hydrology

MDPI

Article

Temporal, Spatial, and Lithological Controls on Groundwater Composition in the Rhône–Mediterranean–Corsica Basin: A Statistical and Machine Learning Approach

Zouhair Zeiki [1], Ismail Mohsine [2], Abderrahim Bousouis [1], Mouna El Jirari [1], Meryem Touzani [3], Abdoul Azize Barry [4], Abdelhak Bouabdli [1], Mohamed Sadiki [1], Vincent Valles [5] and Laurent Barbiero [6,*]

[1] Geosciences Laboratory, Faculty of Sciences, Ibn Tofaïl University, Kenitra 14000, Morocco; zouhair.zeiki@uit.ac.ma (Z.Z.); abderrahim.bousouis@uit.ac.ma (A.B.); mouna.eljirari@uit.ac.ma (M.E.J.); bouabdli@uit.ac.ma (A.B.); mohamed.sadiki@uit.ac.ma (M.S.)
[2] Geosciences, Water and Environment Laboratory, Faculty of Sciences Rabat, Mohammed V University, Rabat 10000, Morocco; ismail.mohsine@um5r.ac.ma
[3] National Institute of Agronomic Research, Rabat 10060, Morocco; meryem.touzani@inra.ma
[4] Geoscience and Environment Laboratory (LaGE), Department of Earth Sciences, Joseph Ki-Zerbo University, Ouagadougou 7021, Burkina Faso; azize.barry@ujkz.bf
[5] Mixed Research Unit EMMAH (Environnement Méditerranéen et Modélisation des Agro-Hydrosystèmes), Hydrogeology Laboratory, Avignon University, 84916 Avignon, France; vincent.valles@outlook.fr
[6] Institut de Recherche pour le Développement, Géoscience Environnement Toulouse, Centre National de la Recherche Scientifique, University of Toulouse, Observatoire Midi-Pyrénées, UMR 5563, 14 Avenue Edouard Belin, 31400 Toulouse, France
* Correspondence: laurent.barbiero@get.omp.eu

Abstract

This study analyses groundwater quality in the Rhône–Mediterranean–Corsica (RMC) basin using 27,741 samples collected from 7625 sampling points and 224 groundwater bodies, covering a 34-year period (1987–2021). The objective is to quantify, using statistical and machine learning methods, the respective contributions of temporal variability, spatial variability, and lithology to the hydrochemical and microbiological composition of aquifers. The results reveal a marked dichotomy between microbiological and physico-chemical parameters. Fecal contamination (*E. coli*, enterococci) is dominated by a strong temporal component (33–36% of variance), whereas spatial structuring by lithology remains very limited (<11%). Conversely, major physico-chemical parameters (calcium, conductivity, sulphates, etc.) exhibit near-zero temporal variability (<7%) and a spatial structuring strongly controlled by lithology, with gains in explained variance of up to 20% through lithological grouping. Linear discriminant analysis achieves an overall classification rate of 56.2%, but with wide disparities: the crystalline basement shows the highest discrimination rate (76%), while alluvial formations are not discriminated at all (0%), as their chemical signature is more strongly influenced by anthropogenic pressures than by lithology. Machine learning methods, particularly k-nearest neighbors (k-NN), significantly improve performance (76.2% vs. 56.2% for LDA), confirming the existence of nonlinear structures in the data. Unsupervised validation via K-means confirms the reality of the lithological signal. From an operational standpoint, these findings justify a differentiated approach to groundwater management: a geological approach for chemical parameters, and an approach based on land use and agropastoral practices for bacteriological parameters. The proposed methodology—combining PCA, HCA, nested ANOVA, LDA and machine learning—provides a transferable analytical framework applicable to other major European river basins.

Check for updates

Academic Editor: Peiyue Li

Received: 21 September 2026
Revised: 5 November 2026
Accepted: 25 November 2026
Published: 28 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: groundwater; water quality; lithology; Rhône–Mediterranean–Corsica basin; fecal contamination; nitrates; PCA; ANOVA; machine learning; SISE-Eaux database; France

1. Introduction

Worldwide, groundwater is the primary source for drinking water supply, accounting for about 50% of total withdrawals on average [1]. This share varies greatly depending on the context, frequently exceeding 90% in many countries and even approaching 100% in some developing regions [2]. In Europe, and more specifically in France, nearly two-thirds of the water distributed to populations comes from aquifers—a daily volume of 18 million cubic meters [3]. This essential resource supplies millions of households across the country, yet its quality remains fragile.

Groundwater quality results from a complex interplay between natural factors—reservoir lithology, residence time, water–rock interactions—and anthropogenic pressures: over-abstraction, agricultural pollution, industrial discharges, and population density [4,5]. Preserving groundwater is therefore a major strategic issue: any lasting contamination of aquifers jeopardizes the water supply of millions of people.

The Rhône–Mediterranean–Corsica (RMC) basin, France's second-largest hydrographic district, with 130,000 km^2 and 14 million inhabitants [6], perfectly illustrates this vulnerability. Its exceptionally heterogeneous geology—Hercynian crystalline basement, Alpine and sub-Alpine formations, Mediterranean karstic carbonate systems, and Tertiary and Quaternary sedimentary deposits—imparts a remarkable hydrogeochemical diversity to its waters. Yet this same lithological heterogeneity, which controls recharge, flow paths, and water–rock interactions [7,8], also exposes aquifers to differentiated risks.

Two major threats compromise the drinkability of RMC basin waters. The first, of fecal origin, stems from extensive livestock farming and pastoral practices, particularly in karstic and crystalline areas: bacterial indicators (*E. coli*, enterococci) travel through rapid, unfiltered pathways and reach aquifers directly [9,10]. The second, chemical, is nitrate pollution from livestock effluents and nitrogen fertilizers in arable plains, which massively degrades the quality of sedimentary and alluvial water bodies [11,12]. Added to these are contaminations by iron and manganese, whose mobility depends closely on aquifer redox conditions [13].

While regional or thematic studies have explored the link between lithology and water quality [14–18], none has yet provided a rigorous statistical quantification of temporal, spatial, and local variability across the entire RMC basin. Under the WFD, the basin is subdivided into homogeneous management units called "water bodies," which constitute the reference scale for assessing and managing water quality. In particular, the discriminating role of lithology with respect to fecal contamination and nitrates—across the basin's 224 water bodies—has never been assessed within a unified hierarchical framework. This gap is all the more critical given that the Water Framework Directive (WFD, 2000/60/EC) and the River Basin Management Plan (Schéma Directeur d'Aménagement et de Gestion des Eaux: SDAGE) 2022–2027 require management based on robust quantitative assessments at the water-body scale [19,20].

Linear statistical or Machine Learning methods are now widely used for predicting the groundwater quality index and contamination risk [21–24]. In parallel, these methods have also been applied, as a complement to other analyses (Gibbs diagrams, End-member mixing analysis, spatial and geostatistical analysis…), to highlight the role of lithological heterogeneity, water pathway, and prolonged water–rock interaction on the chemical profile of groundwater [25–28].

https://doi.org/10.3390/hydrology13120320

This paper builds on a previous study [6] that established a multivariate diagnosis of water quality in the RMC basin's water bodies, by cross-referencing bacteriological and physicochemical parameters with land-use data (CORINE Land Cover). It takes a step further by mobilizing an expanded dataset of 27,741 analyses from 7625 sampling points. Its objectives are: (i) to characterize the hierarchical spatio-temporal structure of water quality variability, (ii) to assess lithology's relevance as a regionalization framework, and (iii) to test the potential of statistical and machine learning methods for predicting the lithological and hydrochemical affiliation of water bodies.

