

Article

Integrating Hydrochemistry and Explainable Machine Learning for Groundwater Quality Assessment in the Bismil Plain, Türkiye

Sevgi Özgür Geter ^{1,*}, Süreyya Betül Rufaioğlu ², Ali Volkan Bilgili ² and Güzel Yılmaz ¹

¹ The Department of Environmental Engineering, Faculty of Engineering, Harran University, Sanliurfa 63300, Türkiye; gyilmaz@harran.edu.tr

² The Department of Plant Nutrition and Soil Science, Faculty of Agriculture, Harran University, Sanliurfa 63300, Türkiye; sureyyarufaoglu@harran.edu.tr (S.B.R.); vbilgili@harran.edu.tr (A.V.B.)

* Correspondence: sevgiozgurgeter@gmail.com

Abstract

This study evaluates groundwater quality in the Bismil Plain (Diyarbakır, Southeast Türkiye) using a total of 208 samples collected from 26 wells during eight seasonal sampling periods conducted between 2022 and 2024. In each sample, pH, electrical conductivity (EC), and the major ions Ca^{2+} , Mg^{2+} , Na^+ , K^+ , Cl^- , SO_4^{2-} , HCO_3^- and NO_3^- were analyzed, and a WHO-based Water Quality Index (WQI) was calculated for every observation. The study combines classical hydrochemical interpretation methods, including descriptive statistics, hierarchical correlation analysis, variance inflation factor, and Piper and Gibbs diagrams, with an explainable machine learning framework integrating SHAP-based feature selection into Random Forest, XGBoost, support vector regression, and stacking ensemble models. In addition, spatial residuals were evaluated using Moran's I and ordinary kriging, anomalies were identified using Isolation Forest and Local Outlier Factor algorithms, and predictive uncertainty was quantified through bootstrap resampling. WQI values ranged from 79.37 to 125.48 (mean: 99.67), with all samples classified only within the "Good" (49.5%) and "Poor" (50.5%) quality categories, indicating that the aquifer is close to a critical water-quality threshold. Spatially, the highest (poorest-quality) WQI values form a coherent zone in the south-western and central parts of the plain, whereas the central-eastern wells return the lowest values; the same pattern is reproduced by all four models. XGBoost and the stacking ensemble models showed comparable predictive performance ($R^2 = 0.911$ and 0.910 ; $\text{RMSE} = 3.29$ and 3.27 , respectively), while SHAP analysis identified EC as the dominant controlling factor, followed by NO_3^- , SO_4^{2-} , Ca^{2+} , Mg^{2+} and Cl^- (mean $|\text{SHAP}| = 6.86, 1.57, 1.10, 1.09, 0.85$ and 0.72 WQI units, respectively). Moran's I computed on the residual fields was -0.067 ($p = 0.275$) for XGBoost and -0.068 ($p = 0.273$) for the stacking ensemble, so ordinary kriging of these residuals produced an essentially null correction, whereas the SVR residuals remained spatially autocorrelated ($I = 0.242$; $p = 0.001$) and were meaningfully corrected by the geostatistical step. The originality of the study lies in integrating explainable machine learning, geostatistical residual analysis, anomaly detection, and bootstrap-based uncertainty assessment within a unified framework for a multi-season groundwater dataset, while also evaluating the effectiveness of spatial correction using a Moran's I-based approach.



Academic Editors: Changwei Li, Wenchang Liu and Gang Liu

Received: 23 June 2026

Revised: 26 July 2026

Accepted: 29 July 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: groundwater quality; water quality index; Bismil Plain; SHAP; explainable machine learning; hydrochemistry; residual kriging

1. Introduction

Groundwater is a critical resource in semi-arid agricultural regions where surface water alone cannot meet irrigation demand. Its long-term usability depends on natural recharge dynamics, on lithological controls, and on the management of anthropogenic loading from fertilizers, manure and wastewater [1]. The World Health Organization recommends that physico-chemical indicators be interpreted jointly against health-based limits. For this reason the Water Quality Index (WQI), originally developed by [2] and refined by [3] and many subsequent authors, has become a near-standard synthesis tool. It combines ten or more parameters into a single value that decision-makers can interpret directly, while the underlying weighting structure remains explicit [4,5].

The Bismil Plain, located within the Tigris Basin in Southeast Türkiye, is among the most intensively cultivated districts of Diyarbakır Province. Almost the entire territory ($\approx 95\%$ of 1546 km^2) is dedicated to agriculture, dominated by cereals and, in recent decades, increasingly by cotton supported by GAP-related irrigation infrastructure. This combination of irrigation return flows, fertilizer application, and a semi-arid climate regime creates conditions under which both salinity and nitrate accumulate in the aquifer. The surface waters of the same plain were recently characterized, revealing significant salinity stress and elevated SO_4^{2-} and Cl^- concentrations in the streams, with associated local health-risk concerns [6]. Broader hydrogeochemical and rare-earth-element variability has also been reported across thermal and abandoned oil-well systems in Diyarbakır [7]. However, a focused, multi-season groundwater quality assessment of the Bismil aquifer system has remained lacking.

Machine learning has been applied to water-quality indices with increasing frequency over the last three years, and the emphasis has gradually shifted from predictive accuracy alone towards interpretability. Hussein et al. [8] compared XGBoost, support vector regression and k-nearest neighbours for the irrigation water quality index of a semi-arid Algerian aquifer and found gradient boosting to be both the most accurate and the most stable model. Makumbura et al. [9] coupled Random Forest, LightGBM and XGBoost with SHAP and showed that the resulting ranking of drivers is hydrochemically interpretable rather than purely statistical. Choudhary et al. [10] reached a similar conclusion with a stacked ensemble regression, and Yang et al. [11] combined an entropy-weighted index, LightGBM and SHAP with an explicit treatment of spatial uncertainty, reporting that maps produced by models of comparable accuracy may nevertheless differ substantially in space. Ahn et al. [12] obtained the same result at national scale using averaged neural networks and geospatial analysis. These studies establish SHAP-based attribution as a standard component of water-quality modelling, but they also show that the spatial component of the prediction is still treated inconsistently.

Two methodological tendencies in the regional WQI literature limit how much such studies can contribute. First, a substantial number of recent regional papers couple WQI calculation with machine learning and report leaderboard-style accuracy comparisons (e.g., Random Forest vs. XGBoost vs. SVR vs. ANN) without quantifying which input variables actually drive the predicted index. The outcome is a model that predicts the index well but does not indicate what should be managed [13,14]. Second, hybrid frameworks that combine machine learning with residual kriging are commonly applied without verifying whether the residuals carry meaningful spatial structure in the first place. Such a check is straightforward to perform with Moran's I [15], but it is too often skipped, so the geostatistical step is presented as a uniform improvement even in cases where it has almost no effect on the predictions.

The present study addresses both gaps explicitly. We integrate hydrochemical interpretation [16,17], a SHAP-guided machine learning pipeline [18–21], ordinary kriging of

the model residuals conditioned on the outcome of a Moran's I test, two-stage anomaly screening with Isolation Forest [22] and Local Outlier Factor [23], and bootstrap-based uncertainty quantification into a single workflow calibrated for the realities of small-sample agricultural groundwater monitoring. The specific objectives are: (i) to evaluate groundwater quality in the Bismil Plain through a WHO-based WQI computed for eight seasonal periods; (ii) to identify, by means of SHAP, the hydrochemical and anthropogenic variables that control WQI variability; and (iii) to test whether kriging of the model residuals actually improves predictive accuracy, using Moran's I as the decision criterion. The originality of the study lies in combining SHAP-based explainable machine learning, residual kriging, anomaly detection, and bootstrap-based uncertainty analysis within the same analytical framework using a multi-season groundwater dataset, thereby enabling the simultaneous evaluation of both systemic water-quality pressures and locally critical anomalies.

2. Materials and Methods

2.1. Study Area

The study area lies in the Bismil Plain in the eastern part of Diyarbakir Province, within the Tigris Basin in Southeast Türkiye. The sampling network occupies a longitudinal range of approximately 40.805° – 41.301° E and a latitudinal range of 37.688° – 37.896° N (Figure 1). The plain is bounded by the Silvan Plateau to the north and the Mardin Threshold Plateau to the south, with the Tigris River flowing through the central depression. The Bismil Plain is a region where agriculture is intensively practiced, both due to its water resources and its climate. Climate conditions are semi-arid, with mild wet winters and hot dry summers; agriculture occupies almost the entire administrative territory and depends increasingly on irrigation supplied through the GAP infrastructure. The hydrogeological setting is therefore well suited for examining both natural rock-water interaction and the superimposed footprint of agricultural management on groundwater chemistry.

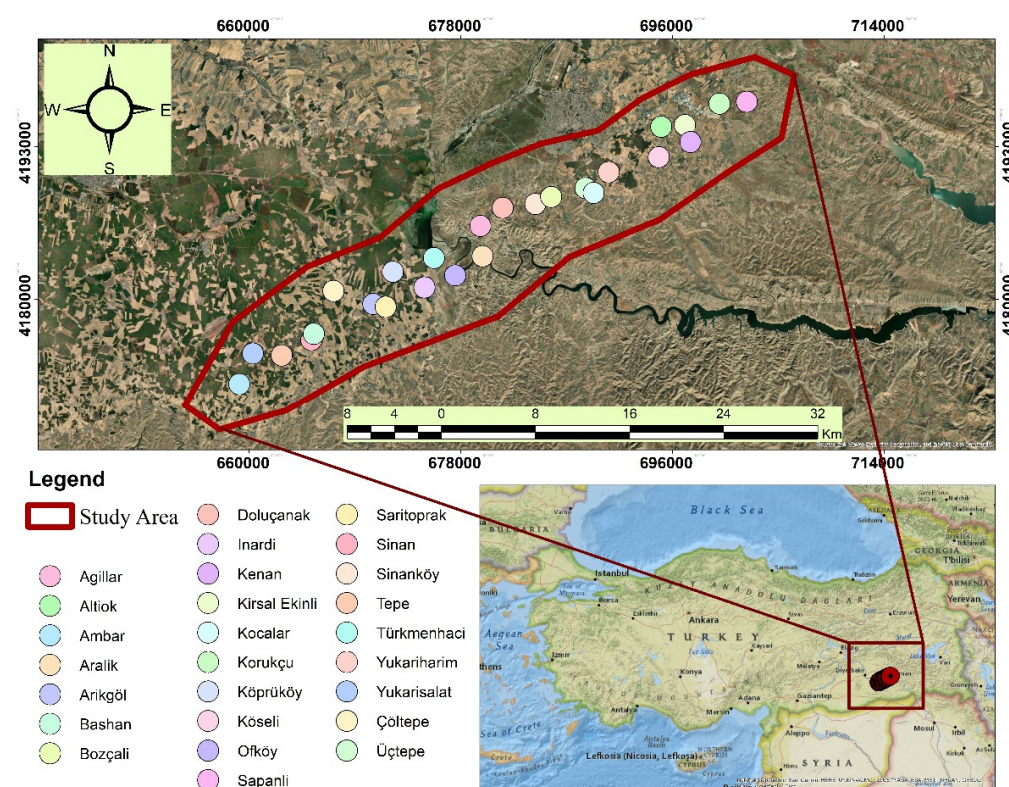


Figure 1. Location of the study area within Diyarbakir Province (Bismil Plain, Tigris Basin, Southeast Türkiye).

