Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning and GIS-Integrated Framework for Irrigation Water Quality Index (IWQI) Mapping and Hydrochemical Clustering in an Algerian Semi-Arid Watershed

Zahra Benkhamallah
Sidi Bel Abbes University
Bachir Hallouche
Sidi Bel Abbes University
Khaled Korichi
kh.korichi@gmail.com

Sidi Bel Abbes University

Research Article

Keywords: IWQI assessment, Hydrochemical Clustering, Machine Learning, GIS, Wadi El Melah

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11330101/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine Learning and GIS-Integrated Framework for Irrigation Water Quality Index (IWQI) Mapping and Hydrochemical Clustering in an Algerian Semi-Arid Watershed

Dr. Zahra Benkhamallah
Valorization of Phytoresources and Eco-Development of Spaces Laboratory, Sidi Bel Abbes University, Sidi Bel Abbes 22000, Algeria
Email: zbenkhamallah@gmail.com
ORICD: 0009-0000-4481-9825

Dr. Bachir Hallouche
Civil Engineering and Environmental Laboratory (LGCE), Sidi Bel Abbes University, Sidi Bel Abbes 22000, Algeria
Email: hallouche_udl@yahoo.fr
ORCID: 0000-0003-3129-3419

Prof. Khaled Korichi
Civil Engineering and Environmental Laboratory (LGCE), Sidi Bel Abbes University, Sidi Bel Abbes 22000, Algeria
Email: kh.korichi@gmail.com (*Corresponding author)
ORCID: 0000-0002-2663-2757

Abstract

In the Ain Temouchent region, surface water is severely limited, leaving groundwater as the primary source for both domestic supply and agricultural irrigation. This study evaluates the suitability of El Melah watershed groundwater for irrigation by computing the Irrigation Water Quality Index (IWQI) following the Meireles approach, mapping its spatial distribution through GIS, and complementing the analysis with two machine learning methods: (i) a Random Forest (RF) model with Leave-One-Out Cross-Validation (LOOCV) and SHAP explainability analysis to identify the dominant drivers of IWQI degradation; and (ii) K-means and hierarchical clustering to delineate hydrochemical facies and provide a statistically grounded alternative to the conventional Wilcox diagram. Salinity is elevated throughout the study area (EC: 1,034–5,010 µS cm^{-1}). The RF model achieved LOOCV $R^2 = 0.61$, with Na^+ and EC confirmed as the dominant IWQI degradation drivers by SHAP analysis. Hierarchical clustering resolved three hydrochemical facies: a Ca^{2+}-HCO_3^- continental group, a Na^+-Cl^- evaporite-influenced group, and a seawater-intrusion group, all confirmed as statistically distinct by one-way ANOVA ($p < 0.05$). Overall, groundwater quality in the watershed ranges from moderate to severe restriction for irrigation use, with the northern sector most critically affected by marine infiltration.

Keywords: IWQI assessment, Hydrochemical Clustering, Machine Learning, GIS, Wadi El Melah.

1. Introduction

In arid and semi-arid regions, groundwater is typically the primary freshwater source for drinking, industry, and farming (Toumi et al. 2015). Sustained extraction, compounded by climate variability, has progressively degraded both the quantity and quality of these reserves. Meeting food security targets requires more than technology, it depends on good soil quality, water fit for irrigation, and sound agronomic practices working together.

Evaluating water availability for irrigation requires equal attention to both quantity and quality, yet in water-stressed countries quality has historically received far less scrutiny. Water quality is best understood as the degree to which water meets the requirements of its intended use, a definition that necessarily depends on the specific application (Ayers and Westcot 1985).

Groundwater salinization is a widespread problem in coastal zones. It can be primary from natural processes like rock mineralogy or seawater intrusion or secondary, caused by human activity (Ghassemi et al. 1995). Secondary salinization most often results from irrigation systems poorly suited to local conditions (Badraoui et al. 1998). According to the World Resources Institute, more than half of irrigated land worldwide has been affected by salinization and/or sodification due to poor irrigation methods, inadequate drainage, and the use of saline water sources (Van Der Lelij and Poolman 1989; Dehaan and Taylor 2002). As a result, salinization degrades millions of hectares of productive farmland every year (Nishanthiny et al. 2010; Mohammed Muthanna 2011).

Evaluating irrigation water quality requires considering several physicochemical parameters sodium, macronutrients, micronutrients, alkalinity, acidity, and hardness which at high concentrations harm both soil structure and plant growth. The Sodium Adsorption Ratio (SAR) quantifies the sodicity hazard of irrigation water, measuring the propensity of sodium to drive cation exchange reactions in the soil matrix (Richards 1968). Soil sodicity causes clay particle dispersion, reduces porosity, and blocks aeration and water infiltration, all critical for normal plant growth (Gupta and Abrol 1990; Ghassemi et al. 1995).

Composite water quality indices such as the IWQI (Meireles et al. 2010) integrate multiple physicochemical parameters into a single dimensionless score, enabling straightforward spatial mapping and informed management decisions. The IWQI has been applied successfully in semi-arid regions across Brazil, North Africa (Hadji et al. 2013), and the Middle East (Kizar 2018). Classical index-based approaches do have two notable limitations, however: they cannot identify which specific parameters most drive quality degradation at individual boreholes, nor can they reveal the underlying geochemical processes responsible for spatial patterns in water quality.

Machine learning methods have recently been applied to groundwater quality assessment to overcome these limitations. Random Forest (RF) models have demonstrated strong predictive performance for water quality indices in data-sparse environments (Knoll et al. 2019; Rahmati et al. 2019). Integrating SHAP (SHapley Additive exPlanations) values with RF models, grounded in cooperative game theory (Lundberg and Lee 2017), allows each prediction to be decomposed into individual parameter contributions, providing a mechanistic link between measured chemistry and computed index values that classical methods cannot offer.

Unsupervised hydrochemical clustering has similarly advanced beyond classical Piper and Wilcox diagrams. K-means and hierarchical clustering applied to the full ionic dataset can identify hydrochemical facies, groups of boreholes sharing a common geochemical origin, without imposing pre-defined classification boundaries (Helena et al. 2000; Cloutier et al. 2008; Ismail et al. 2021). Combined with spatial analysis, these facies can be directly linked to geological and hydrogeological processes such as water–rock interaction, evaporite dissolution, and seawater intrusion (Balathandayutham et al. 2015; Marghade et al. 2019).

Single-campaign groundwater quality studies inevitably carry temporal uncertainty. In semi-arid Mediterranean catchments, dry-season evaporative concentration has been shown to raise EC by factors of 1.3–2.1 relative to spring values (Azlaoui et al. 2021; Benmerabet et al. 2025), with the potential to shift boreholes from moderate to severe restriction classes. Accounting for this variability is therefore an important consideration in any rigorous water quality assessment.

This study applies an integrated framework to the Wadi El Melah watershed (Ain Temouchent, northwestern Algeria), a coastal semi-arid catchment where seawater intrusion, evaporite dissolution, and agricultural water demand converge to create complex groundwater quality challenges. The specific objectives are: (1) to compute and map IWQI following Meireles et al. (2010); (2) to predict IWQI using a Random Forest model with LOOCV and identify dominant degradation drivers through SHAP analysis; and (3) to delineate hydrochemical facies through K-means and hierarchical clustering.

2. Study Area

The Wadi El Melah catchment pertains to the Oranais coastal basin complex and is situated in the northwestern sector of Algeria, between latitudes 35°07’43"N - 35°31’8"N and longitudes 0°52’40"W - 1°10’45"W (Figure 1). The watershed encompasses a total area of 870 km² with a Gravelius compactness coefficient KC = 1.31 and drainage density Dd = 1.48 km/km². Elevations range from 0 to 824 m above sea level. The prevailing climate is semi-arid Mediterranean, with mean inter-annual rainfall of 362.9 mm and mean annual temperature of 17.8°C (range: 10.5°C in January to 25.6°C in August).

