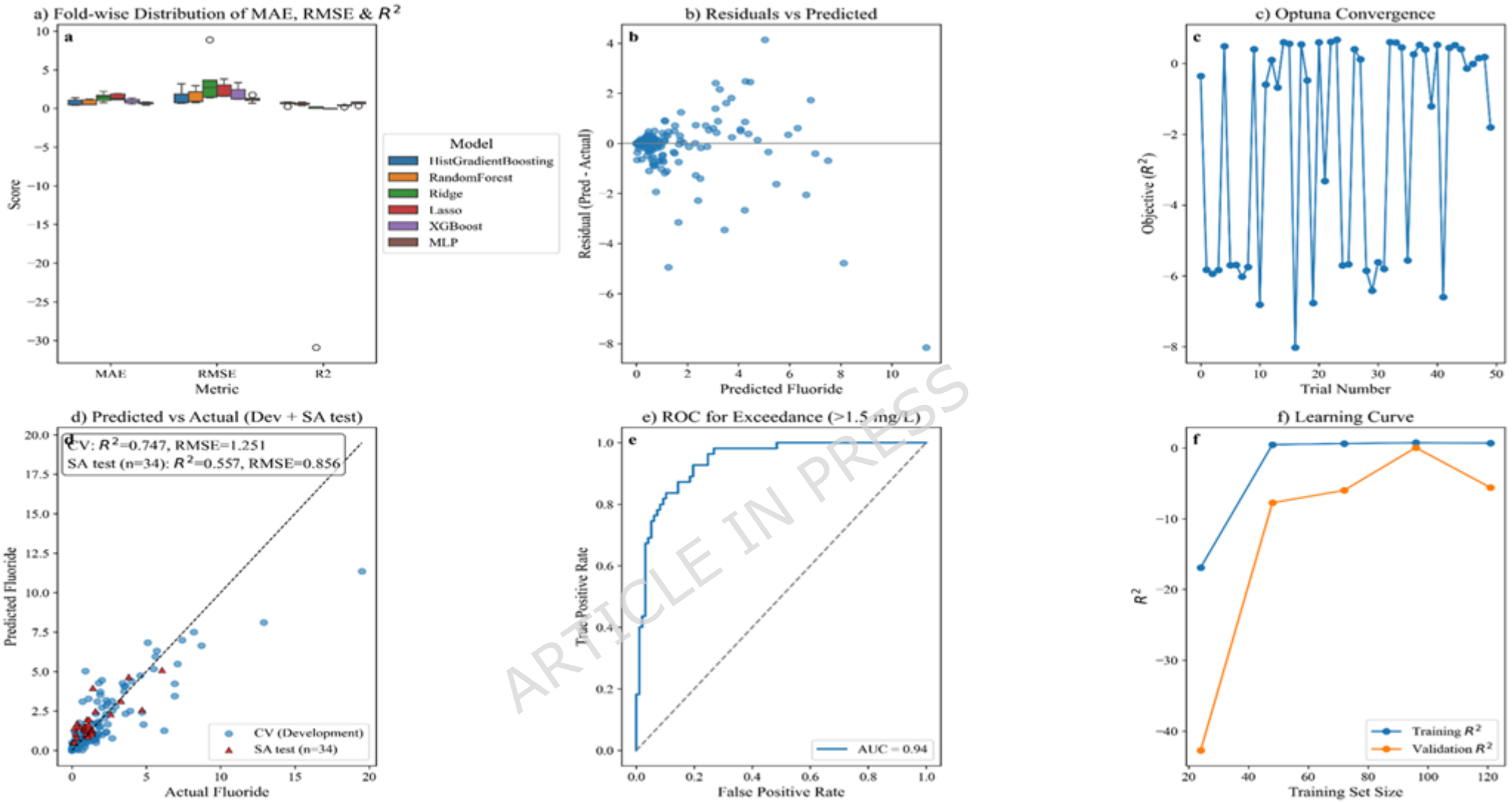


Fig. 5. Comprehensive performance evaluation of the selected fluoride-prediction model. (a) Fold-wise distributions of MAE, RMSE, and R^2 across six algorithms; (b) Residuals versus predicted values for hold-out samples; (c) Optuna hyperparameter-search convergence (objective = R^2);
(d) Predicted versus actual fluoride concentrations with overall R^2 and RMSE; (e) ROC curve and AUC for binary

classification of fluoride exceedance above 1.5 mg/L; (f) Learning curves showing training and validation R^2 as a function of training set size.