Geologically, the plain is floored by Quaternary alluvial deposits of the Tigris River, which overlie the Oligocene–Miocene succession of the Southeast Anatolian foreland basin. The irrigation and supply wells sampled in this study are relatively deep (Section 2.2) and draw on this deeper basin sequence rather than on the surficial alluvium alone. The regional sequence comprises the shallow-marine and reefal carbonates of the Firat Formation and, in the eastern and northern parts of the region, evaporitic units—principally the Germik Formation (Oligocene) and the Gerçüş Formation (Palaeocene)—in which gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$) and anhydrite (CaSO_4) occur as massive, laminated, banded and nodular lithofacies [24,25]. Because the wells reach these carbonate and evaporite-bearing units, dissolution of calcite, dolomite, gypsum and anhydrite is the expected geogenic source of Ca^{2+} , Mg^{2+} , HCO_3^- and SO_4^{2-} , whereas the irrigated land surface is the main anthropogenic source of NO_3^- and of part of the Cl^- .

2.2. Sampling and Analytical Programme

The dataset consists of 208 groundwater observations collected from 26 wells over eight sampling campaigns carried out between 2022 and 2024. The campaigns were organized into two annual cycles labelled Y1-Winter, Y1-Spring, Y1-Summer, Y1-Autumn, Y2-Winter, Y2-Spring, Y2-Summer and Y2-Autumn, so that every well contributes both spatial and seasonal variability to the same analytical frame. The wells are irrigation and domestic supply wells operated by the State Hydraulic Works (DSİ) and by local users; their estimated depths range from 135 to 269 m, with a mean of about 200 m, so they draw on the deeper basin aquifer that includes the carbonate and evaporite-bearing units described in Section 2.1. Before each sampling the wells were pumped until the pH and EC readings stabilised, which removed stagnant borehole water. Samples were then collected in pre-rinsed polyethylene bottles, stored at approximately 4 °C and transported to the laboratory in accordance with the relevant Turkish Standards (TS 5090 and TS 5106) and the national guidelines for water-quality sampling. Well coordinates (X, Y) were recorded with handheld GPS receivers, and well identifiers and campaign labels were preserved for repeated-measures analysis.

The analytical programme followed national and international standard methods. pH was measured with a glass-electrode digital pH meter according to TS EN ISO 10523, the instrument being calibrated with pH 4.00, 7.00 and 10.00 buffer solutions before each measurement session. Electrical conductivity was determined conductometrically according to TS EN 27888 (ISO 7888) and is reported in $\mu\text{S}/\text{cm}$ at a reference temperature of 25 °C. Bicarbonate was calculated from total alkalinity, measured by strong-acid titration according to TS 3790 EN ISO 9963-1. The anions (NO_3^- , Cl^- , SO_4^{2-}) and the cations (Ca^{2+} , Mg^{2+} , Na^+ , K^+) were determined by ion chromatography: samples were filtered through 0.45 μm membrane filters and injected onto dedicated anion and cation separation columns with conductivity detection, and concentrations were obtained from multi-point calibration curves, following ISO 10304-1 for anions and ISO 14911 for cations. All procedures are consistent with Standard Methods [26]. Quality assurance followed a fixed laboratory protocol: instruments were calibrated daily, quality-control standards and blank samples were analysed with every batch, and replicate analyses were performed where necessary. The accuracy of the complete analysis was additionally verified through the ionic charge-balance error described in Section 2.3.

2.3. Data Preprocessing

No missing values were detected in the dataset; therefore, no imputation procedure was required. The internal consistency of every analysis was verified with the ionic charge-balance error (CBE), calculated in milliequivalents per litre following the ion-balance check

of Standard Methods (APHA 1030 E) [26] as $CBE = 100 \times (\Sigma cations - \Sigma anions) / (\Sigma cations + \Sigma anions)$. The cation sum comprises Ca^{2+} , Mg^{2+} , Na^+ and K^+ and the anion sum comprises Cl^- , SO_4^{2-} , HCO_3^- and NO_3^- . The mean CBE of the dataset is -3.74% , which lies within the $\pm 10\%$ range conventionally regarded as acceptable for groundwater of moderate to high mineralisation [27,28]. The small negative sign indicates a slight anion excess, consistent with the fact that the analytical programme did not include the minor anions; the balance was used as a quality indicator for the dataset as a whole and no sample was excluded on this basis. Outlier screening was performed using a two-stage approach rather than a single rule-based filter. First, potential anomalies were explored through boxplots and z-score distributions. Subsequently, model-based anomaly detection was applied using Isolation Forest [22] and Local Outlier Factor [23]. Isolation Forest is an ensemble of random binary trees in which the feature space is recursively partitioned by selecting a variable at random and a split value at random between its minimum and maximum. Observations that lie far from the bulk of the data are separated after only a few splits, so the average path length required to isolate a sample serves as its anomaly score; the algorithm therefore detects global anomalies, that is, samples with unusual combinations of values. The Local Outlier Factor adopts a density-based view: for each sample it compares the local reachability density of that sample with the densities of its k nearest neighbours, and a ratio appreciably greater than unity indicates that the sample lies in a sparser region than its neighbourhood. It therefore detects local anomalies, which may remain inside the global range of every individual variable. Both algorithms were applied to the standardised matrix of the six retained predictors. Isolation Forest was run with a fixed contamination level of 0.10, whereas the Local Outlier Factor was run with $k = 20$ neighbours and with the automatic contamination threshold of the implementation, so that the number of local anomalies is determined by the data rather than fixed in advance. This difference in setting explains why the two algorithms return sets of very different size (Section 3.8). Because the input variables spanned substantially different numerical ranges (e.g., $pH \approx 7.6$ versus $EC \approx 1500 \mu S/cm$), all numerical features were standardized to zero mean and unit variance prior to machine learning analysis. This preprocessing step was particularly important for scale-sensitive algorithms such as Support Vector Regression (SVR) and also improved the reliability of distance-based anomaly detection methods.

2.4. Hydrochemical Interpretation

Descriptive statistics, including minimum, maximum, mean, standard deviation, coefficient of variation, skewness, and kurtosis, were calculated for all parameters. To evaluate the systematic component of seasonal variability, the seasonal mean and standard deviation values of the two annual sampling cycles were pooled by season. Relationships among the parameters were assessed using a Pearson correlation matrix organized through hierarchical clustering, with the WQI included as an additional variable so that its association with each ion can be read directly from the matrix. Multicollinearity was quantified using the Variance Inflation Factor (VIF), calculated by regressing each parameter against the remaining variables. The commonly accepted threshold of $VIF < 5$ was used to identify potential redundancy among variables and to support the interpretation of SHAP-based feature contributions. To identify hydrochemical facies, the major cation and anion concentrations were normalized in milliequivalents per litre and evaluated using the Piper trilinear diagram [16]. In addition, the Gibbs approach [17] was applied by examining the relationship between EC and the $Na^+ / (Na^+ + Ca^{2+})$ ratio in order to distinguish the dominant controlling mechanisms of groundwater chemistry, including rock–water interaction, evaporation–crystallization processes, and atmospheric contributions.

2.5. Water Quality Index

The Water Quality Index was computed for every sample following the WHO-based weighted arithmetic formulation [29]. For each parameter k , the quality rating q_k is computed as $q_k = [(C_k - I_k)/(S_k - I_k)] \times 100$, where C_k is the measured concentration, S_k is the WHO standard and I_k is the ideal value (taken as 7 for pH and zero for the ions). The relative weight of each parameter is $W_k = w_k/\sum w_k$, where w_k are the unit weights (5 for NO_3^- ; 4 for pH and EC; 3 for Cl and SO_4^{2-} ; 2 for Ca, Mg and Na; 1 for K and HCO_3^-). The final index is $\text{WQI} = \sum (W_k q_k)$. Samples were classified into the conventional five categories: Excellent (<50), Good (50–100), Poor (100–200), Very Poor (200–300) and Unsuitable (>300).

2.6. SHAP-Guided Feature Selection

An initial Random Forest regressor was trained on the full feature set, and SHAP (SHapley Additive exPlanations) [18] values were computed using a tree-based explainer. For this selection step the input matrix comprised the ten measured parameters (pH, EC, Ca^{2+} , Mg^{2+} , Na^+ , K^+ , Cl^- , SO_4^{2-} , HCO_3^- and NO_3^-) and the output variable was the WQI of the same sample. Features whose mean absolute SHAP contribution exceeded the median across all features were retained for the subsequent modelling stage. This conservative selection rule is adapted to small-sample settings: it removes low-importance noise variables while preserving every parameter that contributes more than the typical signal level. The retained features were EC, Ca^{2+} , Mg^{2+} , Cl^- , SO_4^{2-} and NO_3^- .

2.7. Machine Learning Models

2.7.1. Models

Four predictive models were compared. In all four cases the input matrix X consisted of the six retained hydrochemical predictors (EC, Ca^{2+} , Mg^{2+} , Cl^- , SO_4^{2-} and NO_3^-) and the output variable y was the WQI of the corresponding sample. No coordinate or categorical variable was used as a predictor, so that the spatial information remained entirely available for the independent residual analysis described in Section 2.8. Random Forest [19] was selected as a robust, low-variance ensemble of decision trees that captures non-linearities and pairwise interactions without explicit specification. XGBoost [20] was included for its sequential boosting and built-in regularization, which often translates into competitive performance on tabular data. Support Vector Regression [21] was included because of its strong generalization properties on small to moderate samples. A stacking ensemble was constructed with Random Forest, XGBoost and SVR as base learners, with their predictions passed to a linear meta-learner.

2.7.2. Hyperparameter Tuning and Validation Strategy

The 208 samples were first assigned to a 3×3 spatial block structure defined by the quantiles of the X and Y coordinates, and the blocks—rather than individual samples—were used as cross-validation units. This is necessary because neighbouring wells that appear simultaneously in the training and the test subsets produce optimistically biased error estimates in dense monitoring networks. Following feature selection, the dataset was divided into 80% training and 20% test subsets, and the test subset was held out from every stage of model development. The same training and test partition, the same standardised input matrix and the same evaluation metrics were used for all four models, so that the comparison reported in Section 3.5 is not affected by differences in the data handling.

Random Forest was taken as the reference model of the study, because it also provides the SHAP attribution reported in Section 3.7, and its hyperparameters were therefore optimised explicitly. The number of trees ($n_estimators$, searched between 50 and 200) and the maximum tree depth (max_depth , searched between 3 and 20) were tuned with

Optuna [30] over 25 trials, the objective being the minimisation of the mean MSE across the nine spatial blocks rather than across conventional random folds; the remaining parameters were kept at their default values. XGBoost was trained with 200 boosting rounds and a maximum depth of 3, a deliberately shallow configuration that limits overfitting on a dataset of this size, and Support Vector Regression was trained with a radial basis function kernel and the standard parameter set ($C = 1$, $\gamma = 1/(n_{\text{features}} \times \text{var}(X))$, $\epsilon = 0.1$) on the standardised feature matrix. These two configurations were kept fixed rather than tuned, so that the comparison in Section 3.5 contrasts model families under a common, widely used specification instead of reflecting how much search effort was spent on each algorithm. The stacking ensemble combines the predictions of the three fitted base learners through a linear meta-learner.

Model performance was evaluated on the held-out test data using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). The complete modelling framework, from field sampling to the management-oriented outputs, is summarised in Figure 2.