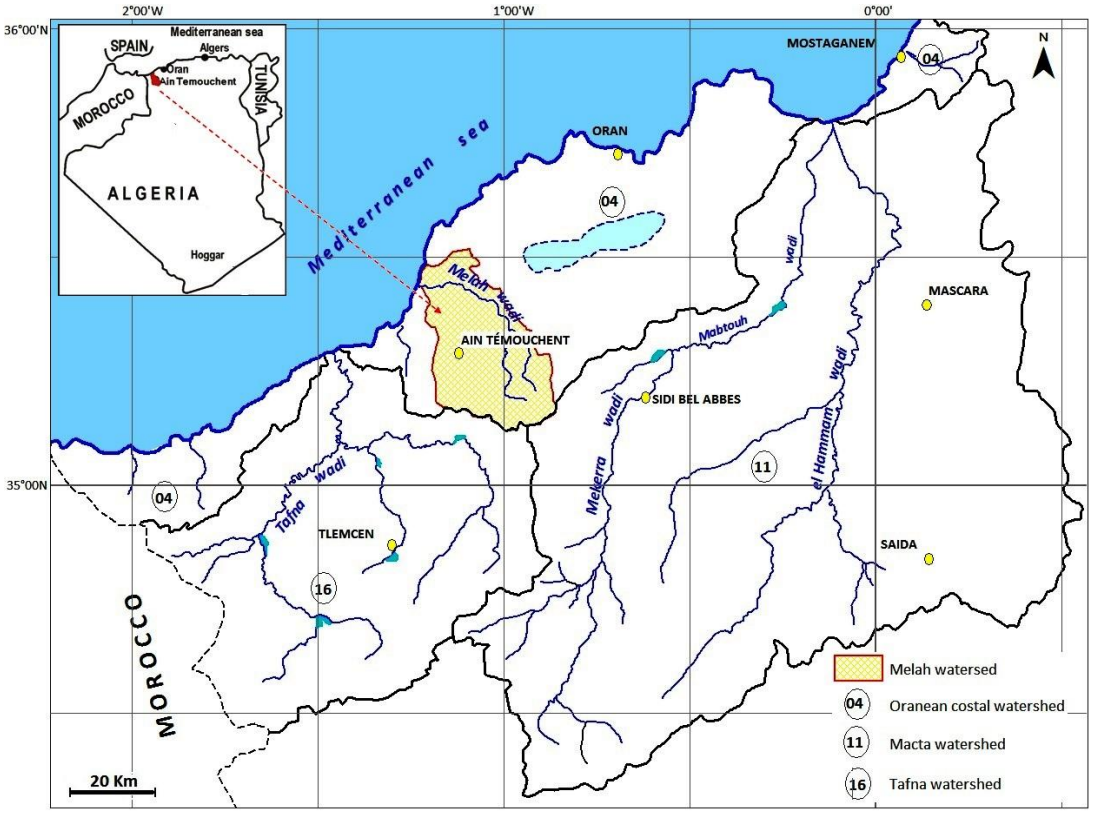


Figure 1. Geographic location of the Wadi El Melah watershed

The geology of the study area is characterized by the occurrence of sedimentary formations spanning from the Upper Jurassic to the Quaternary, alongside volcanic sequences (Figure 2). The Upper Jurassic is represented at Djebel Necissa by micaceous pelites and ferruginous fine-grained sandstones overlain by bioclastic limestones attributed to the Upper Oxfordian, and at Djebel Tounit by massive Kimmeridgian limestones resting upon reddish micaceous pelites. Cretaceous outcrops occur in Djebel Sidi Kacem and Djebel Menjel and are composed of marly limestones, carbonates, and sandstones (Guardia 1975). Eocene deposits, consisting of sandstones and marls, are exposed in the southern portion of the study area. These formations are overlain by transgressive Miocene to Quaternary sequences comprising ochre to brick-red alluvial deposits, surmounted by marine facies consisting of blue-grey marls and oyster-bearing marls, which are in turn capped by Djebel Keroulis limestones. To the east of Ain Temouchent, the marly series are overlain by algal limestones (Lithothamniaceae), scleractinian corals, and echinoids deposited during the Messinian transgression (Fenet 1975). During the Plio-Quaternary, an episode of intense volcanic activity generated alkaline basaltic lava flows and volcanic tuffs that extensively covered large areas of the Ain Temouchent region (Megartsi 1985; Coulon et al. 2002; Louni-Hacini 2002).

Three principal aquifer systems characterise the study area from a hydrogeological standpoint: (1) the Plio-Quaternary aquifer, composed of sandy silty deposits, Quaternary conglomerates, Helix-bearing sands, sandstones, and Belemnite-bearing Pliocene sands resting on a Miocene marly substratum, with a permeability of approximately 4×10^{-4} m/s; (2) the basalt aquifer, whose substratum may consist of Pliocene sands or Miocene limestones and marls, represented by basaltic tuffs of variable particle size and extensively fissured massive basaltic flows; and (3) the Messinian limestone aquifer, comprising friable whitish limestones that locally transition to chalky limestones or gypsum, with a hydraulic conductivity of approximately 2×10^{-3} m/s.

3. Materials and Methods

3.1 Sampling and Analytical Procedures

Water samples are collected from 21 boreholes tapping the various aquifer formations of the watershed (Figure 2) during a piezometric survey conducted between April and May 2018. All samples are collected after a minimum pumping period of 30 minutes and stored in polyethylene bottles at 4°C until analysis. Field measurements (pH, EC, TDS, temperature) are performed in situ using a HACH Sens Ion 156 multi-parameter instrument. Major ions (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Cl^{-}, SO_4^{2-}, HCO_3^{-}) and nitrogenous compounds are analysed at the Algerian Water Authority (ADE) laboratory in Ain Temouchent. Ionic balance errors ranged from −4.85 to +4.76%, within the acceptable ±5% threshold (Freeze and Cherry 1979; Appelo and Postma 2004).

3.2 Classical Irrigation Quality Indices (EC, SAR, Wilcox)

Salinity is characterised by electrical conductivity (EC, µS cm^{-1}). The Sodium Adsorption Ratio (SAR) is computed following Richards (1954) as:

$$\mathrm{SAR} = \frac{\mathrm{Na}^{+}}{\sqrt{(\mathrm{Ca}^{2+}\ \mathrm{Mg}^{2+})/2}} \quad (3)$$

All concentrations are expressed in meq L^{-1}. Samples are projected on the Wilcox diagram for graphical classification (Todd and Mays 2004). These classical indices provided the baseline characterisation against which the machine learning results are subsequently compared.

3.3 Irrigation Water Quality Index (IWQI)

Following Meireles et al. (2010), the IWQI of is computed as:

$$IWQI = \sum_{i=1}^{n} q_i \times w_i \quad (1)$$

where q_i is the quality sub-index of parameter i (ranging 0–100) and w_i is its normalized weight. The five parameters and their corresponding weights are (Table 1): EC (0.211), Na^{+} (0.204), HCO_3^{-} (0.202), Cl^{-} (0.194), and SAR (0.189). All classical indices and IWQI scores are spatially interpolated and mapped using ArcGIS.