2. Materials and Methods

2.1. Study Area

The Rhône–Mediterranean–Corsica (RMC) basin covers approximately 130,000 km^2—nearly 20% of metropolitan France—and is structured around the Rhône River, opening onto the Mediterranean coastline (Figure 1). It is the second largest hydrographic district in France and is home to 14 million inhabitants (22.6% of the French population), nearly half of whom live along the coastal strips of Languedoc-Roussillon, Provence-Alpes-Côte d'Azur, and Corsica [AERMC: https://www.eaurmc.fr/ accessed on 26 January 2026].

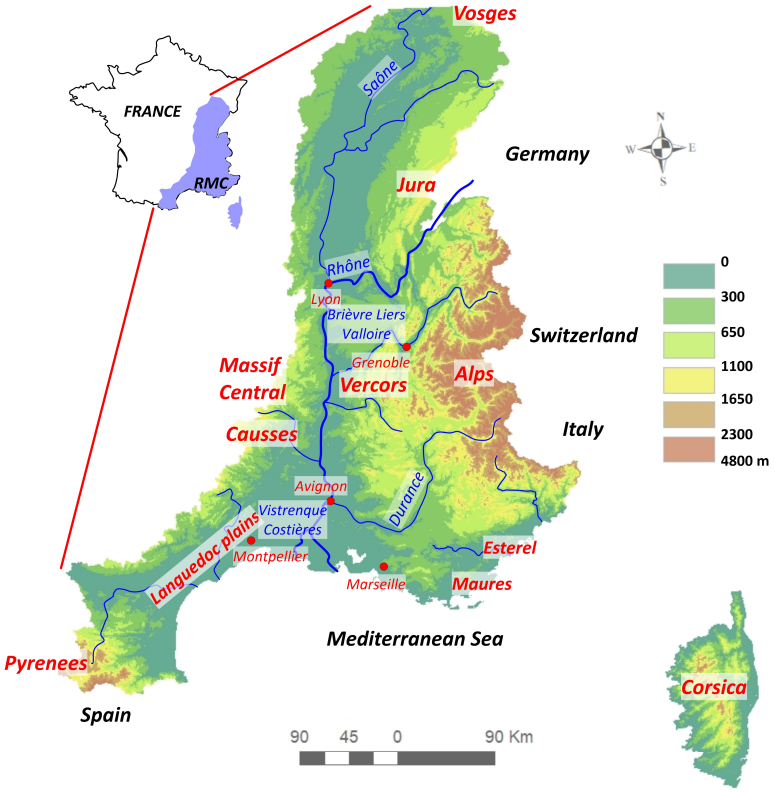

Figure 1. Location and main characteristics of the RMC basin.

Its topography ranges from sea level to Alpine peaks exceeding 4000 m, producing a marked climatic diversity: a Mediterranean regime along the southern perimeter (hot, dry summers and intense Cévennes rainfall episodes in autumn) and a colder, more humid mountain climate in the Alpine and sub-Alpine sectors. Mean annual precipitation reaches approximately 120 billion m^3, of which 60 billion m^3 is effective rainfall; however, recent years have been marked by severe droughts (notably in 2003 and 2005), leading to the drying-up of sensitive small watercourses and heightened water-use conflicts [29–31].

This climatic diversity is reflected in a highly heterogeneous soil cover—from Mediterranean brown soils and rendzinas on carbonate formations to acidic, leached soils on crystalline basements, alongside fertile alluvial soils in the Rhône and Languedoc plains—which directly influences infiltration, runoff, and aquifer recharge dynamics.

The basin also exhibits strong geological and hydrogeological diversity, combining crystalline domains (eastern Massif Central, Corsica, Maures and Estérel massifs), Alpine and sub-Alpine formations, karstic carbonate systems (Vercors, Pre-alps, Causses, Jura), and Tertiary and Quaternary sedimentary formations (Rhône corridor, Languedoc plains). This heterogeneity translates into a wide variety of aquifer types: highly permeable and vulnerable karstic aquifers in carbonate massifs [32]; lower-permeability basement aquifers in crystalline domains; productive alluvial aquifers in plains and valleys (e.g., the Bièvre–Liers–Valloire plain, ~480 km^2 and up to 100 m deep upstream [33]; the Vistrenque and Costières aquifers, >500 km^2 but only 4–20 m deep [34]; and confined or semi-confined deep sedimentary aquifers in Tertiary and Quaternary basins. This heterogeneity directly influences recharge mechanisms, groundwater flow paths, water–rock interactions, and aquifer vulnerability to contamination, as evidenced by the 12,800 water abstraction points unevenly distributed across the basin, with 74% of the total volume abstracted for drinking water supply originating from groundwater.

2.2. Data Used

2.2.1. SISE-Eaux Database

The data used in this study originate primarily from the French national SISE-Eaux database [35], which is fed by the Regional Health Agencies (Agences Régionales de Santé, ARS) as part of the sanitary monitoring of water intended for human consumption. At the national level, this database covers over 32,000 catchment stations, comprising 96% groundwater and 4% surface water. Every analysis is performed by laboratories approved by the Ministry of Health, complying with all international certifications and requirements [36] for analytical quality and reliability. The data used in this study were extracted from the administrative regions of Occitanie, Provence-Alpes-Côte d'Azur, Auvergne-Rhône-Alpes, and Bourgogne-Franche-Comté in order to encompass the study area (RMC) (https://data.ofb.fr/catalogue/data-eaufrance/, accessed on 20 January 2019). The regularity of sanitary controls varies and is influenced chiefly by two factors, first and foremost the number of inhabitants supplied by the water catchments. In densely populated areas, sampling occurs often, whereas catchments serving only a handful of households undergo far fewer controls. Past contamination events or health concerns also play a role: catchments or sectors that repeatedly fail to meet standards are subject to more frequent monitoring. This dataset pertains exclusively to groundwater and only to raw water prior to any potability treatment. It comprises 27,741 raw water samples distributed across 7625 sampling points (springs, wells, and boreholes), covering the entire RMC basin with variable density depending on the region. The analyses span a 34-year period, from 1987 to 2021. Analytical protocols have evolved over the 34-year period (1987–2021), with improvements and lower detection limits; however, these changes remain minor relative to the temporal variability of water quality. Sampling techniques have not evolved over the period, nor geographically across the study area: they are standardized and have remained unchanged, and all samplers were trained by the same company, so operator-dependent variability is also low relative to the temporal variability of water quality.

The selection of parameters is based on a compromise between the number of available data points (rows) for each parameter (columns) and the number of parameters for a multifactorial approach. Because only complete analyses were retained (full ma-

trix), all parameters have the same measurement frequency in the dataset. The selected parameters include:

- Microbiological indicators: *Escherichia coli* (*E. coli*) and enterococci (Enter.), markers of fecal contamination of animal or human origin [37,38];
- Major physicochemical parameters: Ca^{2+}, Mg^{2+}, Na^{+}, HCO_3^{-}, SO_4^{2-}, Cl^{-}, and electrical conductivity (E.C.);
- Nutrients and trace metals: nitrates (NO_3^{-}), iron (Fe), and manganese (Mn), with the latter two being sensitive to aquifer redox conditions.