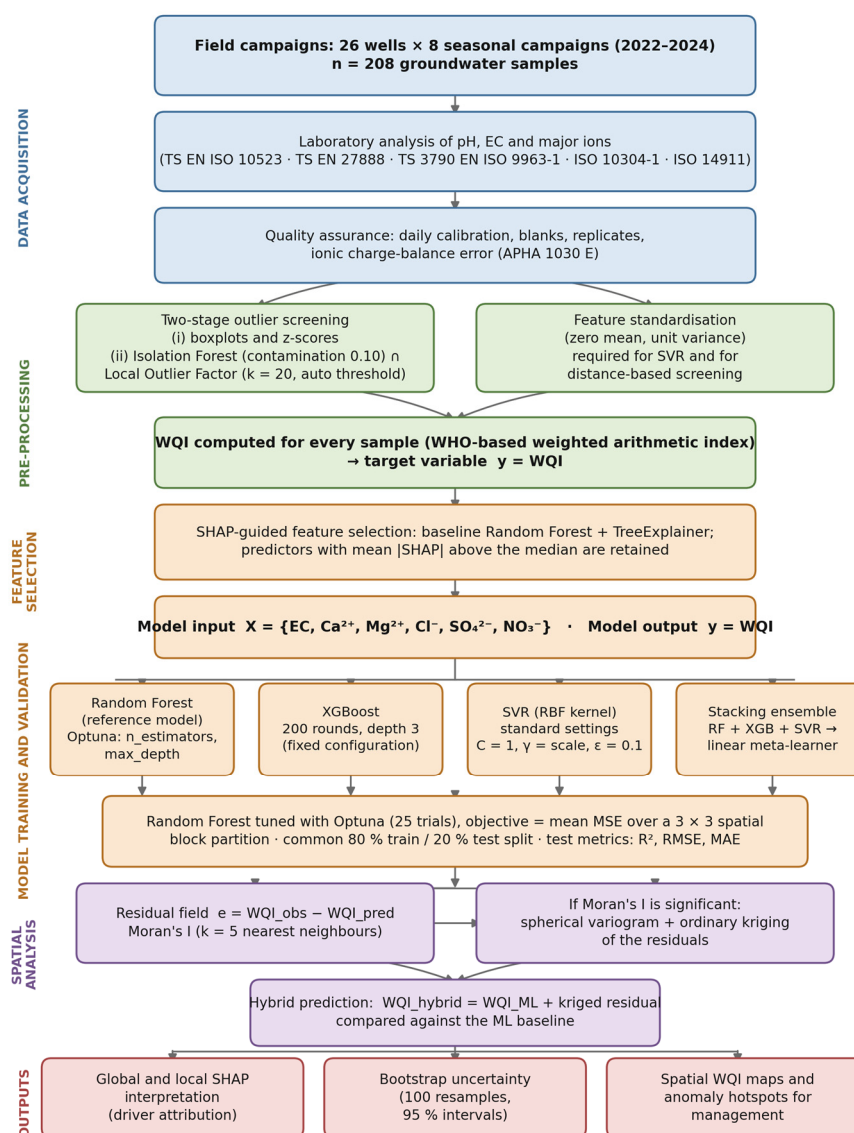


Figure 2. Workflow of the integrated hydrochemical and explainable machine-learning framework applied to the Bismil Plain groundwater dataset.

2.8. Hybrid Spatial Modeling

Spatial residuals were calculated as the difference between the observed WQI values and the predictions obtained from the base machine learning models. These residuals were treated as an independent spatial field, and their spatial dependence was evaluated prior to any geostatistical post-correction. Spatial autocorrelation was assessed using Moran's I [15] based on a $K = 5$ nearest-neighbor weight matrix, the significance of the statistic being obtained from 999 random permutations. Subsequently, a spherical variogram model was fitted under the assumption of second-order stationarity, and ordinary kriging was applied to interpolate the residual field spatially. The final hybrid prediction was reconstructed as the sum of the machine learning prediction and the kriged residual component, allowing the contribution of the kriging-based correction to be directly compared with the original ML baseline. This sequence follows the regression-kriging formulation [31], in which the geostatistical term can only contribute to the extent that the residual random field is genuinely autocorrelated; the Moran's I test is therefore used here as an explicit precondition rather than as a descriptive statistic.

2.9. Anomaly Detection

Two complementary algorithms were combined for anomaly screening. Isolation Forest [22] flags samples that are easy to isolate by random partitioning, which targets samples with extreme combinations of values. Local Outlier Factor [23] flags samples whose local density is anomalous compared with their neighbors, which targets local rather than global anomalies. The two algorithms were combined through a conservative intersection rule. Each of them was run separately on the standardised feature matrix with the settings given in Section 2.3 and returned a binary label for every sample; a sample was accepted as an anomaly only when both algorithms flagged it, that is, when it was simultaneously easy to isolate in the global feature space and located in a locally sparse neighbourhood. The intersection thus discards samples that are flagged only as a consequence of the fixed contamination rate of a single algorithm and retains those for which two independent criteria agree. The union of the two lists was also examined but was not used for interpretation. The intersection of the two lists was treated as the most reliable anomaly set, and these samples were re-examined hydrochemically rather than removed.

2.10. Explainable AI

After the final Random Forest model was retrained on the selected features, SHAP values were re-computed both globally (mean absolute contribution per feature, ranking the systemic drivers of WQI) and locally (per-sample SHAP contributions, which support interpretation of individual problematic wells). Beeswarm summary plots additionally encode the direction of each feature's effect, allowing implicit reading of pairwise interactions.

2.11. Bootstrap Uncertainty Estimation

Predictive uncertainty was estimated through 100 bootstrap resamples of the training set. For every test observation the predictions across the resampled models were used to compute 95% confidence intervals from the 2.5th and 97.5th percentiles. The mean and median width of these intervals are reported as a sample-level stability metric.

2.12. Summary of the Modelling Framework

Figure 2 summarises the complete analytical framework applied in this study. The workflow runs from the field campaigns and the standardised laboratory programme, through quality assurance, outlier screening, standardisation and the computation of the WQI, to the SHAP-guided selection of the predictors. The four models are then

trained and validated under a spatial block scheme, followed by the conditional residual geostatistics. The framework ends with three interpretation outputs used for management: SHAP-based driver attribution, bootstrap uncertainty and spatial WQI mapping with the associated anomaly hotspots. All computations were performed in Python 3 using scikit-learn (Random Forest, SVR, Isolation Forest, Local Outlier Factor, scaling and resampling), XGBoost, SHAP (TreeExplainer), Optuna, libpysal and esda (K-nearest-neighbour weights and Moran's I) and PyKrig (variogram fitting and ordinary kriging). The random seed was fixed at every stochastic step so that the reported results are reproducible, and the analysis script is available from the corresponding author on request.

3. Results

3.1. Hydrochemical Characteristics

The descriptive statistics of the ten measured parameters are summarised in Table 1. The mean sum of the analysed ions is 1486 mg/L, which corresponds to a total dissolved solids content of the order of 1200–1300 mg/L and places the samples slightly above the conventional 1000 mg/L limit between fresh and brackish groundwater [29]; the aquifer is therefore described here as moderately mineralised in this quantitative sense. Mean pH was 7.61 (range 7.20–8.10), confirming a near-neutral to slightly alkaline character throughout the network. Mean EC was 1499 $\mu\text{S}/\text{cm}$ with a maximum of 2174 $\mu\text{S}/\text{cm}$, placing several samples beyond the WHO desirable threshold of 1500 $\mu\text{S}/\text{cm}$. The mean concentrations of the individual ions are listed in Table 1 and are not repeated here. Coefficients of variation are below 21% for the majority of parameters but reach 40.87% for NO_3^- , signaling pronounced spatial heterogeneity in nitrogen loading. The skewness and kurtosis of NO_3^- (1.13 and 3.45 respectively) corroborate the presence of localized contamination peaks. Seven of 208 samples (3.4%) exceeded the WHO drinking-water limit of 50 mg/L for NO_3^- , while 108 samples (51.9%) exceeded the more conservative national and EU guideline value of 25 mg/L. The seasonal mean values compiled in Table 2 and Figure 3 show that EC and Cl^- peak in summer (≈ 1599 $\mu\text{S}/\text{cm}$ and ≈ 315 mg/L) and are lowest in winter (≈ 1425 $\mu\text{S}/\text{cm}$ and ≈ 278 mg/L). This is consistent with evaporative concentration and irrigation return flow during the dry season. In contrast, HCO_3^- follows the opposite pattern, peaking in spring at ≈ 472 mg/L. Inter-annual differences between Y1 and Y2 are negligible (mean WQI = 99.82 and 99.52), so the seasonal signal dominates over the inter-annual one within this two-year window. The spatial distribution of mean parameter concentrations across the 26 wells (Figure 4) reveals that the highest EC, Cl^- and SO_4^{2-} values cluster in the south-western and central portions of the plain, whereas the eastern part of the network shows comparatively lower mineralization.

Table 1. Descriptive statistics of groundwater physico-chemical parameters across the 208 samples collected in the Bismil Plain.

Parameters	Unit	Min	Max	Mean	Std. Dev.	CV (%)	Skewness	Kurtosis
pH	–	7.20	8.10	7.61	0.17	2.18	0.29	0.05
EC	$\mu\text{S}/\text{cm}$	892.39	2174.61	1499.01	310.27	20.70	−0.02	−0.97
Ca	mg/L	126.32	272.19	184.59	29.47	15.97	0.26	−0.62
Mg	mg/L	46.06	129.51	74.85	13.88	18.55	0.56	0.56
Na	mg/L	62.06	196.82	129.41	25.20	19.47	0.21	−0.19
K	mg/L	4.18	8.76	6.05	0.89	14.64	0.34	−0.24
Cl	mg/L	158.20	434.80	292.69	48.27	16.49	0.34	−0.09
SO_4	mg/L	181.20	475.70	324.72	57.21	17.62	0.10	−0.30
HCO_3	mg/L	329.44	578.14	447.46	54.38	12.15	−0.01	−0.52
NO_3	mg/L	3.03	79.59	26.60	10.87	40.87	1.13	3.45

Table 2. Seasonal variability of groundwater physico-chemical parameters across sampling periods. Values are mean $\pm$ standard deviation.