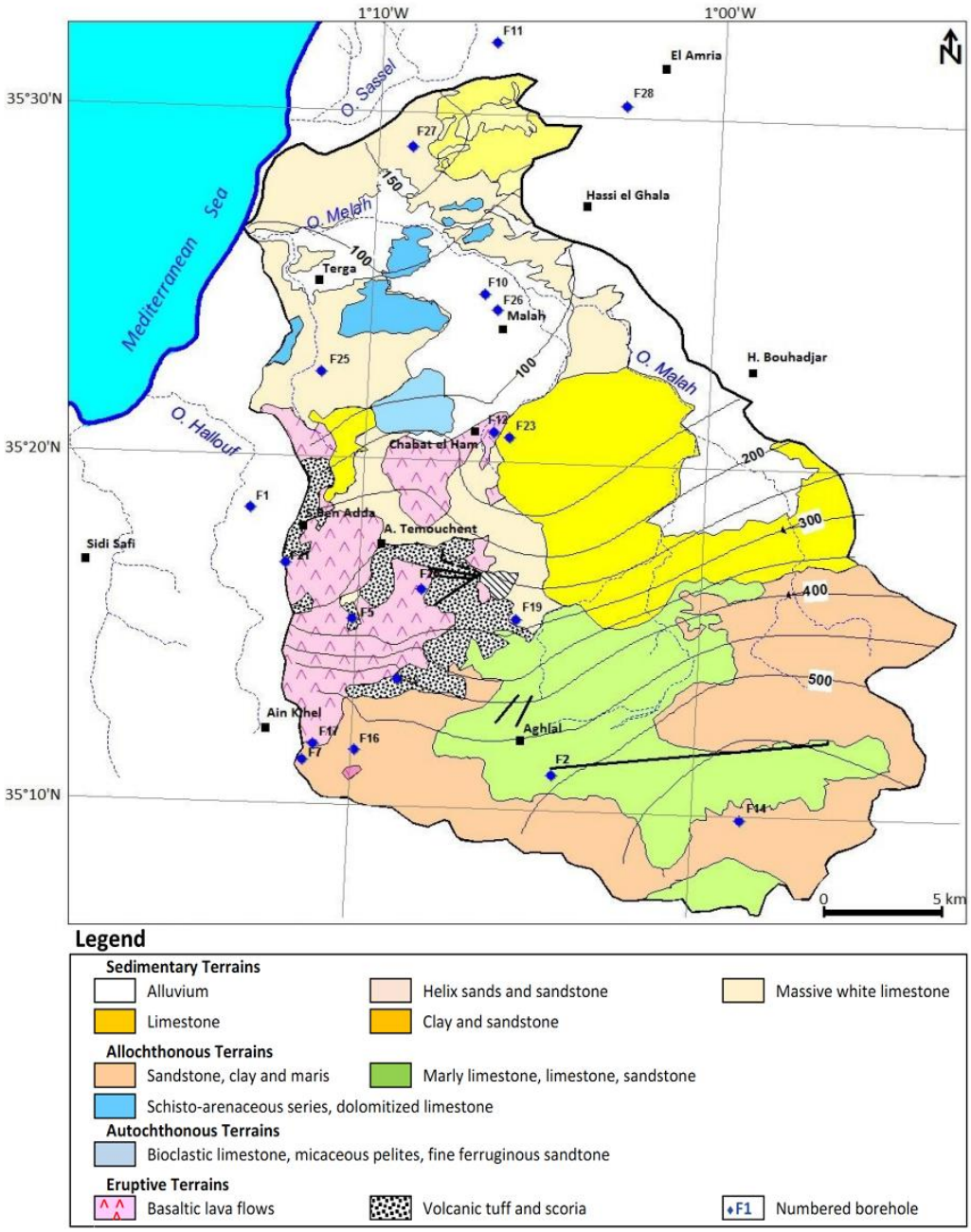


Figure 2. Hydrogeological map and sampling point locations

Table 1: Threshold values of water quality measurement parameters (Meireles et al. 2010)

qi	EC (µS/cm)	SAR $(meq.L^{-1})^{0.5}$	Na^+ $(meq.L^{-1})$	Cl^- $(meq.L^{-1})$	HCO_3^- $(meq.L^{-1})$
85 -100	200 - 750	2 - 3	2 - 3	1 - 4	1 - 1.5
60 - 85	750 - 1500	3 - 6	3 - 6	4 - 7	1.5 - 4.5
35 - 60	1500 - 3000	6 - 12	6 - 9	7 - 10	4.5 - 8.5
0 - 35	< 200 or ≥ 3000	< 2 or ≥ 12	< 2 or ≥ 9	< 1 or ≥ 10	< 1 or ≥ 8.5

The q_i values are computed using Equation 2, based on the tolerance limits presented in Table 1 and the results of the physicochemical analysis of water samples:

$$q_i = q_{i\,max} - \left[\frac{(X_{ij} - X_{inf})}{X_{amp}} \text{ x } q_{i\,amp}\right] \quad (2)$$

where $q_i(max)$ is the maximum value of the class to which q_i belongs; X_{ij} is the observed value of the parameter; $X_{in}f$ is the lower boundary of the class to which the parameter belongs; $q_i(amp)$ is the amplitude of the quality class; and X_{amp} is the amplitude of the class to which the parameter belongs. For the determination of X_{amp} in the terminal class of each parameter, the observed maximum value from the physicochemical analysis is adopted as the upper boundary.

The classification of irrigation water quality into distinct restriction categories is based on the IWQI thresholds, which are delineated considering the risks of salinization, soil permeability reduction, and phytotoxicity. According to the classification proposed by Meireles et al. (2010), five categories are defined. Waters with IWQI values between 85 and 100 are classified as having no restriction (NR) and are suitable for most soils with minimal risk, posing no phytotoxicity to crops. Values between 70 and 85 correspond to low restriction (LR), where use is generally acceptable for light to moderately permeable soils, although salt leaching is recommended and salt-sensitive crops should be avoided. IWQI values ranging from 55 to 70 indicate moderate restriction (MR), suitable for soils with moderate to high hydraulic conductivity and for moderately salt-tolerant crops. High restriction (HR), defined by values between 40 and 55, limits use to highly permeable soils and requires careful irrigation management, particularly under high salinity conditions, with cultivation restricted to salt-tolerant crops. Finally, waters with IWQI values below 40 fall under severe restriction (SR), where irrigation is generally discouraged under normal conditions and is only feasible in exceptional cases with specific soil management practices and the use of highly salt-tolerant or halophytic species.

Following this, cartographic representation of groundwater quality is produced using MapInfo GIS software. The Vertical Mapper spatial analysis extension is used to perform spatial integration and thematic mapping of individual quality parameters and the composite IWQI across the El Melah watershed, as formulated by Equation 1 through the overlay of thematic maps of the contributing physicochemical parameters.

3.4 RF-LOOCV-SHAP Framework

A Random Forest (RF) regression model is trained to predict IWQI from the five input parameters (EC, SAR, Na^+, Cl^-, HCO_3^-). RF is an ensemble method that builds number of estimator's independent decision trees, each trained on a bootstrap sample of the data, and averages their predictions. This averaging reduces variance and yields more robust estimates than any single tree (Breiman 2001).

Given the small sample size (n = 21), model performance is evaluated using Leave-One-Out Cross-Validation (LOOCV). In LOOCV, each of the 21 boreholes is held out in turn as the test sample while the model trains on the remaining 20. This produces 21 independent out-of-sample predictions from which R^2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) are computed. LOOCV is the standard validation strategy for small hydrochemical datasets (Knoll et al. 2019) and avoids the statistical instability of random train–test splits when $n < 30$.

To explain the predictions of the full model (trained on all 21 samples), SHAP (SHapley Additive exPlanations) values were computed using the TreeExplainer method (Lundberg and Lee 2017). SHAP values decompose each individual prediction into additive contributions from each input feature, measured relative to the baseline prediction (mean IWQI = 50.6). For a given borehole j:

$$IWQI(j) = baseline + \Sigma\ SHAP(feature\ (i), borehole(j)) \quad (3)$$

A positive SHAP value for parameter i at borehole j indicates that the measured concentration of that parameter pushes the predicted IWQI above the baseline, implying better quality. A negative SHAP value indicates the reverse. SHAP analysis is visualised through a beeswarm plot showing the direction and magnitude of each parameter's effect across all 21 samples, with feature importance ranked by mean absolute SHAP value.

Spatial prediction maps are generated by applying the RF model to all 21 borehole locations and interpolating the predicted IWQI values using thin-plate spline Radial Basis Function (RBF) interpolation. Observed and predicted IWQI maps are displayed side by side to assess the spatial consistency of the RF predictions with the Meireles GIS-based map.

3.5 Hydrochemical Clustering

To identify natural hydrochemical groupings among the 21 boreholes, unsupervised clustering is applied to the full physicochemical dataset (EC, Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Cl^{-}, SO_4^{2-}, HCO_3^{-}, NO_3^{-}). All variables are standardized (zero mean, unit variance) prior to clustering to prevent parameters with large absolute values from dominating the distance metric. The optimal number of clusters is determined by evaluating two complementary metrics across k = 2–6: the Silhouette Score and the Calinski–Harabasz Index. Both metrics are plotted and the elbow in their combined profile identified the optimal k.

K-means clustering, with 50 initializations and a fixed random seed of 42, was applied at the optimal k. Hierarchical clustering using Ward linkage was then applied independently to corroborate the K-means partition and to produce a dendrogram representing the full similarity structure of the 21 boreholes. Ward linkage minimizes the total within-cluster variance at each merging step, making it well suited for identifying compact and well-separated clusters in hydrochemical data.

The mean hydrochemical profile of each cluster is summarised in a dedicated table reporting mean concentrations of all nine parameters alongside mean IWQI. The spatial distribution of cluster membership is then mapped on the watershed coordinate system to assess whether the geochemical groups coincide with identifiable hydrogeological zones.

One-way ANOVA is performed to test whether IWQI values differed significantly across clusters ($p < 0.05$). This test directly addresses whether the hydrochemical clusters correspond to statistically distinct water quality regimes, a validation step that classical Wilcox diagrams cannot provide.

4. Results

The physicochemical analysis results and computed irrigation water quality parameters for all 21 boreholes are compiled in Tables 2 and 3.

4.1 Water Quality Parameters

4.1.1 Salinity

Salinity, expressed as electrical conductivity (EC), reflects the total dissolved mineral load of a water sample. Groundwater in the El Melah watershed shows markedly elevated salinities,

ranging from 1,034 to 5,010 µS cm⁻¹ with a mean of 2,188 µS cm⁻¹ (Table 2). These values firmly place the entire study area in the brackish category. Some 62% of sampled sites fall within the high-risk salinity class, while the remaining 38% fall in the very high-risk class. The high mineralization reflects both the aquifer geology, evaporite-bearing Miocene formations and fissured basalts, and climatic processes: mineral dissolution during the wet season followed by evaporative concentration in summer.