Only analyses with no missing data (complete matrix) were retained. The procedure for obtaining this complete matrix has been described in detail in a previous paper [6].

2.2.2. Groundwater Body Reference Framework

Based on the geographic coordinates of the sampling points, each analysis was assigned to the corresponding groundwater body (GWB), as defined under the Water Framework Directive (2000/60/EC) [19]. This reference framework, which subdivides the territory into coherent hydrogeological units, makes it possible to respect natural discontinuities between aquifers and to avoid biases inherent to continuous spatial interpolation methods (kriging, inverse distance weighting) that artificially smooth variations between distinct geological units. In total, 224 GWBs were identified within the RMC basin, referenced as FRDGXXX in accordance with the European WISE information system. In this code, FR refers to French territory, D to the RMC basin (District Rhône–Méditerranée–Corse), G to groundwater, and XXX is the unique reference number assigned to each groundwater body.

2.3. Lithological Classification

Particular attention was paid to the lithological structuring of the groundwater bodies. The geological diversity of the RMC basin was synthesized into seven major lithological groups representing the main regional geological contexts:

- Z1—Pure crystalline basement (granites, gneisses);
- Z2—Weathered basement, metamorphic formations, and folded domains;
- Z3—Limestones and karstic systems;
- Z4—Tertiary formations, molassic deposits, and varied formations;
- Z5—Deep, uncontaminated alluvial deposits;
- Z6a—Alluvial and coastal formations of the Mediterranean shoreline;
- Z6b—Vulnerable lowland alluvial deposits.

Each groundwater body (GWB) was assigned a dominant lithology, defined as the lithological group covering the largest proportion of its surface area. This was determined in two ways. First, in the groundwater body reference system, a dominant aquifer lithology was assigned to each GWB. The information was then verified visually in a GIS environment by intersecting the GWB boundaries with the harmonized geological map of the basin. Where a well may intersect multiple aquifers, its assignment to a GWB follows the WFD dominant lithology, which may locally overlook secondary aquifer contributions. This classification aims to relate the geological characteristics of aquifers to the observed hydrochemical signatures, by grouping formations with similar geochemical behavior. The Z1–Z6b coding provides a synthetic nomenclature that facilitates the integration of lithological information into statistical and machine learning models, while remaining directly interpretable in hydrogeological terms.

2.4. Statistical Tools and Methodology

The analytical strategy adopted in this study is structured in several complementary steps, summarized in the flow diagram of Figure 2. This approach aims successively to: (i) reduce the dimensionality of the dataset, (ii) identify significant spatial structures, (iii) quantify the variance components at different nested scales, and (iv) assess the predictive power of hydrochemical parameters with respect to lithological context.

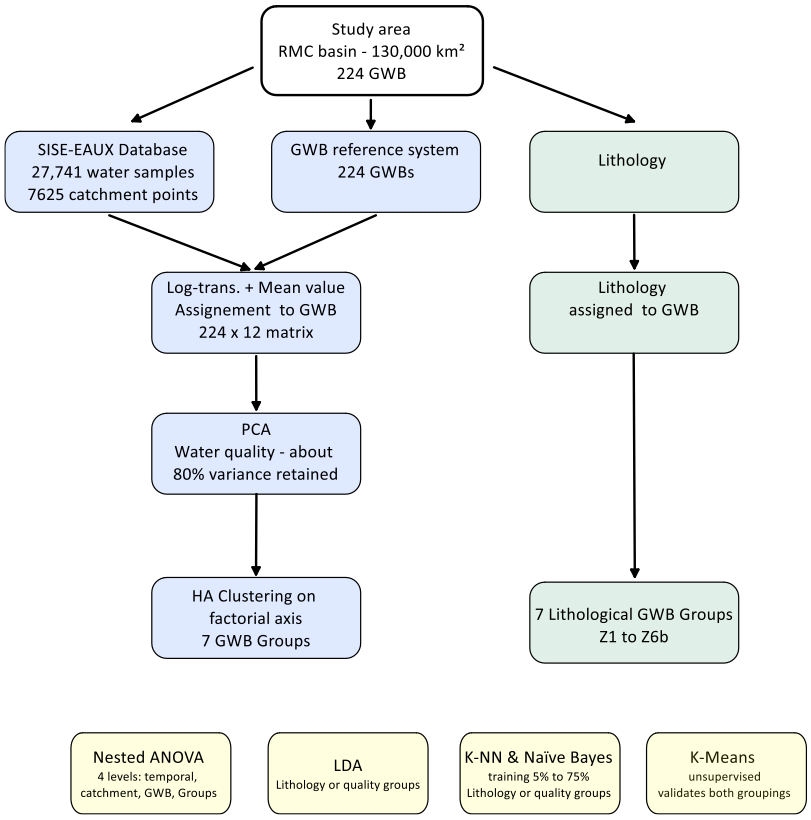

Figure 2. Flow diagram of the methodological procedure. Two independent classifications of the groundwater bodies (GWBs) are built in parallel: a geology-based lithological classification (Z1 to Z6b) and an empirical hydrochemical classification derived from PCA and HCA on the SISE-Eaux data. They are tested through nested ANOVA and supervised/unsupervised classification methods to characterize the spatio-temporal variability of groundwater quality in the RMC basin.

A Principal Component Analysis (PCA) was performed to reduce the dimensionality of the variable space and to identify the main sources of variability within the dataset [39]. The PCA was conducted using the correlation matrix, thus standardizing the variables and allowing the integration of parameters of different natures and units. By diagonalizing this matrix, the resulting principal components (PCs) are orthogonal to each other, enabling each factorial axis to be theoretically associated with independent processes controlling water quality. Only the factorial axes cumulating approximately 80% of the total inertia were retained for interpretation; the remaining axes, explaining a small fraction of the variance and considered statistical noise [40], were discarded from the analysis.

The PCA was applied to water quality data from the SISE-Eaux database. Previous studies conducted on regional extractions from the Rhône–Mediterranean–Corsica (RMC) basin or neighboring areas [41–43] have shown that parameter distributions tend toward normality after logarithmic transformation. Accordingly, a logarithmic transformation (log10) was applied to all parameters prior to PCA analysis, which brought the distributions of all parameters closer to normality and reduced the weight of extreme values during

processing [44]. At the sampling point scale, the distribution is generally multimodal, with each mode corresponding to a distinct site. Consequently, the subsequent part of the analysis (unsupervised clustering) was conducted after data aggregation by the mean of each factor axis for each GWB, which was used as the reference hydrological unit. Since GWBs are delineated in a harmonized manner at the European scale (WFD reference framework [45,46], the methodology consisted of calculating, for each parameter, a representative mean value at the GWB scale, derived from all sampling points it contains. This value was then assigned to the entire geographic polygon of the GWB. The resulting mapping thus exclusively reflected inter-GWB spatial variability, i.e., chemical quality differences between hydrogeological units, consistent with their individual characteristics and geological context [6].