Season	pH	EC	Ca	Mg	Na	K	Cl	SO ₄	HCO ₃	NO ₃
Autumn	7.61 $\pm$ 0.15	1523.77 $\pm$ 309.54	182.06 $\pm$ 28.03	73.98 $\pm$ 13.42	129.69 $\pm$ 23.17	6.10 $\pm$ 0.86	297.41 $\pm$ 45.65	323.45 $\pm$ 57.70	449.07 $\pm$ 55.11	27.05 $\pm$ 10.50
Spring	7.69 $\pm$ 0.18	1448.55 $\pm$ 309.33	181.90 $\pm$ 29.57	76.74 $\pm$ 16.32	126.54 $\pm$ 25.37	5.96 $\pm$ 0.88	281.02 $\pm$ 45.20	321.19 $\pm$ 55.82	471.75 $\pm$ 49.39	27.17 $\pm$ 13.75
Summer	7.57 $\pm$ 0.15	1599.05 $\pm$ 313.66	187.41 $\pm$ 33.37	75.89 $\pm$ 12.93	131.30 $\pm$ 27.88	5.95 $\pm$ 0.84	314.55 $\pm$ 52.31	345.66 $\pm$ 52.60	425.99 $\pm$ 50.79	26.75 $\pm$ 11.20
Winter	7.59 $\pm$ 0.15	1424.68 $\pm$ 286.00	186.99 $\pm$ 26.90	72.80 $\pm$ 12.61	130.14 $\pm$ 24.62	6.19 $\pm$ 0.96	277.79 $\pm$ 41.43	308.56 $\pm$ 57.83	443.03 $\pm$ 53.42	25.45 $\pm$ 7.30

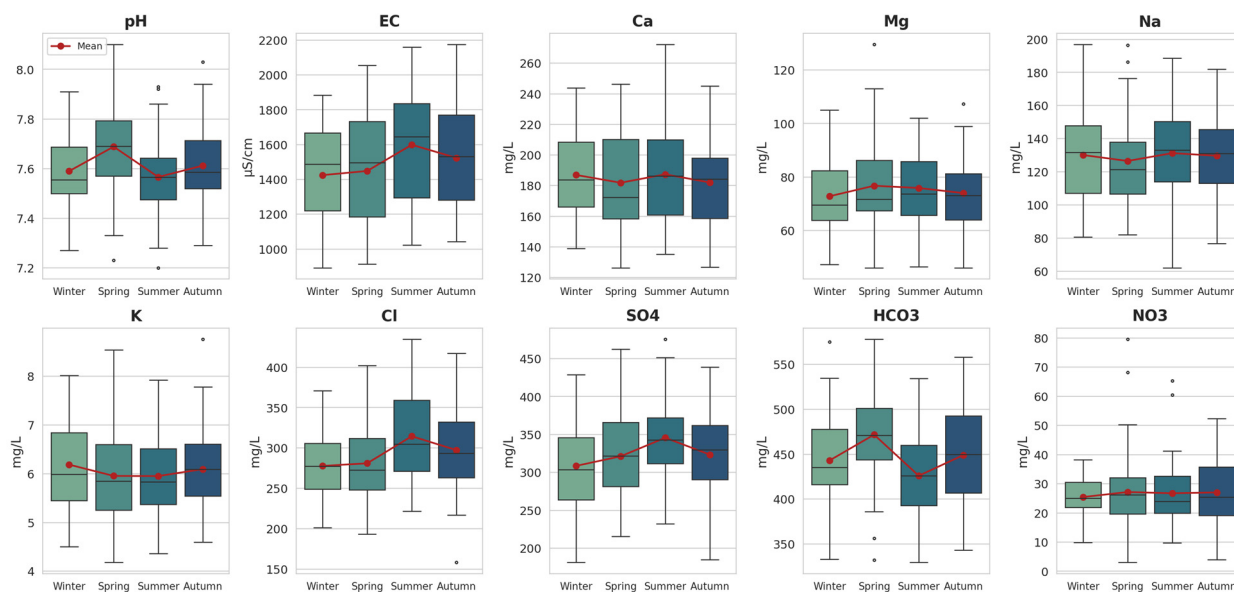


Figure 3. Seasonal variability of the ten physico-chemical parameters illustrated by boxplots; the red line connects seasonal means.

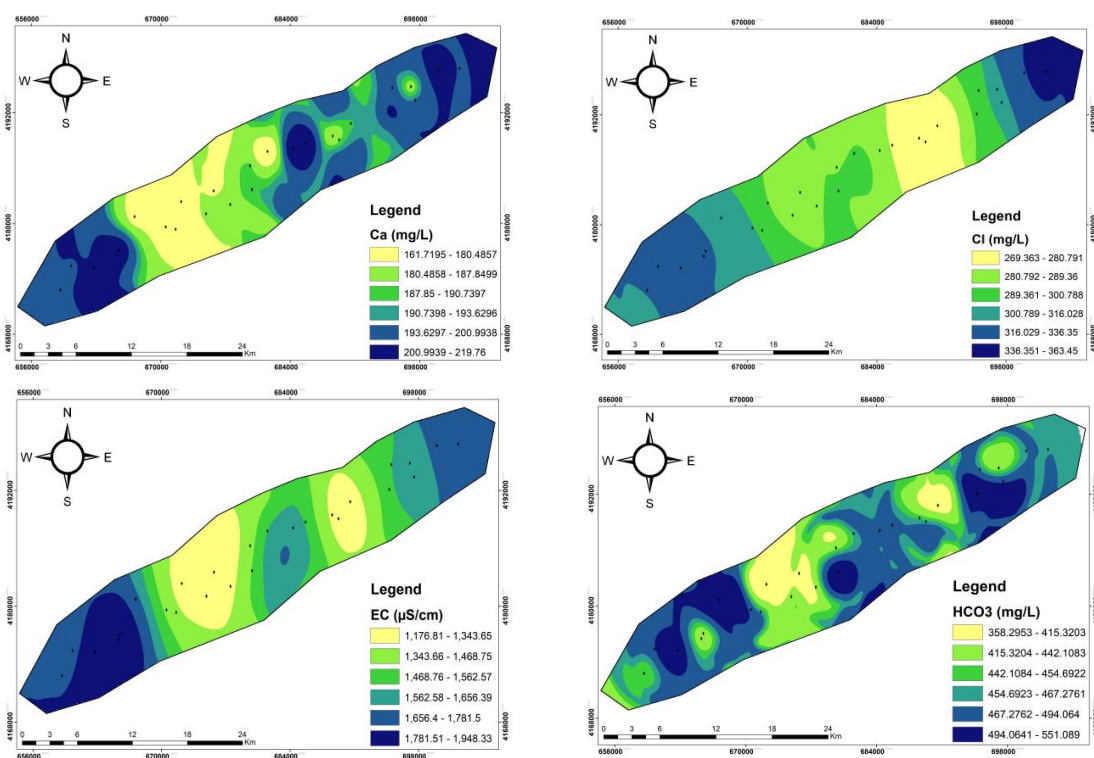


Figure 4. Cont.

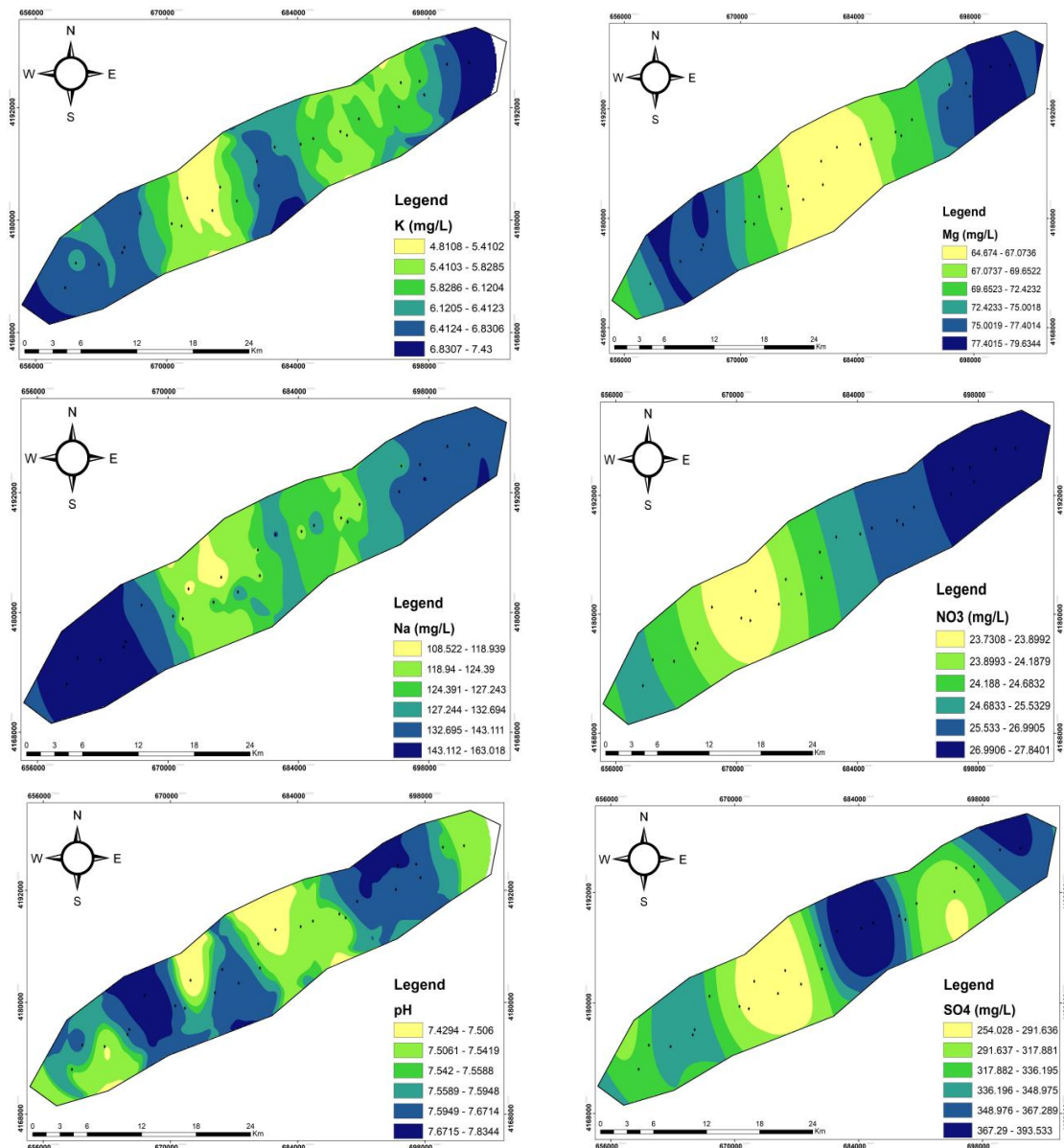


Figure 4. Spatial distribution of well-mean values for the ten hydrochemical parameters, showing the heterogeneity within the Bismil Plain.

3.2. Correlation Structure and Multicollinearity

The hierarchical correlation matrix (Figure 5) places EC at the centre of a strongly correlated cluster that includes SO_4^{2-} , Cl^- , Ca^{2+} , Mg^{2+} and HCO_3^- , with Pearson coefficients between EC and these parameters ranging from 0.55 to 0.86. NO_3^- shows only a weak correlation with this group of salinity-related variables and with the index itself ($r = 0.19$ with WQI). Nitrate variability is therefore governed by a process that differs from the rock–water interaction controlling the remaining ions, most plausibly by agricultural loading applied at the land surface. The pH variable is essentially uncorrelated with WQI ($r \approx 0$), consistent with the narrow alkaline range observed in the dataset. The VIF analysis (Table 3) returns values below 4 for all parameters, with EC at 3.95 marking the upper end of the moderate-collinearity range. No parameter exceeded the conventional $\text{VIF} = 5$ threshold, so all input variables can be retained without collinearity-driven instability.