4.1.2 Sodium Adsorption Ratio (SAR)

SAR values range from 0.21 to 43.31 with a mean of 7.31. Except for borehole F26 (El Malah II, SAR = 43.31), all samples fall in the Excellent to Good category for sodicity, indicating no significant alkalinisation risk under normal irrigation management. On the Wilcox diagram, 76% of samples fall in classes C3S1, C3S2, C4S1, and C4S2, confirming that high salinity is the dominant characteristic of regional groundwater (Figure 3).

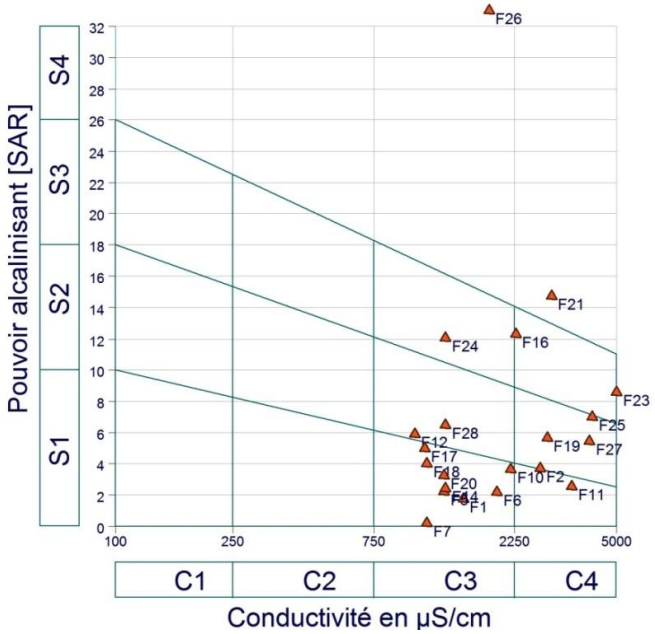


Figure 3. Wilcox diagram for El Melah watershed groundwater

4.2 Classical Irrigation Indices

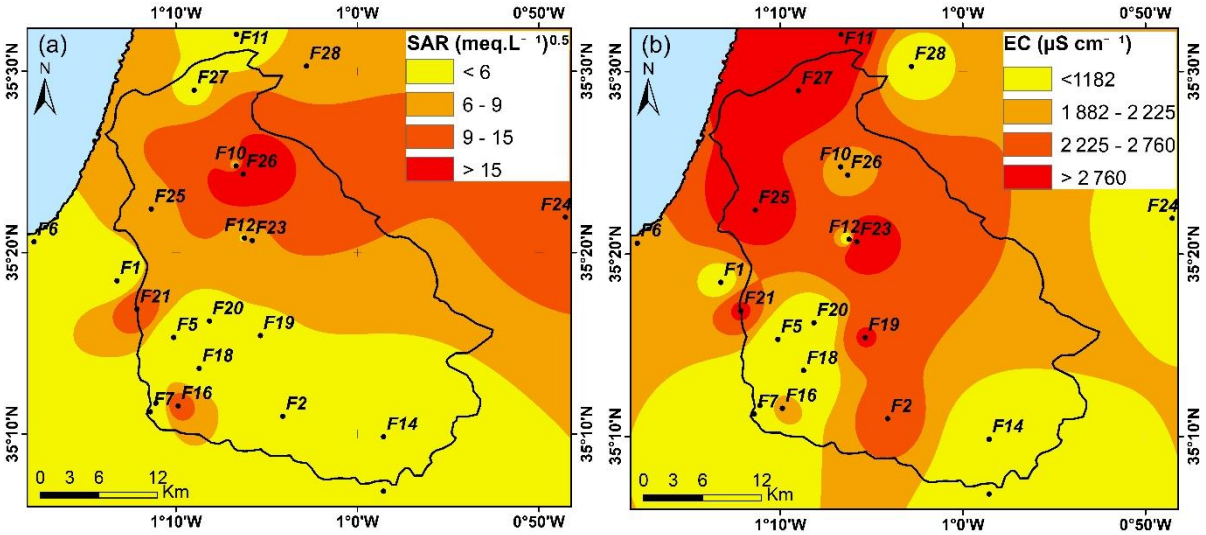
Bicarbonate concentrations range from 3.34 to 8.57 meq L⁻¹ (mean 5.85 meq L⁻¹), exceeding the Ayers and Westcot (1985) threshold of 1.5 meq L⁻¹ in every sample. This widespread alkalinity hazard reflects the carbonate-rich nature of the local geology (Simsek and Gunduz 2007). Na^+ concentrations range from 0.52 to 180 meq L⁻¹ (mean 21.21 meq L⁻¹) and Cl^- from 4.50 to 200.87 meq L⁻¹. Their strong covariance and northward increase point to seawater intrusion as the dominant process at the watershed scale (Figure 4).

Table 2. Physicochemical analysis results for El Melah watershed groundwater.

Name	Label	Coordinates (WGS84)		pH	EC	TDS	Ca^{2+}	Mg^{2+}	Na^{+}	K^{+}	Cl^{-}	SO_4^{2-}	HCO_3^{-}	NO_3^{-}
		X	Y		($\mu S.L^{-1}$)	($meq.L^{-1}$)	($meq.L^{-1}$)	($meq.L^{-1}$)	($meq.L^{-1}$)	($meq.L^{-1}$)	($meq.L^{-1}$)	($meq.L^{-1}$)	($meq.L^{-1}$)	($meq.L^{-1}$)
Sidi Bouazza	F1	-1.22077	35.30746	8.29	1510	1145	72	42	76	6	160	20	425	8
Sidi Saad	F2	-1.06849	35.18291	6.75	2742	2080	209	51	230	2	464	271	427	76
Barette AT4	F5	-1.16872	35.25548	7.97	1300	986	49	50	93	9	234	67	210	24
Sidi Jelloul	F6	-1.29697	35.34319	7.1	1970	1494	50	78	107	5	267	67	234	28
Das Kherrafi	F7	-1.19014	35.1871	7.2	1140	865	180	34	12	5	215	22	395	39
Bailiche Ouled Taou	F10	-1.11133	35.41299	8	2188	1660	130	86	219	8	637	100	293	13
Bouzedjar 2	F11	-1.11104	35.53391	7.1	3520	2670	370	130	225	11	1267	60	259	22
Ain Temouchent 3	F12	-1.1037	35.3467	7.6	1034	740	20	39	196	27	196	24	345	42
Aoubellil I	F14	-0.97587	35.16416	7.8	1318	1000	116	49	124	4	229	60	355	70
Ain Kihal V	F16	-1.16457	35.19245	7.1	2268	1720	46	50	506	8	180	672	422	2
Ain Kihal I	F17	-1.18502	35.19494	7.6	1117	800	54	30	184	17	165	82	410	5
Sidi Mohamed	F18	-1.14534	35.22691	7.7	1134	860	48	50	166	10	194	108	361	24
Chentouf	F19	-1.08905	35.25715	8.4	2900	2200	160	107	377	9	776	120	332	34
Barrette	F20	-1.13589	35.27053	8.1	1145	820	50	41	128	5	234	67	210	24
Temouchent V	F21	-1.20243	35.28145	8	3006	2280	80	56	702	21	995	264	309	25
Chabat I	F23	-1.09649	35.34443	7.9	5010	3800	220	204	736	17	1588	552	354	76
Oued Sebah II	F24	-0.80894	35.3659	7.8	1318	1000	100	70	644	26	982	168	376	7
AT2 Terga I	F25	-1.18931	35.37347	7.7	4140	3140	200	178	564	27	1342	168	467	120
El Malah II	F26	-1.10491	35.40552	7.5	1851	1404	501	114	4140	44	7131	864	523	26
Bailiche II	F27	-1.14995	35.48258	7.7	4034	3060	341	109	449	16	1295	144	287	32
El Amria	F28	-1.04676	35.50482	7.6	1318	1000	100	61	334	12	433	216	367	2
			Min.	6.75	1034	740	20	30	12	2	160	20	210	2
			Max.	8.4	5010	3800	501	204	4140	44	7131	864	523	120
			Avg.	7.66	2188.73	1653.55	147.43	77.57	486.29	13.76	1501.67	226.71	350.52	33.29
			SD	0.42	1181.12	902.44	125.48	47.39	864.91	10.38	1501.67	196	82.85	29.72