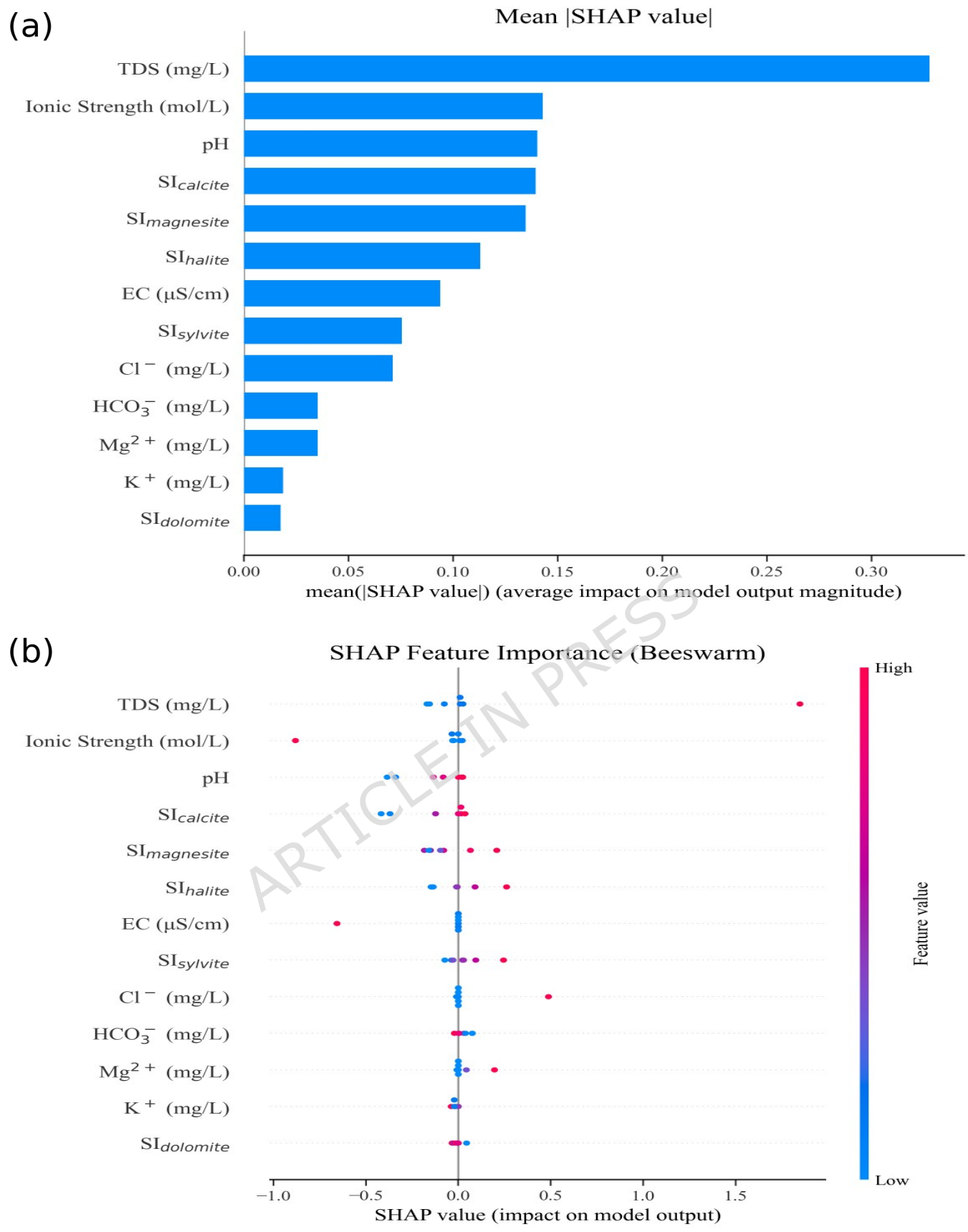


Fig. 6. (a) Mean absolute SHAP values for all features, showing each variable's average impact on the model's predictions of groundwater fluoride behaviour. (b) SHAP beeswarm plot showing feature importance for the groundwater-fluoride model.