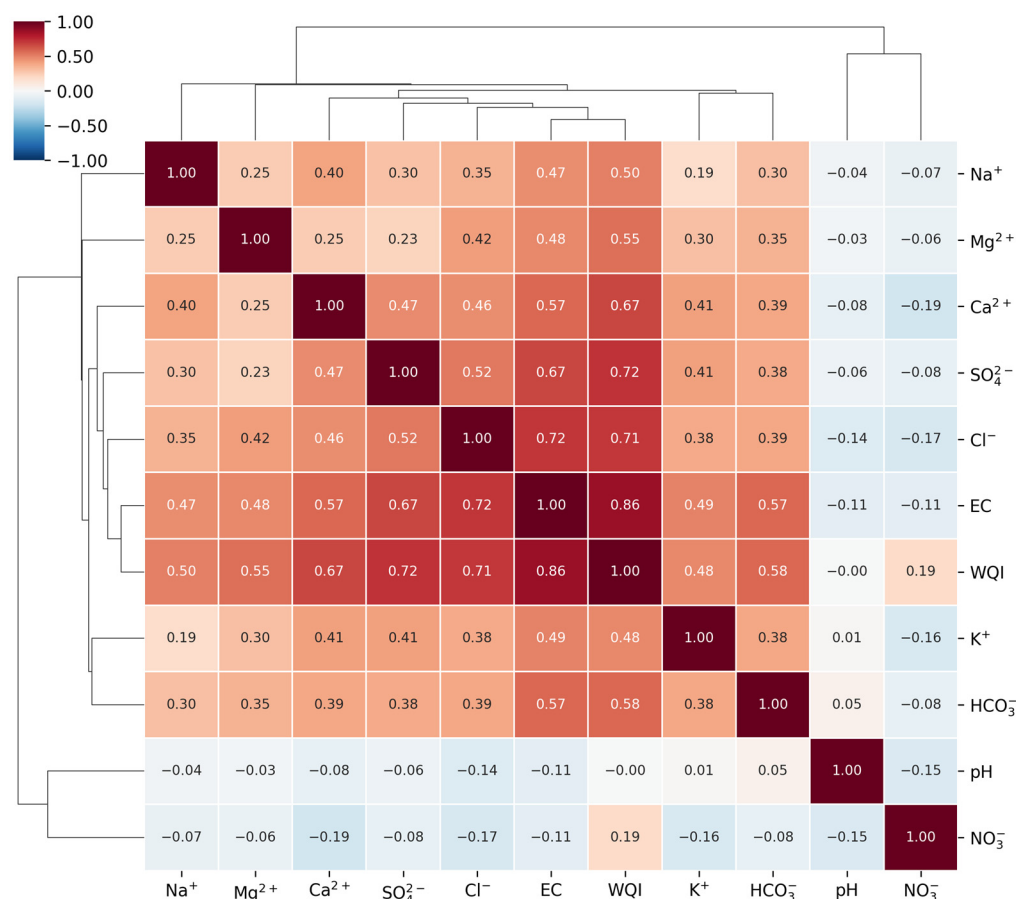


Figure 5. Hierarchically clustered Pearson correlation matrix of groundwater parameters and the WQI.

Table 3. Pearson correlation with WQI, mean SHAP rank in the final Random Forest model, and Variance Inflation Factor (VIF) of each input parameters.

Parameters	Pearson r vs. WQI	Mean SHAP Rank	VIF	Multicollinearity
EC	0.860	1	3.95	Moderate
NO ₃	0.188	2	1.10	Low
SO ₄	0.719	3	1.92	Low
Ca	0.667	4	1.68	Low
Mg	0.553	5	1.39	Low
Cl	0.708	6	2.19	Low
HCO ₃	0.576	—	1.57	Low
Na	0.504	—	1.34	Low
K	0.484	—	1.45	Low
pH	−0.003	—	1.08	Low

3.3. Hydrochemical Facies and Governing Processes

Expressed as percentages of the total cation concentration in meq/L, Ca²⁺ dominates (≈43.5%), followed by Mg²⁺ (≈29.2%) and Na⁺ + K⁺ (≈27.3%). The anions, expressed in the same way as percentages of the total anion concentration in meq/L, are more evenly distributed: Cl⁻ (≈36.9%) slightly exceeds HCO₃⁻ (≈33.0%) and SO₄²⁻ (≈30.2%). On the Piper template (Figure 6a) the samples plot in the mixed Ca²⁺–Mg²⁺–Cl⁻–SO₄²⁻ region, with a subtle drift toward the chloride end-member during summer periods. No sample displays a purely Ca²⁺–HCO₃⁻ composition, which would indicate exclusively carbonate-controlled recharge, and none displays a purely Na⁺–Cl⁻ composition, which would indicate dominant evaporite dissolution or anthropogenic salinization. The Gibbs

interpretation (Figure 6b) places the samples within the rock–water interaction field, with a measurable shift toward the evaporation–crystallization end-member during summer. The combination supports a hydrogeochemical model in which carbonate dissolution provides the background, gypsum/anhydrite contribution explains the elevated SO_4^{2-} , and seasonal evaporative concentration superimposes a Cl^- -driven component. This interpretation is consistent both with the regional stratigraphy described in Section 2.1, in which gypsum- and anhydrite-bearing evaporitic units crop out around the plain [27,28], and with the ionic ratios of the samples: the mean molar $\text{Ca}^{2+}/\text{SO}_4^{2-}$ ratio is 1.36, so calcium exceeds the 1:1 proportion expected from gypsum dissolution alone and an additional carbonate source of Ca^{2+} is superimposed on the evaporite signal.

3.4. WQI Distribution and Classification

Computed across 208 samples, the WQI ranges from 79.37 to 125.48 with a mean of 99.67. The distribution is therefore confined to a narrow band straddling the boundary between the Good and Poor classes (Table 4); 49.5% of the samples fall in the Good class ($50 \leq \text{WQI} < 100$) and 50.5% fall in the Poor class ($100 \leq \text{WQI} < 200$). No samples reach Excellent quality and none falls into Very Poor or Unsuitable. The seasonal pattern is consistent with the EC and Cl^- values: summer samples have a mean WQI of 102.34 against a winter mean of 96.83, a 5.5-unit shift that is enough to move a substantial fraction of samples across the 100 boundary.

Table 4. Distribution of the 208 groundwater samples across the WQI quality classes.

WQI Class	Range	Samples	Percentage (%)
Excellent	<50	0	0.0
Good	50–100	103	49.5
Poor	100–200	105	50.5
Very Poor	200–300	0	0.0
Unsuitable	>300	0	0.0

3.5. Performance of the Machine-Learning Models

Test-set performance of the four models is reported in Table 5. XGBoost and the stacking ensemble are tied at the top, with R^2 values of 0.911 and 0.910, RMSE values of 3.29 and 3.27, and MAE values of 2.38 and 2.36, respectively. Random Forest follows at $R^2 = 0.859$ (RMSE = 4.16; MAE = 3.24), and SVR completes the ranking at $R^2 = 0.819$ (RMSE = 4.70; MAE = 3.08). The differences between the leading models are small relative to the bootstrap confidence ranges discussed in Section 3.9, so the four models can be regarded as performing comparably. What separates the models is the spatial structure of their residuals (last two columns of Table 5): for the XGBoost, ensemble and Random Forest models Moran's I is mildly negative and statistically insignificant ($p > 0.17$), whereas SVR produces residuals with a positive Moran's I of 0.242 ($p = 0.001$), indicating that the SVR baseline still leaves spatial structure unexplained.

Table 5. Comparative test-set performance of the four machine-learning models for WQI prediction, including the Moran's I and p -value of the corresponding residual fields.

Model	R^2	RMSE	MAE	Moran's I	p -Value
Random Forest	0.859	4.16	3.24	−0.079	0.178
SVR	0.819	4.70	3.08	0.242	0.001
XGBoost	0.911	3.29	2.38	−0.067	0.275
Ensemble	0.910	3.27	2.36	−0.068	0.273

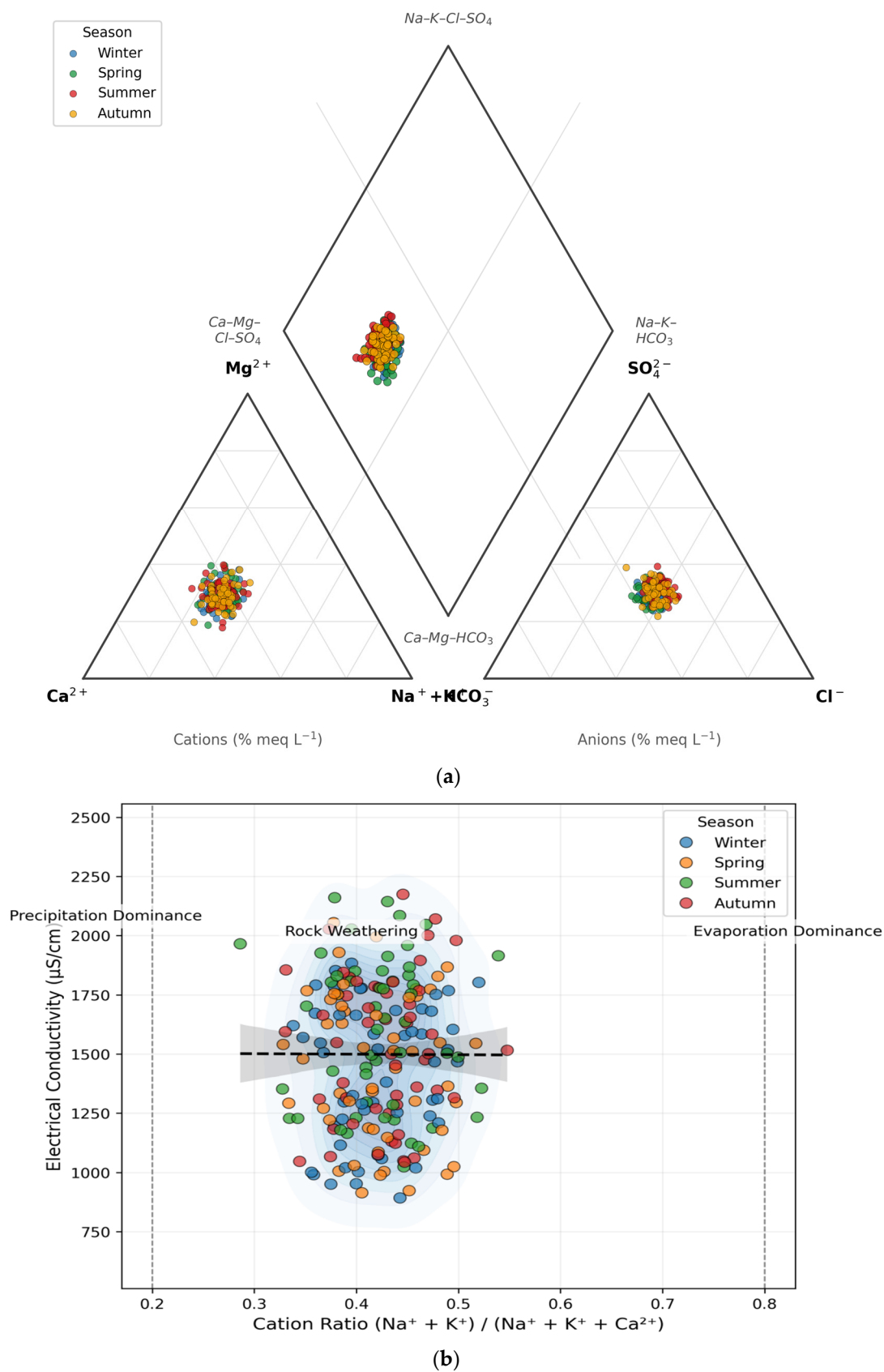


Figure 6. (a) Piper trilinear diagram of the 208 groundwater samples and (b) Gibbs diagram of EC versus $\text{Na}^+ / (\text{Na}^+ + \text{Ca}^{2+})$, illustrating the dominant hydrochemical processes governing groundwater evolution in the Bismil Plain.