An unsupervised Hierarchical Cluster Analysis (HCA) [47,48] was then applied to the contributions of the variables to the retained PCA factorial axes, weighted by their respective inertia (i.e., their contribution to the total variance of the dataset). This approach allows groundwater bodies (GWBs) to be grouped into homogeneous hydrochemical classes, based on their spectral similarities in the reduced space of the principal components. The HCA produces a dendrogram whose cut was determined according to an inter-class inertia jump criterion, enabling the definition of GWB groups sharing similar chemical signatures.

A nested analysis of variance (ANOVA) was applied to quantify the different components of variability of the studied parameters [49]. The model distinguishes four nested hierarchical levels: temporal variability (between sampling dates), variability within sampling points (catchment point), variability between groundwater bodies (GWB), and finally variability between GWB groups defined either by lithology (GWB Lithology Group) or by the hydrochemical classification derived from HCA (GWB HCA-PCA Group). Results are expressed as the η^2 coefficient (correlation ratio), corresponding to the fraction of total variance cumulatively explained at each spatial aggregation level [50]. This hierarchical variance decomposition allows identification of the spatial and temporal scales at which the processes controlling water quality are primarily structured [51].

Finally, linear discriminant analysis (LDA) as well as various supervised and unsupervised Machine Learning methods were implemented to assess the robustness of the obtained classifications and the ability of hydrochemical parameters to predict the membership of observations to the different GWB groups [52]. LDA seeks linear combinations of predictor variables that maximize separation between predefined classes while minimizing within-class variability [53,54]. It was applied using the 11 factorial axes from PCA as predictor variables and lithological group membership (Z1–Z6b) as the response variable. For this discriminant analysis, the filtering retains the first 100 rows (i.e., 100 observations), the covariance matrices are assumed to be equal, the prior probabilities are considered, and the significance level is set at 5%.

In addition, the k-Nearest Neighbors (k-NN) method—a non-parametric approach that makes no assumptions about the shape of class boundaries [55–57]—was tested with several training/validation split ratios. To this end, the waters of the whole dataset were ranked by increasing EC, and training subsets were built by selecting 1/20, 1/4, 1/2, and 3/4 of the samples along this ranking (5%, 25%, 50%, and 75% training data, respectively). The remaining samples formed the test set used to assess classification performance with respect to the GWB group. This EC-based selection ensures that each training subset covers the entire concentration range, allowing evaluation of performance sensitivity to training sample size and detection of potential nonlinear structures in the hydrochemical space. For this K-NN analysis, the number of neighbors is set to 4, the distance used is the Euclidean distance, and ties are handled by random choice.

The naïve Bayesian classifier [58–60], based on Bayes' theorem under the assumption of conditional independence of predictors, was also applied in external validation. The prior distributions of the parameters are assumed to be normal after logarithmic transformation of the data; ties are resolved by random choice; and an empirical Bayesian approach is adopted, with the prior distribution being estimated from the data.

Finally, the K-means method [61,62] was applied as an unsupervised validation, to verify whether natural structures emerge from the data without recourse to predefined class labels. All these methods were applied to the 11 factorial axes derived from PCA, ensuring predictor orthogonality and removal of statistical noise from the original variables.

3. Results

Table 1 shows the variance explained by the first five factor axes (PCs), which together account for more than 80%. These first five PCs were therefore the ones used for the unsupervised clustering of the GWBs.

Table 1. Eigenvalues and explained variance for the first 5 principal components resulting from the PCA.

	PC1	PC2	PC3	PC4	PC5
Eigenvalue	3.96	2.08	1.55	1.29	0.91
Variance (%)	33.01	17.36	12.90	10.76	7.59
Cumulative %	33.01	50.37	63.28	74.03	81.62

The contribution of the different parameters to these five PCs is detailed in Table 2. The first factor axis, which alone accounts for one third of the total information, contrasts old, mineralized groundwater with more recent rainwater affected by bacterial contamination. The second factor axis represents the bacterial contamination of nitrate-rich waters with higher Fe and Mn contents, and of a sodium-chloride facies characteristic of the downstream sectors of the study area.

Table 2. Contribution of the different parameters to the 5 first PCs.

Parameter	PC1	PC2	PC3	PC4	PC5
Enteroc.	−0.368	0.522	0.595	−0.332	0.084
E. coli	−0.346	0.514	0.565	−0.407	0.131
EC	0.871	0.010	−0.096	−0.376	−0.037
Na^+	0.503	0.593	0.137	0.510	−0.018
Ca^{2+}	0.779	−0.130	−0.134	−0.505	0.040
Mg^{2+}	0.734	0.040	0.215	−0.021	−0.434
Cl^-	0.557	0.645	−0.014	0.363	0.134
SO_4^{2-}	0.769	0.130	0.222	0.041	−0.273
HCO_3^-	0.546	−0.121	−0.014	−0.333	0.264
Fe	−0.367	0.543	−0.536	−0.236	−0.240
Mn	−0.315	0.561	−0.524	−0.241	−0.316
NO_3^-	0.339	0.345	−0.416	0.008	0.612

Figure 3 shows the hierarchical classification dendrogram derived from the main factorial axes of the PCA. The cut-off line allows the identification of 7 classes of GWBs.

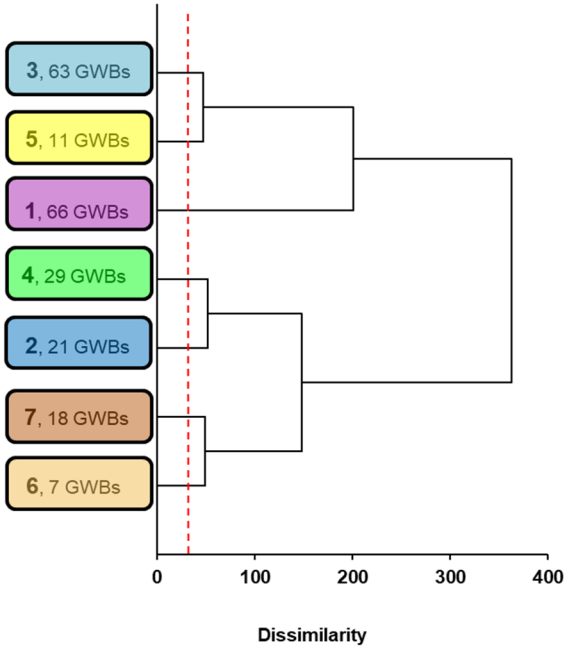

Figure 3. Dendrogram from the PCA + HCA processing, showing the distinction of 7 classes of GWBs, based on physicochemical and bacteriological compositions.

The distributions of the different classes of GWBs, that is, on the one hand, the classification from the dendrogram in Figure 3, and on the other hand the dominant lithology, are presented in Figure 4. Nine GWBs lacking sufficient data were excluded from the processing and appeared in white on the maps.