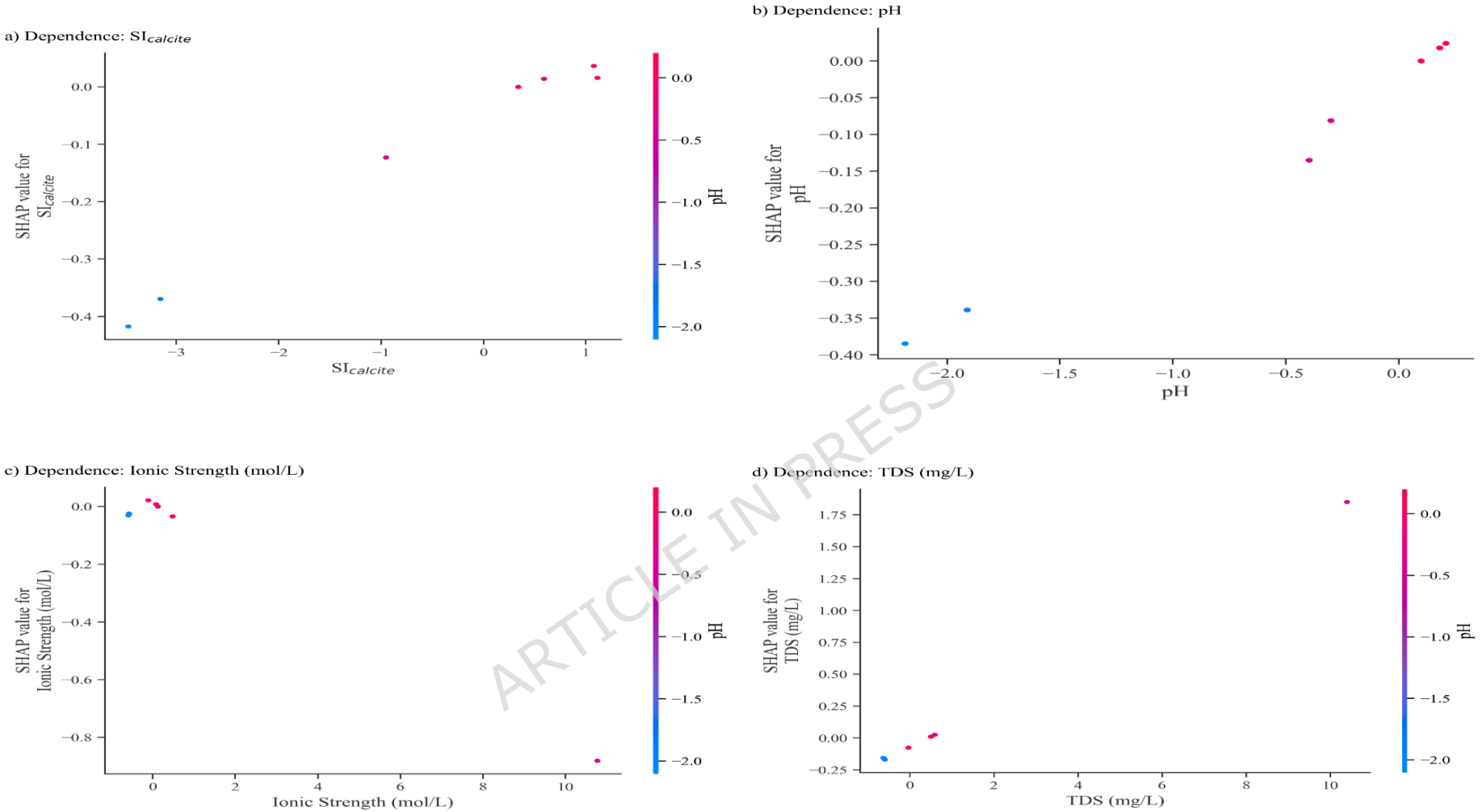


Fig. 7. SHAP dependence plots for four key hydrogeochemical predictors: (a) $\mathbf{SI_{calcite}}$, (b) pH, (c) Ionic Strength, and (d) Total Dissolved Solid. Each point represents one groundwater sample; the x-axis shows the (standardized) feature value, the y-axis its SHAP contribution to the model prediction, and the colour scale encodes pH.