3.6. Spatial Residual Analysis

The kriged spatial residual map for the ensemble baseline (Figure 7) shows that residuals are small in absolute magnitude and lack a coherent regional gradient, which is consistent with the low Moran's I reported in Table 5. Localized residual maxima are confined to isolated wells. For the SVR baseline, the spherical-variogram fit produced a finite range of autocorrelation (nugget = 14.78, partial sill = 4.79 WQI², range $\approx 0.48^\circ$, i.e., of the order of the extent of the monitoring network); the corresponding fit for the stacking ensemble returned a much smaller structured component (nugget = 1.55, partial sill = 2.17 WQI²), and the ordinary kriging step generated a non-trivial residual surface. For XGBoost and the stacking ensemble, the kriged residual field collapsed to near-zero values almost everywhere, and the hybrid ML + kriging prediction was numerically indistinguishable from the ML baseline. The spatial WQI prediction maps in Figure 8 are therefore presented for all four models for completeness, but they should be interpreted together with Table 5, which shows that only the SVR map carries a meaningful contribution from the geostatistical post-correction. Across all four models, the south-western corner and the central portion of the plain consistently emerge as higher-WQI (poorer-quality) zones, whereas the central-eastern wells show the lowest WQI values.

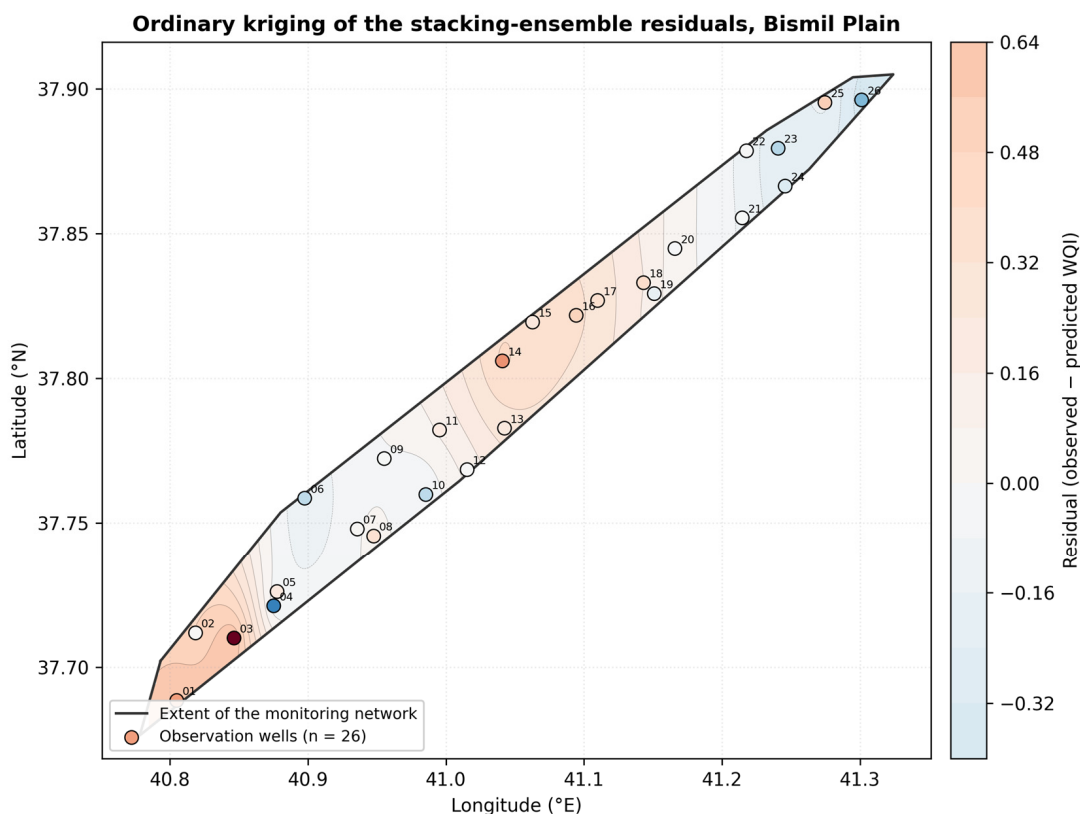


Figure 7. Spatial distribution of model residuals (observed—predicted WQI), kriged from the ensemble baseline.

3.7. SHAP-Based Driver Attribution

Figure 9 summarizes the global SHAP attribution from the final Random Forest model. EC dominates the contribution by a wide margin, followed by NO_3^- , SO_4^{2-} , Ca^{2+} , Mg^{2+} and Cl^- . The corresponding mean absolute SHAP values are 6.86 WQI units for EC, 1.57 for NO_3^- , 1.10 for SO_4^{2-} , 1.09 for Ca^{2+} , 0.85 for Mg^{2+} and 0.72 for Cl^- , so EC alone accounts for roughly four times the combined contribution of the five remaining predictors. The horizontal spread of EC is the largest in the beeswarm, indicating that EC alone moves the

predicted WQI across most of its observed range; the color gradient on this row confirms that higher EC values systematically push the model toward higher predicted WQI. NO_3^- emerges in second place even though its Pearson correlation with WQI is only 0.19, which is informative: the model exploits NO_3^- mainly through its non-linear, threshold-like effect at high concentrations rather than through a global linear association. SO_4^{2-} and Cl^- , which carry stronger linear correlations with WQI, modulate the salinity baseline and rank below NO_3^- once non-linearities are accounted for. The position of NO_3^- in the global ranking is therefore consistent with the field observation that 7/208 samples exceed the 50 mg/L WHO limit and 108/208 exceed the 25 mg/L guideline value: nitrate matters at the upper tail of its distribution rather than as a uniform contributor across the network.

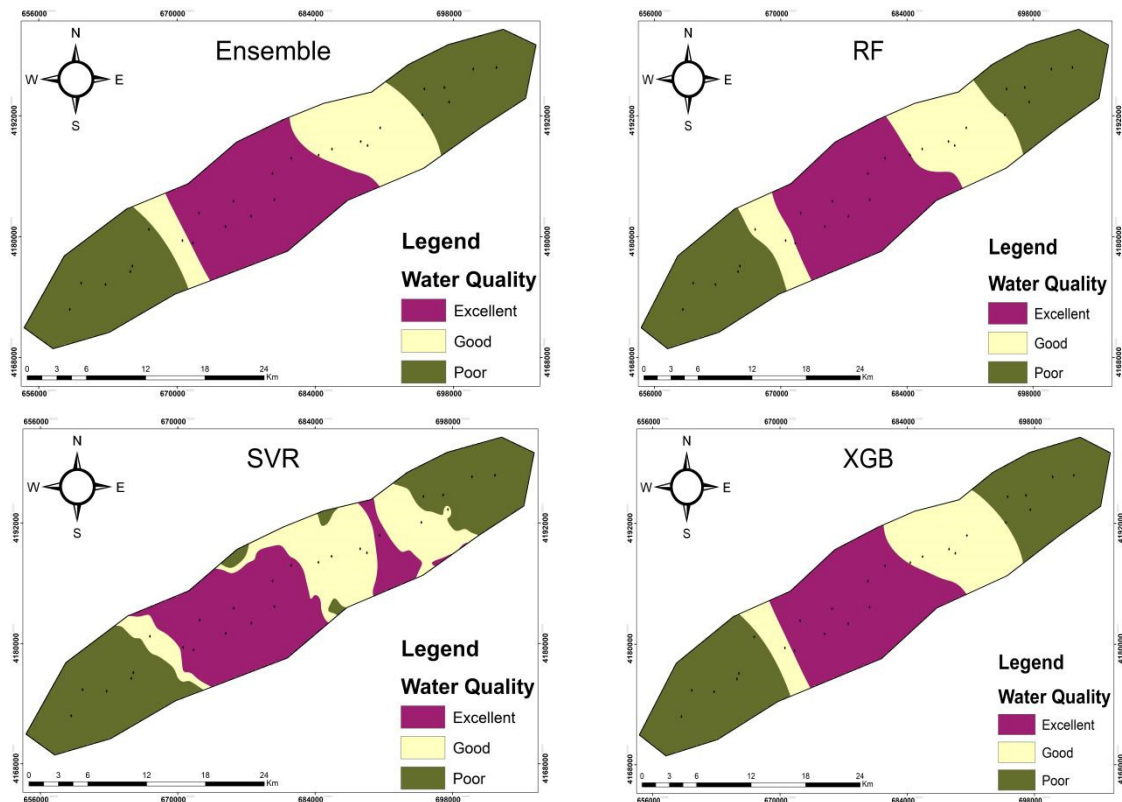


Figure 8. Spatial prediction of the Water Quality Index across the Bismil Plain using the Random Forest, XGBoost, SVR and stacking ensemble models, after the residual-kriging post-correction.

3.8. Anomalous Samples

Isolation Forest and Local Outlier Factor flagged complementary subsets of samples (Figure 10). Isolation Forest, with a fixed contamination parameter of 0.10, identified 21 samples as global anomalies, whereas the Local Outlier Factor, which uses a data-driven threshold, flagged a smaller set of four samples that are anomalous within their local density neighbourhood; the difference in set size follows directly from the two threshold conventions described in Section 2.3 and not from a disagreement about which samples are unusual. These observations are dominated by cases in which EC, Cl^- , SO_4^{2-} or NO_3^- simultaneously approach the upper deciles of their respective distributions, indicating that the intersection set captures hydrochemically meaningful anomalies rather than statistical artifacts. Several flagged samples correspond to summer periods at wells located in the central, intensively cultivated portion of the plain, which suggests that agronomic loading and seasonal evaporative concentration coincide in space and time.



Figure 9. SHAP beeswarm summary plot of the final Random Forest model. Each dot is one sample; horizontal position is the SHAP contribution to WQI, color encodes the feature value.

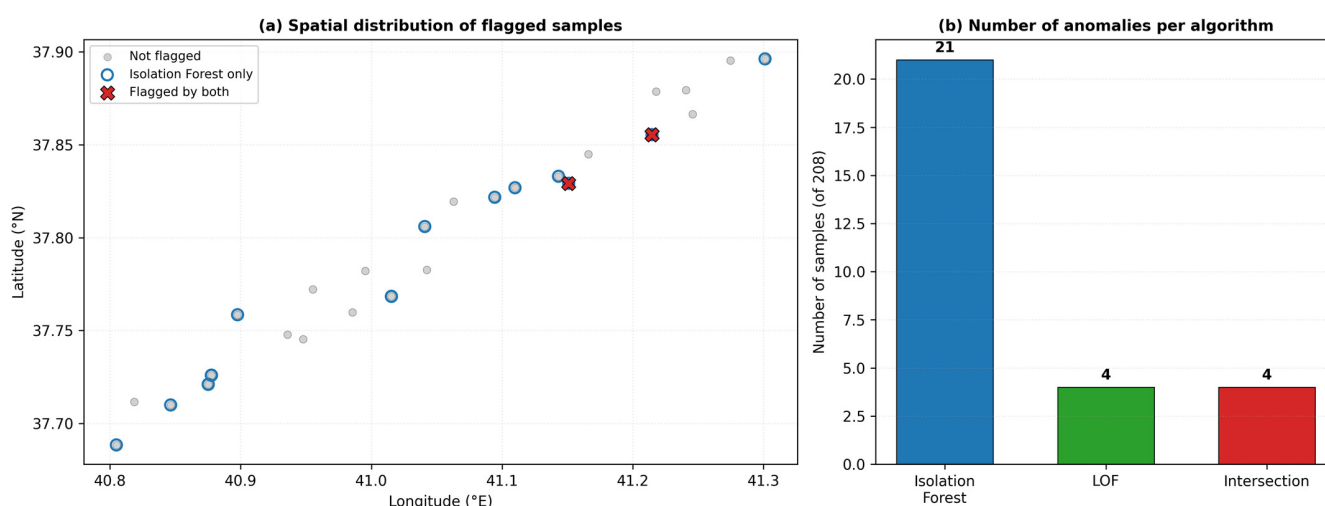


Figure 10. (a) Spatial distribution of groundwater samples flagged as anomalous by Isolation Forest (IF) and Local Outlier Factor (LOF) across the Bismil Plain. (b) Number of anomalous samples identified by each anomaly-detection algorithm and their intersection.