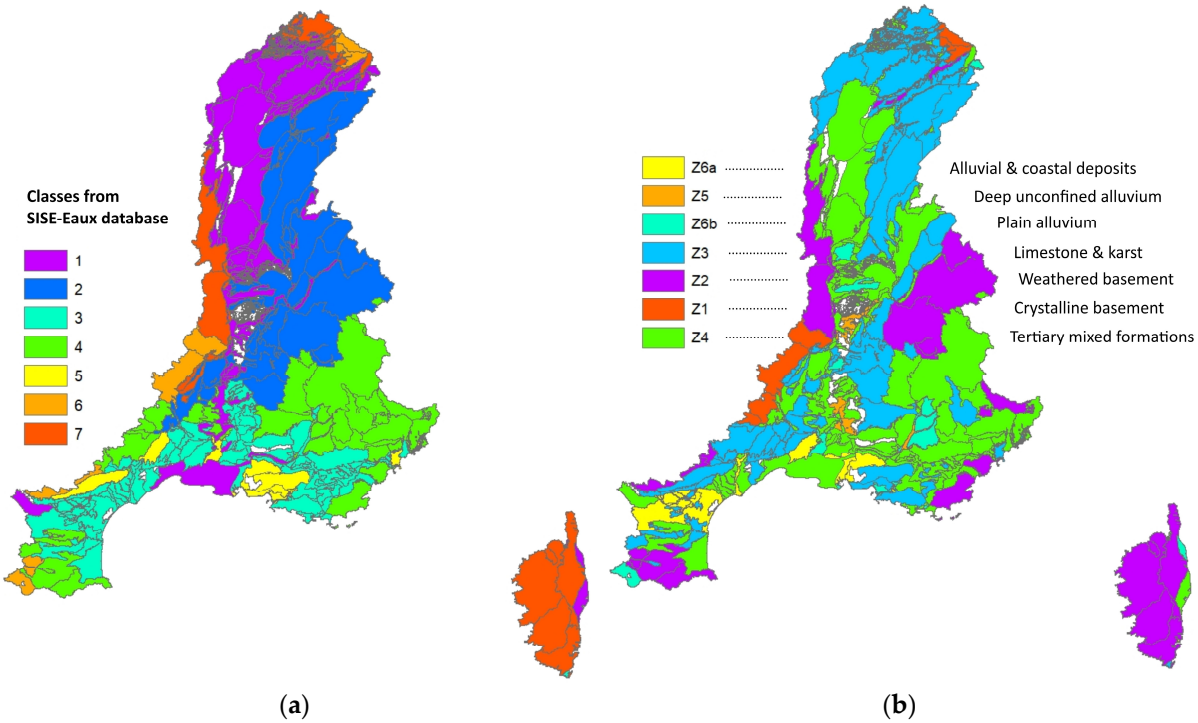

Figure 4. Distribution of groundwater body classes based on: (**a**) processing of the SISE-Eaux database by PCA followed by HCA, and (**b**) the dominant lithology of each of the 224 GWBs. The 9 GWBs shown in white were excluded from the processing due to insufficient data in the SISE-Eaux database.

3.1. Structure of the Spatio-Temporal Variability

The nested analysis of variance (ANOVA) enabled the total variability of each chemical parameter to be decomposed into four hierarchical components, corresponding to the spatial and temporal levels defined in the methodology (see Section 2.4). The results, summarized in Table 3, reveal a clear hierarchical organization, with markedly contrasting patterns depending on the nature of the parameters considered. They highlight that there is indeed a link between the lithological groups and the water quality groups, but this link is only partial, since the η^2 coefficient is far from 1.

Table 3. Proportions of explained variance (η^2) for each parameter, cumulated at each spatial aggregation level. The net variance attributable to each level is obtained by subtracting the cumulated variance of the lower level, thus allowing the specific contribution of each nested scale to be isolated.

Parameter	Temp. Variance	Catchment Point	GWB	Gr. GWB Lithology	Gr GWB HCA-PCA
Enter.	0.360	0.640	0.377	0.107	0.242
E. coli	0.334	0.666	0.382	0.110	0.193
E.C.	0.004	0.996	0.656	0.450	0.542
Ca^{2+}	0.029	0.971	0.779	0.505	0.631
HCO_3^-	0.034	0.966	0.694	0.221	0.349
Mg^{2+}	0.055	0.945	0.485	0.147	0.297
SO_4^{2-}	0.070	0.930	0.534	0.186	0.376
Na^+	0.063	0.937	0.568	0.126	0.390
Cl^-	0.064	0.936	0.618	0.116	0.448
Fe	0.308	0.692	0.681	0.213	0.258
Mn	0.216	0.784	0.695	0.113	0.441
NO_3^-	0.121	0.879	0.512	0.102	0.347

3.2. Contrast Between Microbiological and Physico-Chemical Parameters

Microbiological parameters (*E. coli* and enterococci) are characterized by a strong temporal component, accounting for 33.4% and 36.0% of the total variance, respectively—a proportion incomparable with that of physico-chemical parameters, whose temporal share is below 7%, and even drops to 0.4% for electrical conductivity. In contrast, the spatial structuring of these microbiological indicators remains very limited across all scales: the explained variance is low at the GWB scale ($\eta^2 = 0.377$ and 0.382), and aggregation by lithological groups provides only a marginal gain ($\eta^2 = 0.107$ and 0.110). These findings confirm that regional lithology exerts very little influence on the spatial distribution of microbiological contamination. This behavior reflects the episodic and localized transmission of these microorganisms into aquifers, directly linked to surface inputs and intense recharge events.

3.3. Contribution of Lithology to Spatial Structuring

Conversely, major physico-chemical parameters exhibit nearly zero temporal variability (0.3% to 7.0%) and a spatial structuring that progressively intensifies with changing scale, attesting to the temporal stability of geochemical processes (mineral dissolution, water–rock interactions, residence time). Calcium is an emblematic case ($\eta^2 = 0.971$ at the sampling-point scale; $\eta^2 = 0.505$ at the lithological-group level), a pattern consistent with calcite dissolution in extensive carbonate and karst systems, as reported in previous studies [42,43,63,64]. The increase in explained variance between the GWB scale and lithological groups demonstrates that lithology constitutes a regional control structure on groundwater composition. Comparing the two classifications (Lithology Group vs. HCA-PCA Group) shows that the empirical quality-based classification systematically explains

more variance, with the gap being particularly pronounced for manganese (ratio 3.9) and nitrates (ratio 3.4).

3.4. Intermediate Parameters: Iron, Manganese and Nitrates

Iron (30.8% temporal variability) and manganese (21.6%) reflect their redox sensitivity to aquifer oxygenation conditions. Nitrates exhibit a distribution co-determined by diffuse agricultural pressures and hydrogeological vulnerability: the low variance explained by lithological groups (0.102) compared to HCA-PCA groups (0.347) indicates that vulnerability to nitrogen inputs depends on fine-scale hydrodynamic parameters—porosity, permeability, connectivity with surface horizons—which lithology alone cannot capture.

3.5. Discrimination of Lithological Groups

Linear discriminant analysis (LDA) applied to the 7625 observations achieves an overall success rate of 56.2%, but this masks considerable disparities among lithological groups (Table 4). The pure crystalline basement (Z1) shows the highest recognition rate (76.0%), owing to its highly specific hydrochemical signature (very soft, bicarbonate-rich waters, with virtually no soluble minerals). The weathered basement (Z2; 64.0%) and Tertiary molasses (Z4; 65.2%) are also well discriminated. Limestones and karsts (Z3; 50.2%) are subject to substantial confusion with Z4 (702 misclassified observations), due to the similarity of calcium-bicarbonate facies in both formations. In contrast, deep alluvium (Z5; 0%) and vulnerable lowland alluvium (Z6b; 0%) are not discriminated at all, as their chemical signatures are driven more by anthropogenic pressures and exchanges with the surrounding Tertiary formations than by lithology stricto sensu.