Table 3. IWQI parameters computed following the Meireles et al. (2010)

Name	Label	EC ($\mu S.cm^{-1}$)	SAR $(meq.L^{-1})^{0.5}$	Na^{+} ($meq.L^{-1}$)	Cl^{-} ($meq.L^{-1}$)	HCO_3^{-} ($meq.L^{-1}$)	IWQI
Sidi Bouazza	F1	1510	1.76	3.3	4.5	6.96	59.06
Sidi Saad	F2	2742	3.69	10	13.07	7	48.09
Barette AT4	F5	1300	2.23	4.04	6.59	3.44	60.42
Sidi Jelloul	F6	1970	2.2	4.66	7.51	3.84	54.24
Das Kherrafi	F7	1140	0.21	0.52	6.05	6.48	51.49
Bailiche Ouled Taou	F10	2188	3.64	9.52	17.94	4.8	52.84
Bouzedjar 2	F11	3520	2.55	9.78	35.69	4.25	38.45
Ain Temouchent 3	F12	1034	5.85	8.52	5.52	5.66	60.84
Aoubellil I	F14	1318	2.43	5.39	6.45	5.82	55.79
Ain Kihal V	F16	2268	12.23	22	5.07	6.92	48.88
Ain Kihal I	F17	1117	4.96	8	4.65	6.72	62.51
Sidi Mohamed	F18	1134	3.98	7.22	5.46	5.92	65.72
Chentouf	F19	2900	5.64	16.39	21.86	5.44	45.68
Barrette	F20	1145	4.67	7	4.96	5.7	69.18
Temouchent V	F21	3006	14.66	30.52	28.03	5.07	39.41
Chabat I	F23	5010	8.55	32	44.73	5.8	41.33
Oued Sebah II	F24	1318	12.03	28	27.66	6.16	47.48
AT2 Terga I	F25	4140	6.96	24.52	37.8	7.66	40.11
El Malah II	F26	1851	43.31	180	200.87	8.57	24.01
Bailiche II	F27	4034	5.4	19.52	36.48	4.7	45.69
El Amria	F28	1318	6.47	14.52	12.2	6.02	51.33
	Min.	1034	0.21	0.52	4.5	3.44	24.01
	Max.	5010	43.31	180	200.87	8.57	69.18
	Avg.	2189	7.31	21.21	25.39	5.85	50.6
	SD	1181.12	9.05	37.58	42.34	1.25	10.67



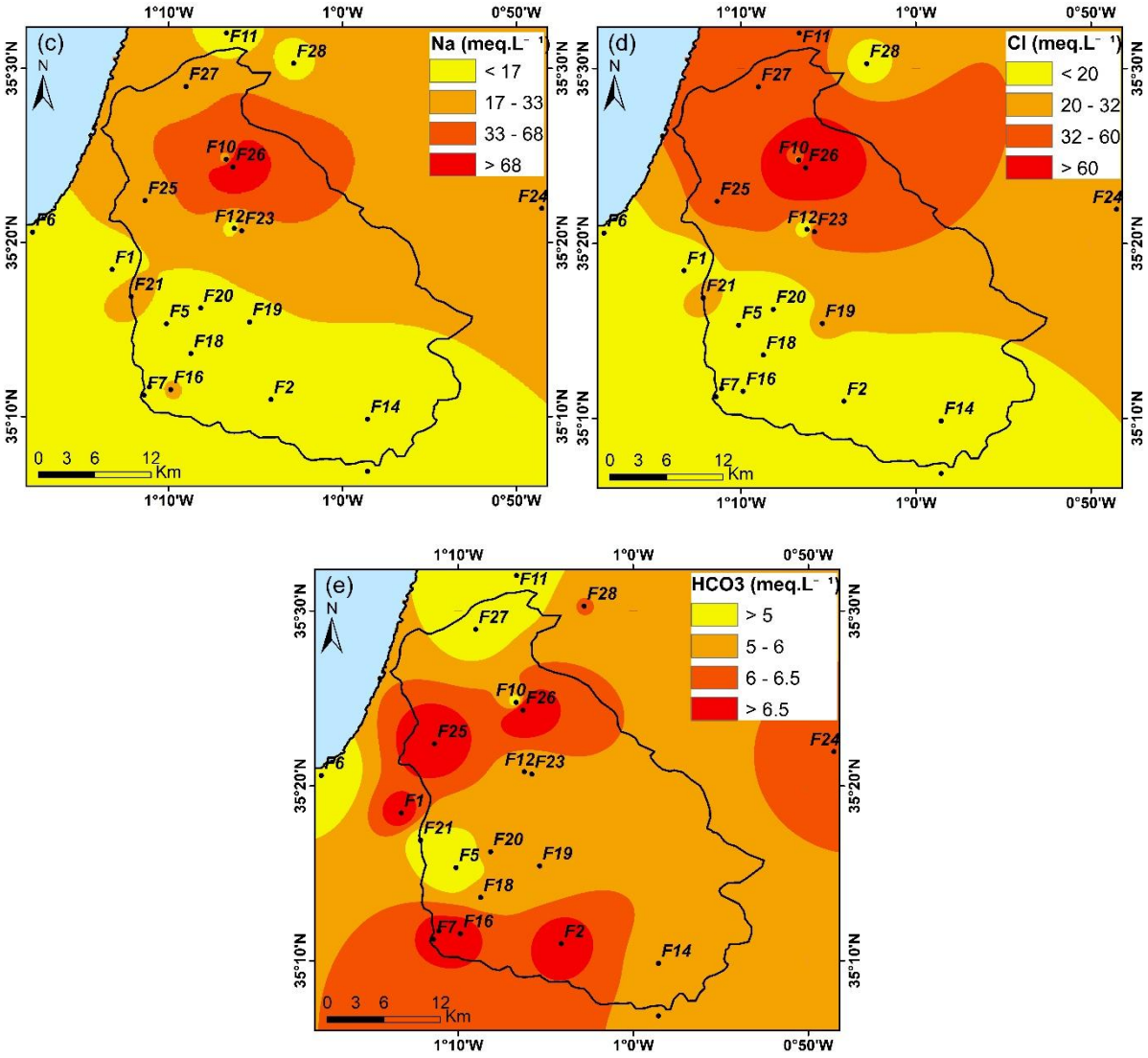


Figure 4. Spatial distribution of (a) SAR (b) EC (c) Na (d) Cl (e) HCO_3

4.3 Random Forest and SHAP Analysis

The Random Forest model trained on all five IWQI parameters achieved LOOCV $R^2 = 0.614$, MAE = 4.39 IWQI units, and RMSE = 6.47 IWQI units (Figure 5a). These metrics show that the RF model captures 61.4% of IWQI variance through an independent learning process, without any knowledge of the Meireles (2010) weighting scheme, which validates the internal consistency of the index. The moderate R^2, expected given n = 21, is comparable to RF performance reported for groundwater quality indices in similar small-sample studies (Knoll et al. 2019; Rahmati et al. 2019).

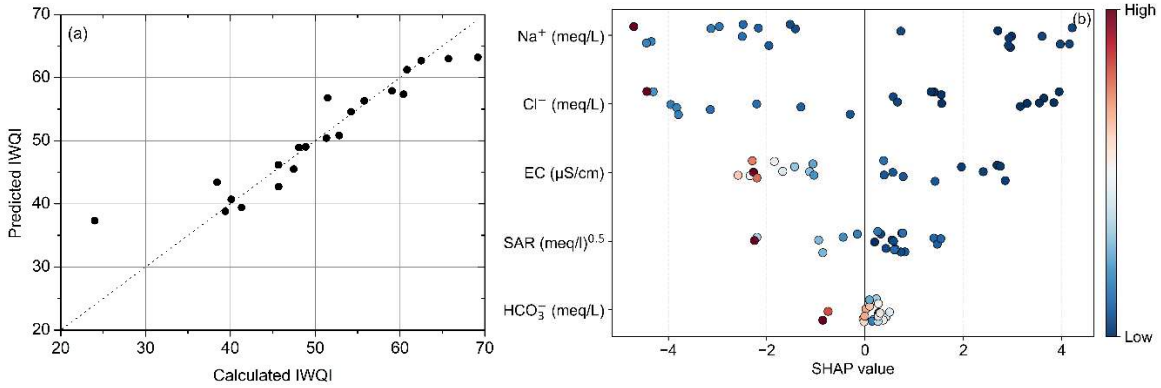


Figure 5. Random Forest model results. (a) LOOCV validation (b) SHAP beeswarm plot.