3.9. Predictive Uncertainty

The bootstrap-based 95% confidence intervals (Figure 11a) describe the stability of the predictions across the test set. The mean interval width is 6.97 WQI units, comparable to about twice the RMSE of the leading models (Figure 11b). These intervals quantify the uncertainty of the fitted model, that is, the variability of the prediction under resampling of the training set, and they are therefore interpreted here as a relative measure of sample-level stability rather than as prediction intervals for individual future observations. Wells with the widest intervals correspond to high- NO_3^- anomalies and to summer periods close to the upper EC range, that is, to the samples that are most strongly affected by the non-linearities discussed in Section 3.7. Uncertainty width therefore correlates with the locally critical anomalies rather than being uniform across the network, which is a useful

operational signal: the wells with the largest predictive uncertainty are also the wells that most warrant targeted re-sampling.

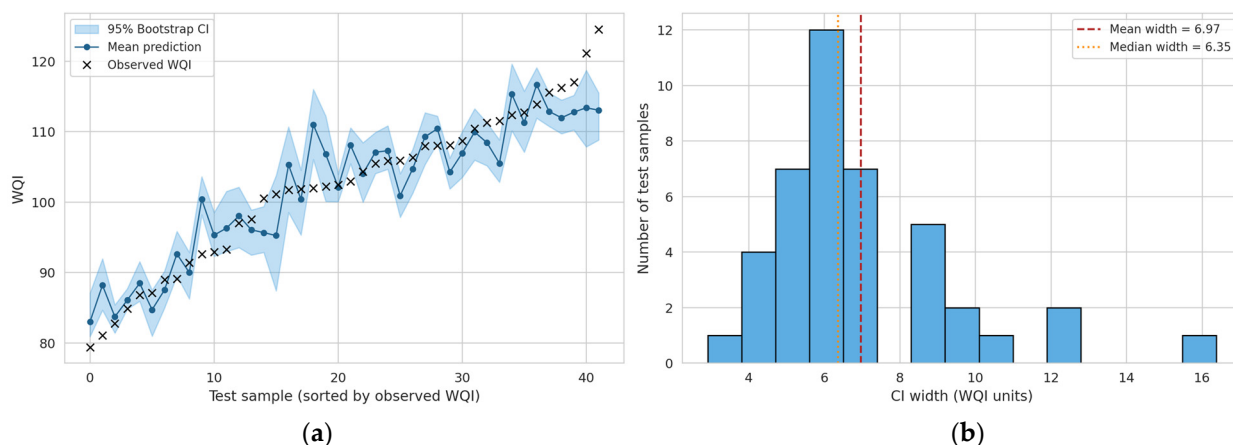


Figure 11. (a) Predictive 95% bootstrap intervals for the Random Forest model on the test set, sorted by observed WQI. (b) Distribution of CI widths across the test samples.

3.10. Integrated Hydrochemistry–Machine-Learning Interpretation

The classical and data-driven analyses converge on the same physical picture. The hydrochemical facies analysis identifies a mixed Ca^{2+} – Mg^{2+} – Cl^- – SO_4^{2-} system in which carbonate weathering, gypsum/anhydrite dissolution and seasonal evaporative concentration are jointly active, none of these three processes being dominant on its own. The SHAP analysis ranks EC as the dominant driver of WQI variability, followed by NO_3^- , SO_4^{2-} , Ca^{2+} , Mg^{2+} and Cl^- . The high SHAP rank of nitrate—second only to EC, despite its modest Pearson correlation with the WQI—shows the value of an explainable non-linear model: it recognises the threshold-like contribution of nitrate at the upper tail of its distribution, which a linear model would miss. Taken together, the hydrochemical and the SHAP results make the workflow operationally useful: they point to salinity as the target of network-level management and to nitrate hotspots as the target of local intervention.

4. Discussion

4.1. Hydrochemical Drivers and Processes

The Bismil aquifer occupies a moderately mineralized regime in which the geochemical fingerprint is jointly set by lithological controls and irrigation-driven processes. The dominance of Ca^{2+} and Mg^{2+} among cations and the near-equipartition of Cl^- , HCO_3^- and SO_4^{2-} among anions are diagnostic of a mixed carbonate–sulfate system, in which limestone and dolomite weathering provides the alkalinity baseline and gypsum/anhydrite dissolution raises the SO_4^{2-} concentration above the level that carbonate dissolution alone could produce. Gypsum- and anhydrite-bearing evaporites occur in the Oligocene and Palaeocene units of the basin [24,25]. Because the sampled wells are deep (mean about 200 m), they intersect this evaporite-bearing sequence directly. The mean molar $\text{Ca}^{2+}/\text{SO}_4^{2-}$ ratio of 1.36 further supports gypsum dissolution together with an additional carbonate source of calcium. The Cl^- signal is more difficult to attribute to a single source: the basin sequence does not host abundant chloride evaporite formations, so a meaningful fraction of the Cl^- is most plausibly anthropogenic, deriving from fertilizer salts, irrigation return flows and possibly domestic effluent reaching the unconfined aquifer through recharge.

The nitrate signal requires a separate interpretation. Agricultural activity influences groundwater chemistry in this intensively cultivated plain. The land is irrigated mainly

under wheat, maize and cotton, and nitrogen fertilisers are applied at high rates. Although the sampled wells are relatively deep, the permeable alluvial cover and widespread irrigation return flow allow surface-derived nitrate to reach the aquifer, and the highly variable nitrate concentrations point to this local, surface-controlled input. The absence of a strong correlation between NO_3^- and the other major ions does not contradict an anthropogenic origin; it reflects instead the different controls acting on these solutes. Nitrate has by far the highest coefficient of variation in the dataset (40.87%, against 12–21% for the remaining major ions), which indicates a spatially heterogeneous input governed by local agricultural practice rather than by a basin-wide geochemical process. Consistently with this, NO_3^- shows only a weak linear relationship with the WQI, yet SHAP identifies it as the second most influential variable, which means that it contributes mainly through localised, threshold-like effects rather than through a uniform gradient across the aquifer. Sulfate and chloride, by contrast, are controlled by more than one process at the same time. The Piper and Gibbs diagrams show that rock–water interaction dominates, with additional contributions from gypsum/anhydrite dissolution and from seasonal evaporative concentration, so that SO_4^{2-} is largely geogenic whereas Cl^- simultaneously reflects evaporation, irrigation return flow and anthropogenic inputs. These mixed origins explain why SO_4^{2-} and Cl^- do not necessarily co-vary with nitrate. We therefore interpret NO_3^- as a marker of localised agricultural contamination, and SO_4^{2-} and Cl^- as mixed geogenic–anthropogenic signals.

The seasonal signal, although moderate on average (mean WQI varying between 96.8 and 102.3 across the four seasons), is internally coherent. Summer increases in EC and Cl^- are consistent with evapotranspirative concentration and the redistribution of irrigation return flow during the dry months; winter values relax toward recharge-dominated conditions. This pattern repeats across both annual cycles, with an inter-annual difference in mean WQI of less than 0.3 units, which shows that seasonal rather than year-to-year variation is the dominant component of temporal variability during the two years covered by the monitoring programme. A single-snapshot survey would have missed the regularity of this signal and could have classified the network into the wrong WQI category by chance, depending on whether sampling fell in the wet or the dry season.

4.2. Why Explainable Machine Learning Matters

The four models converged on similar accuracy ranges (R^2 between 0.82 and 0.91), which suggests that the WQI signal is largely tractable from the input chemistry once feature selection is applied. What distinguishes the models in practice is not their order in the ranking but the fact that their accuracies are almost equivalent: the difference between XGBoost ($R^2 = 0.911$) and Random Forest ($R^2 = 0.859$) lies inside the bootstrap confidence envelope, and a different random split or feature-selection threshold could reverse it. Treating differences of this magnitude as substantive is a common pitfall in regional ML–WQI studies, and it has clear consequences when such models are used for decision support.

The SHAP analysis, by contrast, is informative regardless of leaderboard position. EC dominates the contribution to WQI by a wide margin, in line with both its Pearson correlation ($r = 0.86$) and its explicit weighting in the WQI formulation. The second rank of NO_3^- , despite its weak linear correlation with WQI ($r = 0.19$), is the most important diagnostic outcome of the analysis. An analysis based on linear correlation alone would have placed nitrate near the bottom of the ranking and would consequently have given it little weight in management decisions. The non-linear model shows instead that nitrate is decisive at the small number of wells where it exceeds the WHO limit of 50 mg/L, even though those wells are a minority within the network. Sulfate, calcium, magnesium and chloride follow, in an order that is broadly consistent with the group of salinity-related

variables identified by the correlation analysis. NO_3^- is therefore best understood as a sparse but high-impact driver, whereas EC, SO_4^{2-} and Cl^- are systemic drivers that shift the predicted index for the average sample. A comparable divergence between the SHAP ranking and the linear correlation ranking of the same dataset has been reported in recent explainable machine-learning studies of water quality [9,10,12], which supports the view that driver attribution should not be inferred from correlation coefficients alone.

4.3. When Does Residual Kriging Genuinely Help?

Hybrid ML–geostatistics frameworks are commonly applied without checking whether the residual field actually carries spatial structure. Our workflow inverts that assumption: Moran’s I is computed before the kriging step, and the geostatistical post-correction is interpreted in light of the test result. For the leading models XGBoost (Moran’s $I = -0.067$, $p = 0.275$) and the stacking ensemble (-0.068 , $p = 0.273$) the residual field is indistinguishable from spatial noise, so the kriging step is essentially a null operation. For SVR (Moran’s $I = 0.242$, $p = 0.001$), the residuals retain a measurable autocorrelation, and the kriging step produces a real but small correction. This behaviour is exactly what regression-kriging theory predicts: the geostatistical term contributes only to the extent that the residual random field is autocorrelated [31], so a model that has already absorbed the spatial structure leaves nothing for the kriging step to recover. Yang et al. [11] reached a related conclusion from the opposite direction, reporting that maps produced by models of comparable accuracy may nevertheless differ appreciably in space, which argues for treating the spatial component as an object of validation in its own right. Residual kriging is therefore most useful when (i) the ML model has under-fit a spatial component because of limited geographic features, (ii) the sampling network is dense enough to support variogram estimation, and (iii) the residual field actually exhibits autocorrelation; in our setting, only the first condition is fully satisfied, and the third is satisfied for some models but not for the leading ones. We therefore recommend that future small-sample groundwater studies treat residual kriging as a diagnostic rather than a default enhancement, and report Moran’s I alongside any hybrid ML–geostatistics result.