Table 4. Confusion matrix of linear discriminant analysis. Overall correct classification rate in estimation: 56.2%. Diagonal values (in bold) correspond to correctly classified observations.

Obs.\Pred.	Z1	Z2	Z3	Z4	Z5	Z6a	Z6b	Total	% Correct Within lith. Class	Precision per Class %
Z1	**222**	57	6	6	0	1	0	292	76.0	45.87
Z2	181	**1328**	102	459	0	4	0	2074	64.0	69.31
Z3	35	207	**974**	702	1	20	0	1939	50.2	59.64
Z4	28	268	491	**1676**	0	105	1	2569	65.2	50.07
Z5	0	13	7	145	**0**	19	0	184	0	0
Z6a	0	25	18	222	0	**89**	0	354	25.1	36.63
Z6b	18	18	35	137	0	5	**0**	213	0	0
TOTAL	484	1916	1633	3347	1	243	1	7625	56.25	56.25

3.6. Validation Using Machine Learning Methods

The k-nearest neighbors (k-NN) classifier substantially outperforms LDA with sufficient training (76.2% vs. 56.2% at 75% training), confirming the presence of nonlinear structures in the hydrochemical space that linear methods cannot capture. The steady increase in classification rate with training sample size (from 58.3% to 76.2% for lithology) reflects a high within-class variability that can only be properly learned with enough examples. The naïve Bayesian classifier is suitable for water-quality classification (~65% in validation, stable regardless of training ratio) but is entirely unsuitable for lithological prediction (~11% in validation, Table 5), due to the strong multicollinearity among hydrochemical parameters, which violates the conditional independence assumption. Finally, unsupervised K-means classification confirms the reality of the lithological signal within the data: the centroids of the resulting classes show a structure consistent with the lithological groups defined a priori, with the class exhibiting the most negative values

on the first factorial axis systematically corresponding to the oligomineral waters of the crystalline basement.

Table 5. Comparative performances of Machine Learning methods in external validation according to training ratio. Rates for the Naïve Bayes classifier for lithology reflect the violation of the independence assumption of correlated predictors.

Method	Target	Training 5%	Training 25%	Training 50%	Training 75%	Note
k-NN	Water quality (7 groups)	69.0%	74.4%	74.4%	76.2%	Validation
k-NN	Lithology (Z1–Z6b)	58.3%	63.7%	66.2%	76.2%	Validation
Naïve Bayes classifier	Water quality	65.5%	~65%	64.4%	~65%	Validation
Naïve Bayes classifier	Lithology	~11%	~11%	~11%	~11%	Validation
K-means	Unsupervised	–	–	–	–	Qualitative Validation

4. Discussion

4.1. Lithological Control on Physico-Chemical Quality

Pure crystalline basement aquifers (Z1) yield the softest and least mineralized waters in the basin (conductivity generally below 150 µS/cm, with calcium or sodium bicarbonate chemical profile). Microbiological contamination remains moderate in direct basement abstraction areas, but can reach high levels in fractured zones subject to extensive livestock farming and thin soil cover. Karst aquifers (Z3) exhibit high microbiological vulnerability: rapid conduits transmit surface water directly without effective filtration, generating *E. coli* and enterococci contamination peaks following intense rainfall events. Tertiary and molasse formations (Z4), the most represented group (33.7% of observations), are characterized by often elevated SO_4^{2-}, Cl^- and NO_3^- concentrations, reflecting agricultural pressures in the plains of the Rhône corridor and Languedoc.

The weathered basement (Z2), a transitional domain comprising metamorphic formations and folded massifs, displays a hydrochemical signature intermediate between that of the pure crystalline basement and the surrounding sedimentary formations. This intermediate position, consistent with its moderate LDA discrimination rate (64.0%), reflects greater substrate alteration and longer residence times, promoting a gradual enrichment in major elements without reaching the mineralization levels of carbonate or molasse formations. Coastal formations of the Mediterranean shoreline (Z6a) are distinguished by a signature marked by saline inputs, particularly chloride and sodium contents, resulting from salt-water intrusions and aeolian deposits characteristic of the coastal setting. However, their discrimination remains imperfect (25.1%), with most confusion occurring with molasse formations (Z4), with which they share a common Tertiary sedimentary origin.

The failure to discriminate deep alluvium (Z5) and vulnerable lowland alluvium (Z6b)—entirely absorbed into Z4 in the LDA—confirms that the quality of these aquifers is dictated more by exchanges with the surrounding Tertiary formations and by agricultural inputs than by lithology per se. This result has important practical implications: these abstraction points, which cannot be distinguished on chemical signature alone, require enhanced monitoring of nitrates and microbiological parameters, regardless of their geological context.

These findings converge with international observations on the role of lithology as a primary factor in groundwater physico-chemical composition [14,63,65–67], as well as with studies conducted in other large river basins, such as the Upper Rhine hydrosystem, where

major-element variations are primarily correlated with watershed lithology rather than atmospheric inputs or anthropogenic pressures [68,69]. This lithological control is also consistent with the ranking of solute flux and chemical weathering rates established for monolithologic catchments in France, which places carbonate rocks and evaporite-bearing formations at the top of the reactivity gradient [70]. Our discrimination hierarchy—crystalline basement > weathered basement ≈ molasse > karst > alluvium—reflects an increasing gradient of geochemical reactivity and hydraulic connectivity with adjacent formations, a pattern documented beyond the RMC basin and thus reinforcing the generality of the lithological signal [68–70]. Conversely, the weak spatial structuring of fecal contamination by regional lithology, already highlighted by ANOVA (Table 3, variance share below 11% for lithological groups), confirms that this parameter responds more to local determinants—livestock density, distance to contamination sources, degree of catchment protection—than to the nature of the reservoir rock itself [67,71]. This dissociation is corroborated by independent studies in humid temperate settings, which demonstrate that microbial variability in wells is primarily driven by transient hydrologic events (rainfall, snowmelt) and proximity to surface-water bodies, rather than by aquifer lithology [72]. Similarly, research on river–groundwater interaction in alluvial aquifers of Northwest Switzerland has shown that bacterial cell counts and their temporal variability decrease sharply with distance from the river, irrespective of the lithological nature of the aquifer material [73,74].

This dissociation between the drivers of chemical quality and those of microbiological quality constitutes one of the key findings of this study and, from an operational standpoint, justifies a dual framework for water-body management: a geological approach for chemical parameters, and an approach based on land use and agropastoral practices for bacteriological parameters.

In conclusion, at the scale of the RMC basin, the results confirm the existing link between lithology and water quality. However, this link is only partial at this scale, even though this relationship had already been mentioned at the scale of the administrative regions Provence-Alpes-Côte d'Azur, Auvergne-Rhône-Alpes, Occitanie and Bourgogne-Franche-Comté [42,44,63,67]. The larger the study area considered, the more the link between lithology and water quality weakens. Compared with previous studies conducted at finer scales, other factors of variability come into play, in particular a wider climatic range, but also lithological and altitudinal factors, and a greater diversity in terms of land use.