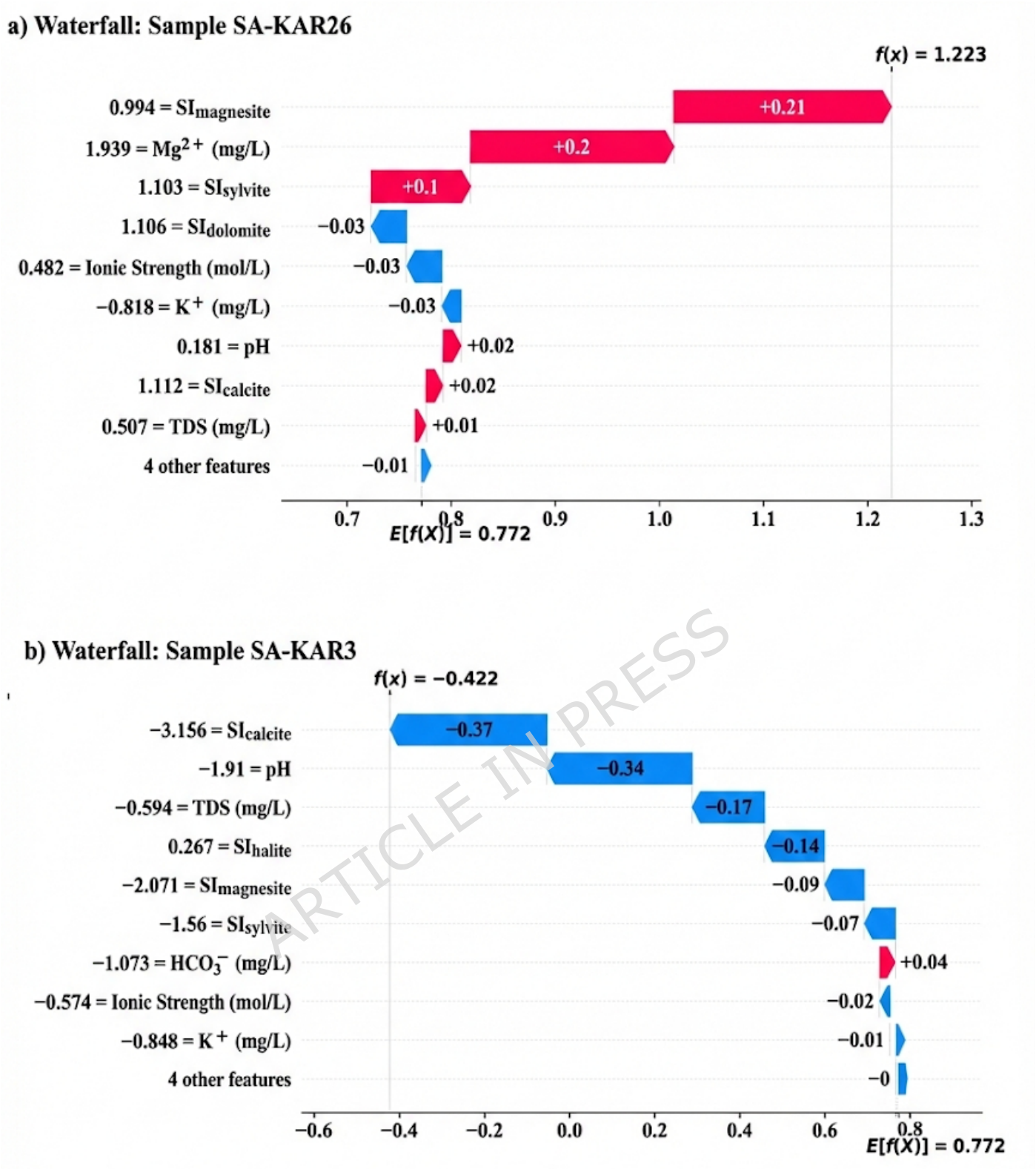


Fig. 8a-b. SHAP waterfall plots for representative groundwater samples (a) KAR26, (b) KAR3. Each panel shows how individual geochemical features drive the model prediction f(x) from the baseline expectation E[f(X)] = 0.772, with blue bars indicating negative contributions and red bars indicating positive contributions.