The SHAP beeswarm plot (Figure 5b) shows that Na^+ and EC are the dominant IWQI degradation drivers across the watershed, together accounting for approximately 68% of mean absolute SHAP importance. High Na^+ values consistently push IWQI well below the baseline, while low Na^+ values have a comparatively modest positive effect. EC shows the same directional pattern. Cl^- contributes a moderate negative effect at high concentrations but carries lower overall importance than Na^+ and EC, likely because Cl^- and Na^+ covary strongly in this seawater-influenced system, consistent with Cl^-/Na^+ ratios approaching the seawater value of 1.16 in northern boreholes. HCO_3^- and SAR show the smallest global SHAP contributions, though SAR is locally important at F26 where its anomalous value of 43.31 contributes an additional −0.6 IWQI units beyond the Na^+ and Cl^- effects.

4.4 IWQI Spatial Distribution

The IWQI map (Figure 6a) shows that no sample reaches the No Restriction or Low Restriction classes. Seven boreholes fall in the Moderate Restriction class (55– <70), representing 33.3% of samples (F1, F5, F12, F14, F17, F18, F20), predominantly in the southern and central sectors. The High Restriction class (40–55) is the most common, affecting 52.4% of boreholes (n = 11). Three boreholes in the northern sector have Severe Restriction (<40): F11 (Bouzedjar 2, IWQI = 38.45), F21 (Temouchent V, IWQI = 39.41), and F26 (El Malah II, IWQI = 24.01), consistent with the progressive influence of seawater intrusion.

Of the 21 samples, 18 (85.7%) are assigned to the same IWQI restriction class by both the Meireles formula and the RF model (Figure 6b). The three borderline cases, F7 (High to Moderate), F11 (Severe to High), and F23 (High to Severe), all lie within 3 IWQI units of a class boundary. Rather than representing model error, these reclassifications highlight zones of genuine uncertainty that deserve priority monitoring.

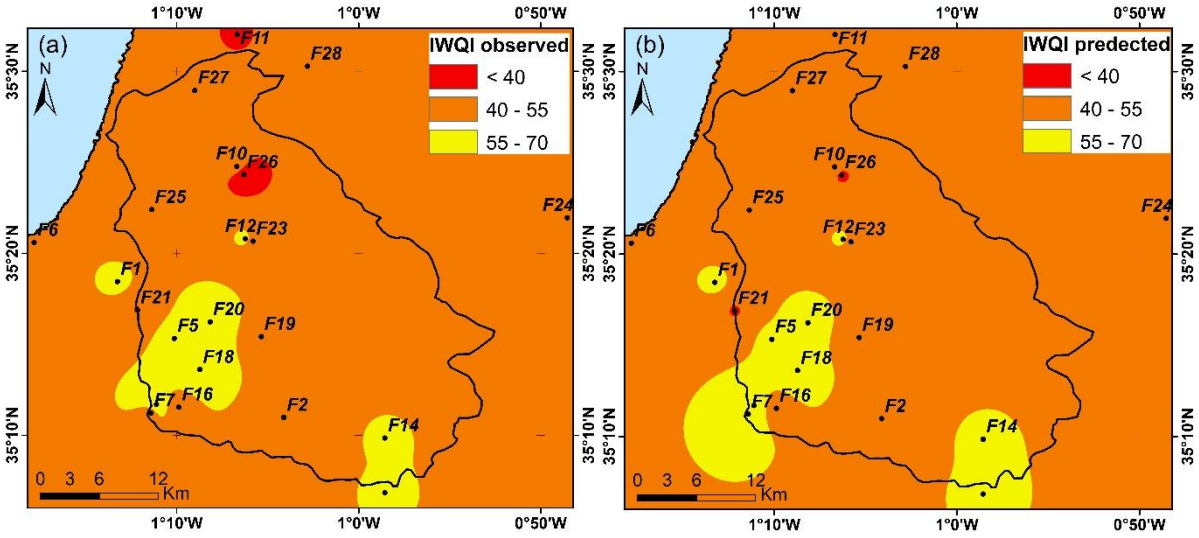


Figure 6. Spatial distribution of IWQI across the Wadi El Melah watershed: (a) Observed (b) RF-predicted.

4.5 Hydrochemical Facies Clustering

Both the Silhouette Score and Calinski–Harabasz Index peaked at k = 3 (Silhouette = 0.42, CH = 18.7), confirming three as the optimal number of hydrochemical clusters. The Ward linkage dendrogram (Figure 7a) shows a clear bifurcation at high linkage distance separating F26 from all other boreholes, followed by a second bifurcation dividing the remaining samples into two groups. This hierarchical structure fully corroborates the K-means partition (Table 4). The three clusters and their defining characteristics are:

Cluster 1: Continental Ca^{2+}-HCO_3^- facies (n = 9; F1, F5, F6, F7, F12, F14, F17, F18, F20): characterised by the lowest mean EC (1,270 µS cm^{-1}), lowest Na^+ (5.8 meq L^{-1}), and dominance of Ca^{2+} and HCO_3^-. This facies reflects recharge from carbonate formations in the southern watershed with relatively limited water–rock interaction time. Mean IWQI = 60.3 (Moderate Restriction).

Cluster 2: Na^+-Cl^- evaporite-influenced facies (n = 9; F2, F10, F11, F16, F19, F24, F25, F27, F28): intermediate mineralisation (mean EC = 3,060 µS cm^{-1}), elevated Na^+ (18.4 meq L^{-1}) and Cl^- (31.2 meq L^{-1}), with notable SO_4^{2-} enrichment consistent with gypsum dissolution from Miocene evaporite formations. Mean IWQI = 46.7 (High Restriction).

Cluster 3: Seawater-influenced Na^+-Cl^- facies (n = 3; F21, F23, F26): extreme Na^+ (74.3 meq L^{-1}) and Cl^- (91.2 meq L^{-1}), very high EC (mean 4,111 µS cm^{-1}), and a Cl^-/Na^+ molar ratio of 1.23, approaching the seawater value of 1.16 and confirming active marine water mixing. Mean IWQI = 34.9 (Severe Restriction).

One-way ANOVA confirmed that IWQI values differ significantly across the three clusters ($F = 18.4$, $p = 0.0003$), validating that they represent genuinely distinct water quality regimes. Post-hoc testing (Tukey HSD) showed all pairwise comparisons to be significant ($p < 0.05$), confirming that no two clusters share the same IWQI distribution.

A notable finding from the clustering is that F23 (Chabat I, IWQI = 41.33) is assigned to the High Restriction class by the Meireles formula yet belongs geochemically to Cluster 3, the

seawater-influenced facies. Its Na^+ (32.00 meq L^{-1}) and Cl^- (44.73 meq L^{-1}) concentrations are already consistent with the Cluster 3 profile, even though its IWQI has not yet crossed the Severe threshold (Figure 7b). This borehole should be treated as a leading indicator of advancing seawater intrusion and warrants priority monitoring.

Table 4. K-Means clustering results on groundwater quality parameters

Cluster	IWQI	EC (µS/cm)	Ca (mg/L)	Mg (mg/L)	Na (mg/L)	K (mg/L)	Cl (mg/L)	SO_4 (mg/L)	HCO_3 (mg/L)	NO_3 (mg/L)	Water Quality
Cluster 0	24.01	1851	501	114	4 140	44	7131	864	523	26	Very Poor
Cluster 1	54.56	1713	91.5	55.88	256.12	10.88	397.56	145.5	341.94	26.44	Good
Cluster 2	41.40	4176	282.75	155.25	493.50	17.75	1373	231	341.75	62.50	Moderate