4.4. Comparison with Previous Regional Work and Broader Implications

Çankaya et al. [6] characterized the surface-water network of the same Bismil Plain and reported salinity stress with elevated SO_4^{2-} and Cl^- , accompanied by health-risk concerns. The groundwater results presented here mirror that pressure pattern but at lower amplitude and with reduced variability. Half of the groundwater samples already classify as Poor, and summer EC reaches $2174 \mu\text{S}/\text{cm}$ in some wells. This indicates that the agricultural plain places pressure on both the surface and the subsurface, with the groundwater responding on a longer timescale. Şener [7] reported substantially more variable hydrogeochemistry across the broader Diyarbakır province with thermal-water and rare-earth-element signatures from deep formations; the present dataset, drawn from a shallow agricultural aquifer, does not show those signals, consistent with a separate hydrogeochemical setting. Beyond the region, the accuracy range obtained here ($R^2 = 0.82\text{--}0.91$) is comparable with that of recent machine-learning assessments of groundwater quality in other semi-arid agricultural basins, such as the XGBoost-based irrigation-quality models of Hussein et al. [8] in Algeria and the national-scale neural-network assessment of Ahn et al. [12], while the dominance of salinity-related predictors in our SHAP ranking parallels the results of Makumbura et al. [9] and Choudhary et al. [10]. What differs is that in the Bismil Plain nitrate enters the ranking through a threshold effect rather than through a monotonic gradient, which we attribute to the localised character of fertiliser application in the plain. The combined picture supports a management framing in which Bismil groundwater is not yet beyond

recovery but is sensitive enough that adverse trajectories could materialize within a few seasons of intensified loading.

4.5. Limitations and Future Directions

Our conclusions should be read in the light of four limitations. First, 26 wells over two years is modest for variogram-based geostatistics, and a denser network would allow stricter spatial validation; in particular, a doubled network would allow a leave-one-block-out validation with finer blocks and a more reliable variogram fit. Second, our parameter list does not include trace metals, NO_2^- , NH_4^+ , dissolved organic carbon or stable isotopes ($\delta^{15}\text{N}$, $\delta^{18}\text{O}$), all of which would help disentangle fertilizer, manure and natural-N sources for the NO_3^- signal. Third, while the SHAP-guided feature reduction improves model stability, removing input variables that have explicit weight in the WQI (such as pH and HCO_3^-) means that the predictive model approximates a slightly compressed version of the index; this is appropriate for driver attribution but would be revisited in a true forecasting application. Fourth, the ionic charge balance shows a small systematic anion excess (mean CBE = -3.74%). This most probably reflects the absence of the minor species (CO_3^{2-} , NO_2^- and PO_4^{3-}) from the analytical programme, and these will be determined in the next campaign. The offset is relevant to the precision of the facies assignments rather than to the WQI itself, which is computed directly from the measured concentrations. Future work should extend the temporal record to capture multi-year climatic variability, integrate isotope tracers and trace elements, and complement the bootstrap uncertainty with a fully spatially-resolved Bayesian ensemble simulation.

5. Conclusions

Groundwater quality in the Bismil Plain was assessed by combining classical hydro-chemical interpretation with explainable machine learning, residual geostatistics, anomaly screening and bootstrap-based uncertainty. The 208 samples collected across 26 wells and eight seasonal periods between 2022 and 2024 occupy a narrow WQI band (79.37–125.48), placing every sample within either the Good or the Poor class and showing that the aquifer lies close to the boundary between these two classes rather than in a severely degraded state. The dominant systemic drivers of WQI variability are EC, SO_4^{2-} and Cl^- , with NO_3^- acting as a high-impact but locally concentrated control: 7/208 samples exceed the WHO 50 mg/L limit and 108/208 exceed the more conservative national and EU guideline value of 25 mg/L, so nitrate management requires a hotspot-targeted strategy rather than a network-wide one. The combined ranking returned by SHAPEC $> \text{NO}_3^- > \text{SO}_4^{2-} > \text{Ca}^{2+} > \text{Mg}^{2+} > \text{Cl}^-$ captures both the salinity-driven systemic signal and the threshold-like contribution of nitrate, which a linear correlation analysis alone would have missed.

Methodologically, the paper integrates explainable ML, residual kriging, anomaly screening and bootstrap uncertainty in a single transparent pipeline calibrated for a small-sample, multi-season groundwater dataset. A key methodological message is that residual kriging contributes meaningfully only when the underlying ML model leaves spatial structure unexplained; for our XGBoost and stacking ensemble baselines the residual field was effectively unstructured, and the hybrid step produced a near-null correction. We report this directly rather than as a uniform spatial improvement, and recommend Moran's I as a routine pre-check for any future ML-plus-kriging study.

For groundwater management in the Bismil Plain, three operational recommendations follow. First, summer monitoring of EC, Cl^- and SO_4^{2-} should be prioritized in wells already approaching the WQI threshold of 100, since seasonal evaporative concentration is sufficient to push borderline samples into the Poor class. Second, NO_3^- hotspots identified through the joint Isolation Forest–LOF screening should be investigated on the ground

rather than addressed through network-wide fertilizer policy. Third, isotopic and trace-element analyses should be added to the existing parameter set in the next monitoring cycle, with the explicit aim of separating natural and anthropogenic loading. More broadly, the workflow described here is portable to other small-sample agricultural aquifers where decisions must be made with limited but well-structured data, and where transparent driver attribution is more useful than leaderboard accuracy.

Author Contributions: Conceptualization, S.Ö.G. and G.Y.; Methodology, S.Ö.G. and A.V.B.; Software, S.Ö.G. and A.V.B.; Validation, S.Ö.G.; Formal Analysis, S.Ö.G. and S.B.R.; Investigation, S.Ö.G.; Resources, S.Ö.G.; Data Curation, S.Ö.G., S.B.R. and A.V.B.; Writing—Original Draft, S.Ö.G., A.V.B. and G.Y.; Writing—Review and Editing, S.Ö.G., A.V.B. and G.Y.; Visualization, S.Ö.G.; Supervision, S.Ö.G. and G.Y.; Project Administration, S.Ö.G.; Funding Acquisition, S.Ö.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: This study constitutes a part of the first author's dissertation. The authors would like to thank all authors for their continuous help in the field and laboratory studies.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Yesilnacar, M.I.; Sahinkaya, E.; Naz, M.; Ozkaya, B. Neural network prediction of nitrate in groundwater of Harran Plain, Turkey. *Environ. Geol.* **2008**, *56*, 19–25. [\[CrossRef\]](#)
2. Horton, R.K. An index number system for rating water quality. *J. Water Pollut. Control Fed.* **1965**, *37*, 300–306.
3. Brown, R.M.; McClelland, N.I.; Deininger, R.A.; Tozer, R.G. A water quality index—Do we dare? *Water Sew. Works* **1970**, *117*, 339–343.
4. Wagh, V.M.; Panaskar, D.B.; Muley, A.A.; Mukate, S.V.; Lolage, Y.P.; Aamalawar, M.L. Prediction of groundwater suitability for irrigation using artificial neural network model: A case study of Nanded tehsil, Maharashtra, India. *Model. Earth Syst. Environ.* **2016**, *3*, 1013–1023. [\[CrossRef\]](#)
5. Tiwari, A.K.; Singh, A.K.; Mahato, M.K. Assessment of groundwater quality of Pratapgarh district in India for suitability of drinking purpose using water quality index (WQI) and GIS technique. *Sustain. Water Resour. Manag.* **2018**, *4*, 601–616. [\[CrossRef\]](#)
6. Çankaya, Ş.; Varol, M.; Bekleyen, A. Hydrochemistry, water quality and health risk assessment of streams in the Bismil Plain, an important agricultural area in southeast Türkiye. *Environ. Pollut.* **2023**, *331*, 121874. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Şener, M.F. Hydrogeochemical and rare earth element properties of geothermal waters and the use of abandoned oil wells in Diyarbakır. *Environ. Earth Sci.* **2024**, *83*, 474. [\[CrossRef\]](#)
8. Hussein, E.E.; Derdour, A.; Zerouali, B.; Almaliki, A.; Wong, Y.J.; Ballesta-de los Santos, M.; Ngoc, P.M.; Hashim, M.A.; Elbeltagi, A. Groundwater quality assessment and irrigation water quality index prediction using machine learning algorithms. *Water* **2024**, *16*, 264. [\[CrossRef\]](#)
9. Makumbura, R.K.; Mampitiya, L.; Rathnayake, N.; Meddage, D.P.P.; Henna, S.; Dang, T.L.; Hoshino, Y.; Rathnayake, U. Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like Shapley additive explanations (SHAP) for interpreting the black-box nature. *Results Eng.* **2024**, *23*, 102831. [\[CrossRef\]](#)
10. Choudhary, R.; Kumar, A.C.P.; Naik, M.M.; Choudhury, M.; Khan, N.A. Predicting water quality index using stacked ensemble regression and SHAP based explainable artificial intelligence. *Sci. Rep.* **2025**, *15*, 31139. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Yang, S.; Luo, D.; Tan, J.; Li, S.; Song, X.; Xiong, R.; Wang, J.; Ma, C.; Xiong, H. Spatial mapping and prediction of groundwater quality using ensemble learning models and SHapley Additive exPlanations with spatial uncertainty analysis. *Water* **2024**, *16*, 2375. [\[CrossRef\]](#)
12. Ahn, S.H.; Jeong, D.H.; Kim, M.; Lee, T.K.; Kim, H.-K. Prediction of groundwater quality index to assess suitability for drinking purpose using averaged neural network and geospatial analysis. *Ecotoxicol. Environ. Saf.* **2023**, *265*, 115485. [\[CrossRef\]](#) [\[PubMed\]](#)

13. Singha, S.; Pasupuleti, S.; Singha, S.S.; Singh, R.; Kumar, S. Prediction of groundwater quality using efficient machine learning techniques. *Chemosphere* **2021**, *276*, 130265. [[CrossRef](#)] [[PubMed](#)]
14. El Bilali, A.; Taleb, A.; Brouziyne, Y. Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agric. Water Manag.* **2021**, *245*, 106625. [[CrossRef](#)]
15. Moran, P.A.P. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23. [[CrossRef](#)]
16. Piper, A.M. A graphic procedure in the geochemical interpretation of water-analyses. *Eos Trans. Am. Geophys. Union* **1944**, *25*, 914–928. [[CrossRef](#)]
17. Gibbs, R.J. Mechanisms controlling world water chemistry. *Science* **1970**, *170*, 1088–1090. [[CrossRef](#)] [[PubMed](#)]
18. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
19. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
20. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [[CrossRef](#)]
21. Drucker, H.; Burges, C.J.C.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. In Proceedings of the Advances in Neural Information Processing Systems 9, Denver, CO, USA, 3–5 December 1996; pp. 155–161.
22. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422. [[CrossRef](#)]
23. Breunig, M.M.; Kriegel, H.-P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 93–104. [[CrossRef](#)]
24. Güngör-Yeşilova, P.; Helvacı, C. Petrographic study and geochemical investigation of the evaporites associated with the Germik Formation (Siirt Basin, Turkey). *Carbonates Evaporites* **2017**, *32*, 177–194. [[CrossRef](#)]
25. Yeşilova, Ç.; Helvacı, C. Diagenesis and palaeogeographic development of the Oligocene evaporites of the Germik Formation (Kurtalan, SW Siirt), Turkey. *Yerbilimleri* **2013**, *34*, 141–168.
26. APHA. *Standard Methods for the Examination of Water and Wastewater*, 20th ed.; American Public Health Association: Washington, DC, USA, 1998.
27. Freeze, R.A.; Cherry, J.A. *Groundwater*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1979.
28. Appelo, C.A.J.; Postma, D. *Geochemistry, Groundwater and Pollution*, 2nd ed.; A.A. Balkema: Leiden, The Netherlands, 2004.
29. World Health Organization. *Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First and Second Addenda*; WHO: Geneva, Switzerland, 2022.
30. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. [[CrossRef](#)]
31. Hengl, T.; Heuvelink, G.B.M.; Rossiter, D.G. About regression-kriging: From equations to case studies. *Comput. Geosci.* **2007**, *33*, 1301–1315. [[CrossRef](#)]

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