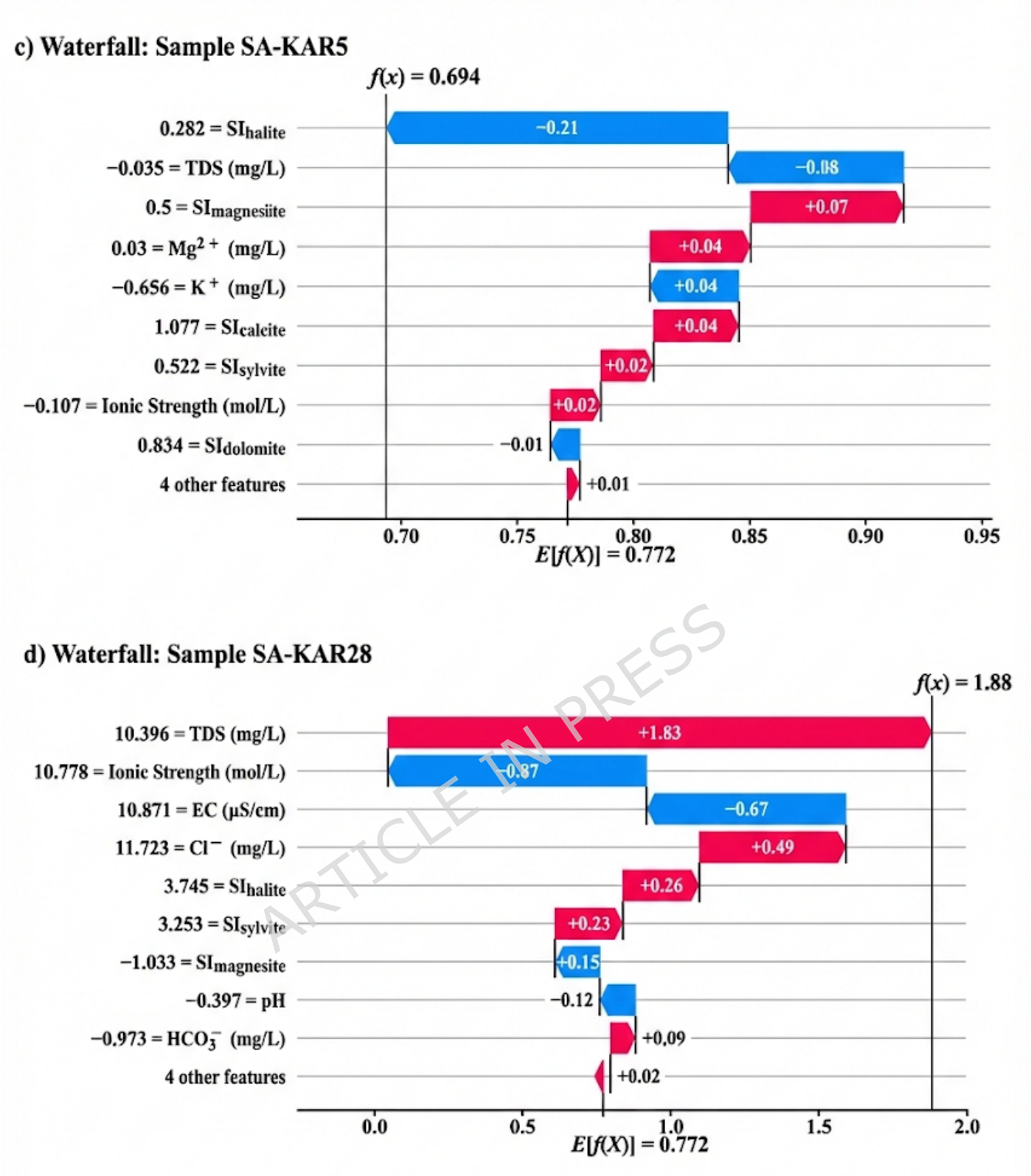


Fig. 8c-d. SHAP waterfall plots for representative groundwater samples (c) KAR5, and (d) KAR28. Each panel shows how individual geochemical features drive the model prediction f(x) from the baseline expectation E[f(X)] = 0.772, with blue bars indicating negative contributions and red bars indicating positive contributions.