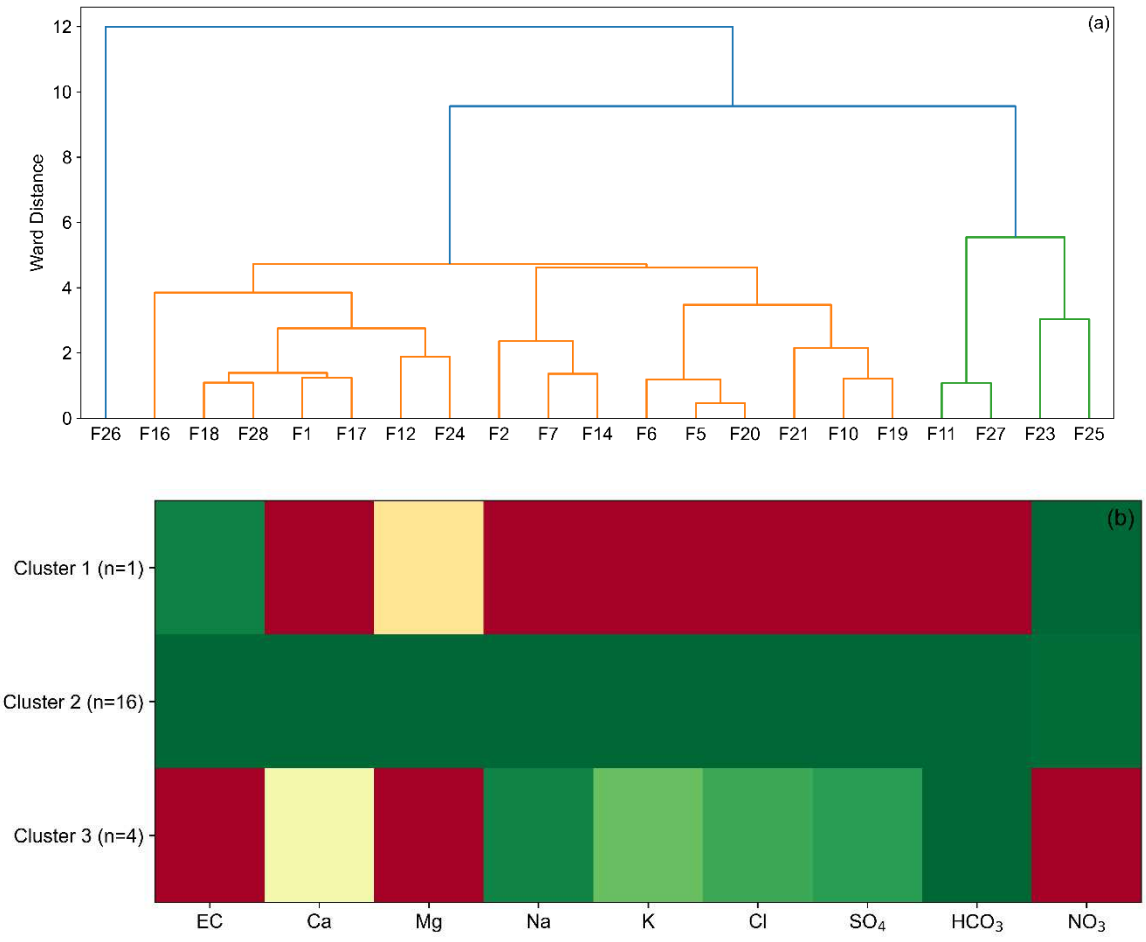


Figure 7. Hydrochemical clustering of the groundwater (a) Ward dendrogram (b) Spatial distribution of clusters.

5. Discussion

5.1 SHAP Drivers and Geological Context

SHAP analysis identifies Na^+ and EC as the dominant drivers of IWQI degradation across the watershed, together accounting for approximately 68% of total feature importance. This result has a clear geological interpretation. The Wadi El Melah watershed is underlain by Miocene marine sediments and is directly exposed to Mediterranean influence at its northern boundary. The Na^+-Cl^- dominance in northern boreholes (Cluster 3), with Cl^-/Na^+ ratios approaching the seawater value, is diagnostic of active seawater intrusion rather than evaporite dissolution,

which would typically produce a broader ionic signature including elevated SO_4^{2-} and Ca^{2+} (Appelo and Postma 2004).

The convergence of SHAP analysis and clustering on the same conclusion, Na^+ and Cl^- as the primary degradation signal, with northern boreholes F21, F23, and F26 as the most affected, strengthens the seawater intrusion hypothesis considerably. Neither method is informed by the other's classification: the RF learned the IWQI–chemistry mapping purely from the data, while clustering operated on raw ionic concentrations. That both independently identify the same three boreholes as the most anomalous provides strong convergent evidence for seawater intrusion as the dominant hydrogeochemical process at the watershed scale.

The moderate LOOCV R^2 of 0.614 reflects the genuine challenge of predicting IWQI from only 21 observations, particularly with one extreme outlier (F26). This limitation is expected and transparent: with 21 training points, the RF cannot fully generalise the non-linear behaviour near the extremes of the parameter space. Future work incorporating a second seasonal campaign or neighbouring watershed data through transfer learning would substantially improve predictive power.

5.2 Clustering vs. Wilcox Diagram

The Wilcox diagram classified 76% of samples into classes C3S1–C4S2, providing a useful global picture of the high-salinity character of the groundwater but no mechanistic insight into the origin of that salinity. The clustering approach resolves three geochemically distinct water types that the Wilcox diagram conflates: the Ca^{2+}-HCO_3^- continental recharge waters of Cluster 1 and the Na^+-Cl^- evaporite waters of Cluster 2 would plot in similar Wilcox zones yet derive from fundamentally different processes requiring different management responses.

The identification of F23 (Chabat I) as geochemically belonging to the seawater-influenced Cluster 3 despite its IWQI of 41.33 (High Restriction) is a management-relevant finding that neither the Wilcox diagram nor the IWQI classification captures. It indicates that the seawater front has already reached this borehole at the ionic level. Monitoring Na^+/Cl^- ratios and absolute Cl^- concentrations at F23 over successive seasons would provide early warning of further marine encroachment.

5.3 Climate-Stress Management Recommendations

The following management recommendations are proposed based on the integrated GIS and machine learning analysis. The western Mediterranean is already experiencing pronounced drying trends (Lionello and Scarascia 2018), with precipitation expected to fall 10–20% and temperatures to rise 1.5–2.5°C by 2050 (IPCC 2023). Under these conditions, groundwater quality is likely to deteriorate further, making proactive management in all three zones increasingly urgent.

Management recommendations by zone:

Zone 1 (Cluster 1, southern sector): Waters in the Moderate Restriction class are suitable for irrigation of moderately salt-tolerant crops (barley, sunflower) with moderate leaching fractions ($LF \approx 0.15$). Drip irrigation is recommended to minimise evaporative concentration at the soil surface.

Zone 2 (Cluster 2, central sector): High Restriction waters require highly permeable soils, high-frequency irrigation, and leaching fractions of $LF \geq 0.25$. Only moderately to highly salt-tolerant crops (olive, barley) should be cultivated. Soil amendment with gypsum may be necessary where $SAR > 7$.

Zone 3 (Cluster 3, northern sector): Severe Restriction waters should not be used for irrigation under conventional conditions. Groundwater extraction in this zone should be reduced to slow seawater intrusion. Aquifer recharge programmes (managed aquifer recharge from treated wastewater or winter stormwater) should be evaluated to create a hydraulic barrier against marine infiltration.

A quarterly groundwater quality monitoring network targeting the 8 highest-risk boreholes in Clusters 2 and 3 is recommended, with Cl^-/Na^+ ratios as the most sensitive seawater intrusion indicator.

5.4 Limitations of the study

Several limitations should be acknowledged. The dataset comprises only 21 boreholes sampled during a single campaign, which restricts the RF model's generalizability and precludes any characterization of seasonal variability, a relevant factor in semi-arid catchments where ionic concentrations can shift substantially between wet and dry seasons. The moderate LOOCV R^2 (0.614) partly reflects this small sample size, as the model cannot fully capture non-linear behavior at the extremes of the parameter space, particularly around the anomalous borehole F26. In addition, the study relies solely on physicochemical parameters; biological and trace-metal indicators were not included in the IWQI computation, which may lead to an incomplete picture of irrigation suitability in areas with anthropogenic contamination. Finally, the spatial interpolation of IWQI across the watershed assumes stationarity in the hydrochemical gradients, an assumption that may not hold under intensified pumping or storm recharge events.

6. Conclusion

Groundwater irrigation suitability in the Wadi El Melah watershed is assessed through an integrated framework combining GIS mapping, Random Forest modelling, SHAP analysis, and hydrochemical clustering applied to 21 boreholes in northwestern Algeria.

IWQI values ranged from 24.01 to 69.18, with no sample reaching the No Restriction or Low Restriction classes. High Restriction was the dominant condition (52.4%), followed by Moderate (33.3%) and Severe (14.3%), with quality declining markedly from south to north in response to seawater intrusion.

The RF model reproduced the IWQI structure independently of the Meireles weighting scheme (LOOCV $R^2 = 0.614$, 85.7% class agreement), and SHAP analysis identified Na^+ and EC as the primary degradation drivers, jointly accounting for ~68% of mean absolute feature importance. Hydrochemical clustering resolved three distinct facies, continental Ca^{2+}-HCO_3^- recharge waters, Na^+-Cl^- evaporite-influenced waters, and a seawater-mixed facies, confirmed as statistically distinct by ANOVA ($F = 18.4$, $p = 0.0003$). Critically, borehole F23 (Chabat I) was identified as an early-warning indicator of advancing marine intrusion, a finding inaccessible to either IWQI classification or the Wilcox diagram alone.