4.2. Methodological Implications and Relevance for Regionalisation

The systematic comparison of methods allows several recommendations to be formulated for similar large-scale studies. LDA provides an excellent starting point due to its interpretability, but its parametric assumptions (multivariate normality, homogeneity of within-class covariance matrices) limit its performance for hydrochemically similar groups. k-NN, the best-performing method in this dataset, demonstrates the existence of nonlinear structures in the hydrochemical space of the RMC basin. Prior PCA transformation proves essential for all methods, as it removes noise and orthogonalizes predictors. Finally, the comparison between lithological classification and empirical hydrochemical classification argues for a complementary use of both approaches in monitoring strategies: lithology as a structuring geological reference framework, and hydrochemical classification as a tool for identifying aquifers most exposed to diffuse contamination or natural geochemical anomalies. This recommendation aligns with the broader trend in groundwater quality assessment, where classical discriminant approaches are increasingly combined with machine-learning classifiers to handle non-linear structures and class imbalance [75,76]. However, our results also caution against over-optimistic performance expectations: while recent SVM-based frameworks report classification accuracies exceeding 93% [77], such figures are typically

obtained on datasets with more balanced class distributions or with feature sets specifically optimized for discrimination, and are therefore not directly transposable to large-scale, imbalanced lithological datasets such as ours.

4.3. Robustness of Lithological Classification: LDA and Machine Learning

The contrast between LDA and k-NN deserves further examination. LDA implicitly assumes that the point clouds associated with each lithological group follow a multivariate Gaussian distribution with a common covariance matrix. However, groups Z3 and Z4 in particular exhibit substantial internal dispersion and likely heterogeneous covariance matrices, which explains the massive confusion observed between these two classes (Table 4). k-NN, by relaxing this assumption and directly exploiting the local structure of the point cloud in the factor space, captures nonlinear decision boundaries that LDA can only imperfectly approximate through hyperplanes. The steady improvement in its performance with increasing training sample size (from 58.3% to 76.2% for lithology) nevertheless indicates that this method remains sensitive to local observation density, warranting caution for underrepresented lithological groups in the dataset, such as Z5 (184 observations, i.e., 2.4% of the total). The relevance of the EC-based ranking used to build the training and validation subsets is confirmed posteriori by the strong contribution of electrical conductivity to the first PCA axis: since the waters were ranked by increasing EC, this selection strategy ensures that each training subset spans the full range of the dominant gradient of hydrochemical variability, thereby justifying the choice of EC as the ordering criterion for the learning and validation subsets.

Unsupervised validation via K-means provides important complementary insight: the consistency between the class centroids obtained without prior labeling and the lithological groups defined a priori demonstrates that the observed structuring is not an artifact of supervised classification, but indeed reflects a real and robust geochemical signal. Conversely, the near-total failure of the naïve Bayesian classifier for lithological prediction (~11% in validation, close to chance for seven classes) illustrates the limits of an approach relying on conditional independence of predictors in a system where major elements are strongly correlated by mineral dissolution equilibria and solution electroneutrality. This seemingly negative result has diagnostic value: it confirms that the covariance structure among hydrochemical parameters itself carries significant geological information, which methods ignoring this covariance cannot exploit. This observation is consistent with methodological reviews emphasizing that the conditional independence assumption of naïve Bayes is rarely met in hydrochemical datasets, where mineral dissolution equilibria and electroneutrality generate strong inter-parameter correlations [75]. It also reinforces the case for k-NN, whose local-structure approach proved superior in our dataset (76.2% vs. 56.2% for LDA), a finding echoed by comparative studies where k-NN performs competitively for hydrochemical classification tasks when the feature space is appropriately structured [77,78].

4.4. Study Limitations and Research Perspectives

Several limitations warrant mention. First, the sampling network of the SISE-Eaux database is not a stratified random sample but results from the historical location of drinking-water supply intakes, with an overrepresentation of springs in mountain areas and wells or boreholes in lowland areas. This network heterogeneity, although partially compensated by aggregation at the GWB scale, may introduce bias in the characterization of certain lithological groups poorly covered by the sanitary monitoring network, particularly formations in low-population-density areas (Z1, Z5). Although most boreholes in the dataset are documented as tapping a single aquifer and are systematically cemented at the top, some wells may potentially intersect multiple aquifers; in such cases, their assignment to a GWB

follows the dominant lithology defined under the WFD, which may locally overlook the contribution of secondary aquifers.

Second, the 34-year observation period (1987–2021) is not covered uniformly: sampling density has evolved over time with the progressive strengthening of sanitary controls, which may affect the representativeness of the temporal component in the ANOVA for the early years of the series. A sub-period analysis, currently beyond the scope of this paper, would allow verification of the stability of variance components over shorter time windows and identification of possible long-term trends related to land-use changes or climatic shifts. In addition, analytical protocols have evolved over the 34-year period, with improvements and lower detection limits. However, these changes remain minor relative to the temporal variability of water quality. Sampling techniques have not evolved over the period, nor geographically across the study area: they are standardized and have remained unchanged, and all samplers were trained by the same company, so operator-dependent variability is also low relative to the temporal variability of water quality.

Third, the logarithmic transformation applied to all parameters, although justified by previous work on regional extractions from the same database [42,44,63], relies on the assumption of an approximately log-normal distribution at the GWB scale. While reasonable for major elements, this assumption is more questionable for microbiological indicators, whose distribution remains highly skewed even after transformation, due to the episodic nature of fecal contamination. It should also be noted that SO_4^{2-} and NO_3^- are redox-sensitive species, in addition to Fe and Mn. In our study area, the dominant source of sulfate is related to Triassic evaporitic gypsum deposits, rather than to pyrite oxidation. Pyrite oxidation does play a role in the Alpine secondary formations (the so-called "terres noires", containing gypsum and some pyrite, abundant throughout the Durance basin), but this process is secondary in our study area compared with the evaporitic Triassic source.

Fourth, unlike the previous study [6], which integrated CORINE Land Cover land-use data, the present study focuses solely on lithological control and does not explicitly incorporate anthropogenic pressures (livestock density, irrigated areas, sanitation). The explicit coupling of lithological classification with land-use and agropastoral-pressure variables represents a natural avenue for refining the understanding of faecal contamination mechanisms and their sensitivity to intense recharge events in a context of climate change. Finally, the proposed methodology, designed to be transferable, would benefit from application to other major European river basins with comparable geological diversity, in order to test the generality of the lithology–quality relationships highlighted in the RMC basin.

A further limitation concerns the unequal data support across lithological groups. Some groups (e.g., Z5, Z6b) are represented by a limited number of observations, which reduces the statistical power of the ANOVA and the reliability of LDA/k-NN performance metrics for these classes. Consequently, part of the observed variability—and of the contrasts between groups—may reflect differences in sampling density rather than solely geochemical or hydrogeological processes. By contrast, because only complete analyses were retained (full matrix), all parameters share the same measurement frequency, so comparisons of variance components between physico-chemical and microbiological indicators are not affected by differential data support.

5. Conclusions

This study makes several original scientific contributions to the understanding of groundwater quality in the RMC basin.

First, it establishes lithology as a fundamental determinant of regional-scale groundwater physico-chemical quality. ANOVA demonstrates that lithological grouping explains up to 20% of additional variance for physico-chemical parameters—a quantitative valida-

tion of a relationship that had hitherto been asserted mainly on a qualitative basis in the hydrogeological literature.