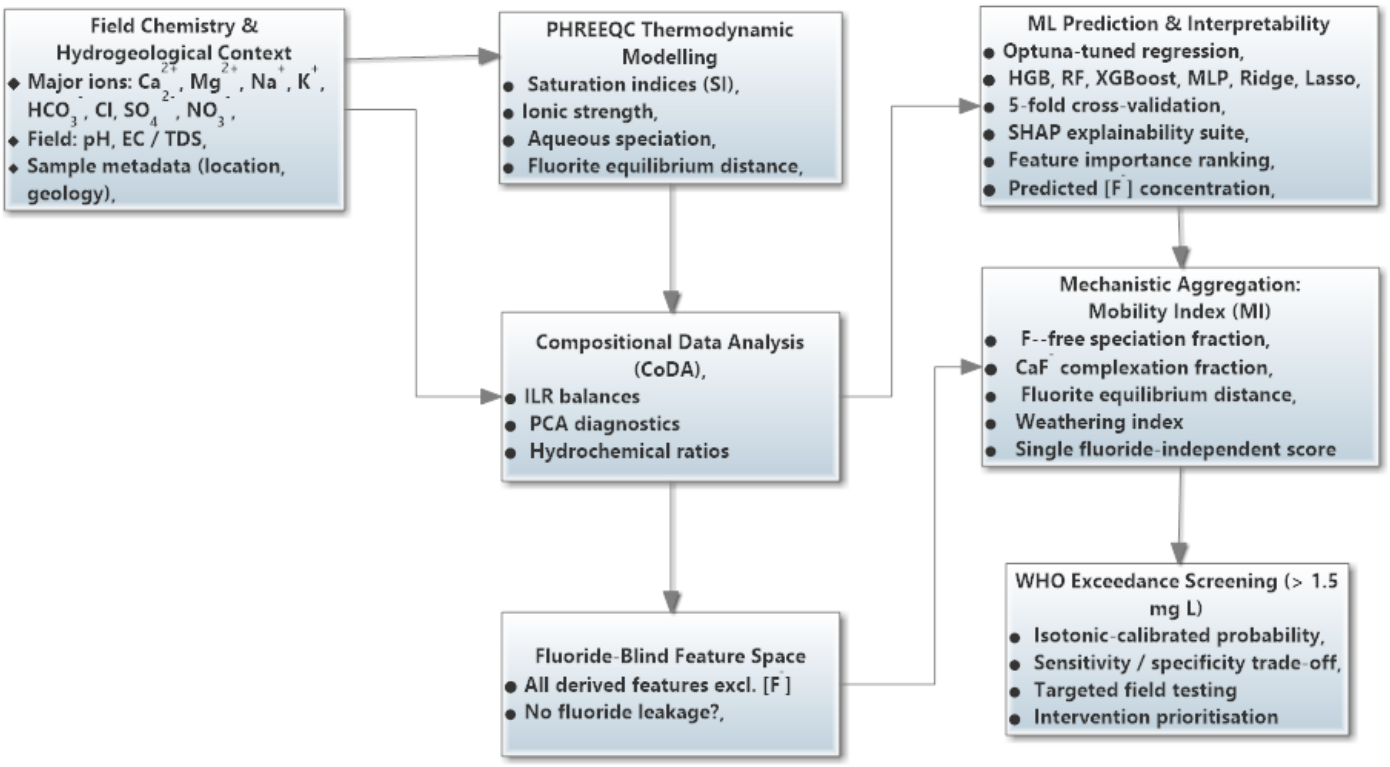


Fig. 9. Conceptual workflow linking hydrogeochemistry to operational screening. Field chemistry and hydrogeological context (major ions, pH, EC/TDS) are interpreted through PHREEQC-derived thermodynamic descriptors and compositional balances (CoDA/ilr). These fluoride-blind features support ML prediction and interpretability (feature importance/SHAP). Mechanistic components are aggregated into the fluoride-independent Mobility Index (MI), which is then calibrated for WHO exceedance screening to support targeted testing and intervention prioritisation