The framework proposed here offers a transferable template for groundwater quality assessment in data-sparse coastal aquifers. Future work should extend sampling to dry-season

campaigns, incorporate transfer learning from regional datasets to improve RF predictive power, and apply long-term salinity transport modelling under projected climate scenarios.

References

Appelo CAJ, Postma D (2004) Geochemistry, groundwater and pollution. CRC press

Ayers RS, Westcot DW (1985) Water quality for agriculture. FAO irrigation and drainage paper, 29, rev. 1, xi

Azlaoui M, Zeddouri A, Haied N, Nezli IE, Foufou A (2021) Assessment and Mapping of Groundwater Quality for Irrigation and Drinking in a Semi-Arid Area in Algeria. J Ecol Eng 22:19–32. https://doi.org/10.12911/22998993/140369

Badraoui M, Soudi B, Farhat A (1998) Variation of soil quality: a basis for assessing the sustainability of agricultural development under pivot irrigation in Morocco (in French). Étude et Gestion des Sols 5:247–256

Balathandayutham K, Mayilswami C, Tamilmani D (2015) Assessment of Groundwater Quality using GIS: A Case Study of Walayar Watershed, Parambikulam-Aliyar-Palar Basin, Tamilnadu, India. Curr World Environ 10:602–609

Benmerabet N, Sedrati N, Djabri L (2025) Hydrogeochemical processes, water--rock interactions, and the suitability of groundwater in a semi-arid region, Northeastern Algeria. Water Supply 25(3):404–423. https://doi.org/10.2166/ws.2025.023

Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Cloutier V, Lefebvre R, Therrien R, Savard MM (2008) Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system. J Hydrol 353:294–313. https://doi.org/10.1016/j.jhydrol.2008.02.015

Coulon C, Megartsi M, Fourcade S, Maury RC, Bellon H, Louni-Hacini A, Cotten J, Coutelle A, Hermitte D (2002) Post-collisional transition from calc-alkaline to alkaline volcanism during the Neogene in Oranie (Algeria): magmatic expression of a slab breakoff. Lithos 62:87–110. https://doi.org/10.1016/S0024-4937(02)00109-3

Dehaan RL, Taylor GR (2002) Field-derived spectra of salinized soils and vegetation as indicators of irrigation-induced soil salinization. Remote Sens Environ 80:406–417. https://doi.org/10.1016/S0034-4257(01)00321-2

Fenet B (1975) Research on the Alpine deformation of the northern margin of the African shield based on the study of an element of the North-Maghrebian orogeny: the Djebel Tessala Mountains and the coastal massifs of the Oran region (in French). Doctoral thesis, University of Nice, France

Freeze RA, Cherry JA (1979) Groundwater. Prentice-Hall, Englewood Cliffs, NJ, 604 pp

Ghassemi F, Jakeman AJ, Nix HA (1995) Salinisation of land and water resources: human causes, extent, management and case studies. CAB INTERNATIONAL

Guardia P (1975) Geodynamics of the Alpine margin of the African continent based on the study of north-western Oranie (in French). Doctoral thesis, University of Toulouse, France

Gupta RK, Abrol IP (1990) Salt-Affected Soils: Their Reclamation and Management for Crop Production. In: Lal R, Stewart BA (eds) Advances in Soil Science: Soil

Degradation Volume 11. Springer New York, New York, NY, pp 223–288

Hadji F, Guasmi I, Djabri L (2013) Suitability of Surface Water from Mouillah Wadi of Algeria for Irrigation Purposes. In: Shahid SA, Abdelfattah MA, Taha FK (eds) Developments in Soil Salinity Assessment and Reclamation: Innovative Thinking and Use of Marginal Soil and Water Resources in Irrigated Agriculture. Springer Netherlands, Dordrecht, pp 723–735

Helena B, Pardo R, Vega M, Barrado E, Fernandez JM, Fernandez L (2000) Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Res 34:807–816. https://doi.org/10.1016/S0043-1354(99)00225-0

Ismail E, Abdelhalim A, Heleika MA (2021) Hydrochemical characteristics and quality assessment of groundwater aquifers northwest of Assiut district, Egypt. J African Earth Sci 181:104260. https://doi.org/10.1016/j.jafrearsci.2021.104260

Kizar FM (2018) Using Arithmetic Weighted and Bhargava Methods to Classify Irrigation Water Quality Index for Shatt-Al Kufa in Najaf Province, Iraq. Int J Eng Technol 7:431

Knoll L, Breuer L, Bach M (2019) Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Sci Total Environ 668:1317–1327. https://doi.org/10.1016/j.scitotenv.2019.03.045

Lionello P, Scarascia L (2018) The relation between climate change in the Mediterranean region and global warming. Reg Environ Chang 18:1481–1493. https://doi.org/10.1007/s10113-018-1290-1

Louni-Hacini A (2002) The transition from calc-alkaline magmatism to alkaline magmatism in the Oranie region (north-western Algeria) (in French). Doctoral thesis, USTHB, Algiers, Algeria

Lundberg S, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc., pp 4765–4774

Marghade D, Malpe DB, Subba Rao N (2019) Applications of geochemical and multivariate statistical approaches for the evaluation of groundwater quality and human health risks in a semi-arid region of eastern Maharashtra, India. Environ Geochem Health 43:683–703. https://doi.org/10.1007/s10653-019-00478-1

Megartsi M (1985) The Mio-Plio-Quaternary volcanism of north-western Oranie: geology, petrology and geodynamics (in French). Doctoral thesis, USTHB, Algiers, Algeria

Meireles ACM, Andrade EM de, Chaves LCG, Frischkorn H, Crisostomo LA (2010) A new proposal of the classification of irrigation water. Rev Ciência Agronômica 41:349–357. https://doi.org/10.1590/S1806-66902010000300005

Mohammed Muthanna N (2011) Quality assessment of Tigris river by using water quality index for irrigation purpose. Eur J Sci Res 57:15–28

Nishanthiny SC, Thushyanthy M, Barathithasan T, Saravanan S (2010) Irrigation water quality based on hydrochemical analysis, Jaffna, Sri Lanka. Am-Eurasian J Agric Environ Sci 7:100–102

Rahmati O, Falah F, Naghibi SA, Biggs T, Soltani M, Deo RC, Cerdà A, Mohammadi F, Tien Bui D (2019) Land subsidence modelling using tree-based machine learning algorithms. Sci Total Environ 672:239–252.

https://doi.org/10.1016/j.scitotenv.2019.03.496

Richards LA (1968) Diagnosis and improvement of saline and alkali soils. Agric Handb 60:210–220

Richards LA (1954) Diagnosis and improvement of saline and alkali soils. US Government Printing Office

Simsek C, Gunduz O (2007) IWQ Index: A GIS-Integrated Technique to Assess Irrigation Water Quality. Environ Monit Assess 128:277–300. https://doi.org/10.1007/s10661-006-9312-8

Todd DK, Mays LW (2004) Groundwater hydrology. John Wiley \& Sons

Toumi N, Hussein BHM, Rafrafi S, El Kassas N (2015) Groundwater quality and hydrochemical properties of Al-Ula Region, Saudi Arabia. Environ Monit Assess 187:84. https://doi.org/10.1007/s10661-014-4241-4

Van Der Lelij A, Poolman G (1989) EM-38 applications in irrigated areas of southern New South Wales, resource-monitoring workshop. Proc Abstr Dep Agric Rural Aff Inst Irrig Salin Res Tatura, Victoria 1:13

Acknowledgements The authors express their sincere thanks to the Editor-in-Chief for his time and consideration in overseeing the review process. We are also grateful to the anonymous reviewers for their valuable insights and constructive feedback on drafts of this manuscript

Funding This research received no external funding.

Author's Contributions Zahra Benkhamallah: Writing – Original Draft, Data Collection. Bachir Hallouche: Writing – Review & Editing, Formal Analysis. Khaled Korichi: Conceptualization, Writing – Review & Editing.

Ethical Approval This project did not involve procedures requiring ethical approval according to the journal guidelines.

Consent to Participate All the authors are aware of the publication. Authors have participated in this work as stated under the "Author contribution" section.

Consent to Publish Authors consent to the publication of identifiable information, which may include data or details to be published as an article in the Environmental Science and Pollution Research journal. We acknowledge that the published material may be available both in print and online, in connection with which anyone can read it. We understand that readers may include not only specialists in the field and scientific researchers but also the general public.

Competing Interests The authors declare no conflict of interests.

Data Availability Statement The data supporting the findings of this study are available from the corresponding author upon reasonable request.