Second, the pure crystalline basement (Z1) stands out with a major qualitative difference, achieving a 76% discrimination rate in LDA and displaying a hydrochemical signature that is perfectly distinct from all other geological contexts.

Third, molasse and alluvial sedimentary formations are much less clearly differentiated from one another, reflecting the strong influence of anthropogenic pressures and hydrodynamic exchanges between formations.

Fourth, microbiological variability is governed by factors other than lithology—episodic hydrological events and local pastoral practices—underscoring the need for specific approaches to bacteriological risk management, particularly in karst systems.

These findings have direct implications for groundwater body management under the River Basin Management Plan (Schéma Directeur d'Aménagement et de Gestion des Eaux, SDAGE RMC 2022–2027): definition of quality standards differentiated by lithological context, enhanced monitoring of vulnerable alluvial abstractions for nitrates and microbiological parameters, and implementation of extended protection perimeters in karst systems.

Several limitations nevertheless qualify the scope of these conclusions. The SISE-Eaux sampling network is not a stratified random sample but reflects the historical location of drinking-water intakes, so lithological groups poorly covered by sanitary monitoring—particularly in low-population-density areas (Z1, Z5)—remain less robustly characterized. The 34-year observation period (1987–2021) is also unevenly covered, and a sub-period analysis would be needed to confirm the stability of variance components and to detect long-term trends related to land-use or climatic change. In addition, the data support is unequal across lithological groups: some groups (e.g., Z5, Z6b) are represented by a limited number of observations, which reduces the statistical power of the ANOVA and the reliability of LDA/k-NN performance metrics for these classes, so part of the observed variability may reflect differences in sampling density rather than solely geochemical or hydrogeological processes. By contrast, because only complete analyses were retained (full matrix), all parameters share the same measurement frequency, so comparisons between physicochemical and microbiological indicators are not affected by differential data support. The log-normal assumption underlying the logarithmic transformation is reasonable for major elements but more questionable for microbiological indicators, whose distribution remains highly skewed because of the episodic nature of fecal contamination. Finally, the present study focuses on lithological control and does not explicitly incorporate anthropogenic pressures (livestock density, irrigated areas, sanitation); coupling lithological classification with land-use and agropastoral-pressure variables therefore represents a natural avenue for refining the understanding of fecal contamination mechanisms, particularly under intense recharge events in a climate-change context.

Beyond these limitations, the proposed methodology—combining PCA, HCA, nested ANOVA, LDA and Machine Learning methods—provides an analytical framework that can be transferred to other major European river basins with comparable geological diversity, in order to test the generality of the lithology–quality relationships highlighted here.

Author Contributions: Conceptualization, V.V., A.B. (Abdelhak Bouabdli) and M.S.; methodology, V.V., L.B. and Z.Z.; software, Z.Z.; validation, A.B. (Abderrahim Bousouis), I.M., M.E.J., A.A.B. and M.T.; formal analysis, Z.Z.; investigation, Z.Z., V.V. and L.B.; resources, L.B., M.S., A.B. (Abderrahim Bousouis), and I.M.; data curation, I.M., M.T., A.A.B. and M.E.J.; writing—original draft preparation, V.V. and Z.Z.; writing—review and editing, L.B.; visualization, Z.Z., V.V. and L.B.; supervision, V.V. and M.S.; project administration, A.B. (Abdelhak Bouabdli); funding acquisition, L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are only available on request from the Health Agency ARS of the Provence-Alpes-Côte d'Azur, Occitanie, Auvergne-Rhône-Alpes and Bourgogne-Franche-Comté regions due to sensitivity to possible malicious acts.

Acknowledgments: The authors thank the RMC Basin Agency for its administrative and technical support. We are grateful to the regional Health Agencies of Provence-Alpes-Côte d'Azur, Occitanie, Auvergne-Rhône-Alpes and Bourgogne-Franche-Comté for extracting the data from the Size-Eaux database.

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

Abbreviations

The following abbreviations are used in this manuscript:

RMC	Rhône–Mediterranean–Corsica (basin)
GWB	Groundwater body
WFD	Water Framework Directive (2000/60/EC)
SDAGE	Schéma Directeur d'Aménagement et de Gestion des Eaux (River Basin Management Plan)
PCA	Principal Component Analysis
HCA	Hierarchical Cluster Analysis
ANOVA	Analysis of Variance
LDA	Linear Discriminant Analysis
k-NN	k-Nearest Neighbours
ARS	Agence Régionale de Santé (Regional Health Agency)

References

1. Katsanou, K.; Karapanagioti, H.K. Surface Water and Groundwater Sources for Drinking Water. In *Applications of Advanced Oxidation Processes (AOPs) in Drinking Water Treatment*; Gil, A., Galeano, L.A., Vicente, M.Á., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 1–19, ISBN 978-3-319-76882-3.
2. Carrard, N.; Foster, T.; Willetts, J. Groundwater as a Source of Drinking Water in Southeast Asia and the Pacific: A Multi-Country Review of Current Reliance and Resource Concerns. *Water* **2019**, *11*, 1605. Correction in *Water* **2020**, *12*, 298. https://doi.org/10.3390/w12010298. [CrossRef]
3. Maréchal, J.-C.; Rouillard, J. Groundwater in France: Resources, Use and Management Issues. In *Sustainable Groundwater Management: A Comparative Analysis of French and Australian Policies and Implications to Other Countries*; Rinaudo, J.-D., Holley, C., Barnett, S., Montginoul, M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 17–45, ISBN 978-3-030-32766-8.
4. Khatri, N.; Tyagi, S. Influences of Natural and Anthropogenic Factors on Surface and Groundwater Quality in Rural and Urban Areas. *Front. Life Sci.* **2015**, *8*, 23–39. [CrossRef]
5. Abanyie, S.K.; Apea, O.B.; Abagale, S.A.; Amuah, E.E.Y.; Sunkari, E.D. Sources and Factors Influencing Groundwater Quality and Associated Health Implications: A Review. *Emerg. Contam.* **2023**, *9*, 100207. [CrossRef]
6. Zeiki, Z.; Mohsine, I.; Bousouis, A.; El Jirari, M.; Touzani, M.; Bouabdli, A.; Sadiki, M.; Valles, V.; Barbiero, L. Deciphering Groundwater Quality Mechanisms in the Rhône-Mediterranean-Corsica Basin (RMC) Through Multi-Source Data Integration. *Water* **2026**, *18*, 1228. [CrossRef]
7. Appelo, C.A.J.; Postma, D. *Geochemistry, Groundwater and Pollution*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2005.
8. Edmunds, W.M.; Shand, P.; Hart, P.; Ward, R.S. The Natural (Baseline) Quality of Groundwater: A UK Pilot Study. *Sci. Total Environ.* **2003**, *310*, 25–35. [CrossRef] [PubMed]
9. Burch, T.; Stokdyk, J.; Durso, L.; Borchardt, M. Quantitative Microbial Risk Assessment for Ingestion of Antibiotic Resistance Genes from Private Wells Contaminated by Human and Livestock Fecal Sources. *Appl. Environ. Microbiol.* **2024**, *90*, e01629-23. [CrossRef] [PubMed]
10. Delpla, I.; Jung, A.-V.; Baures, E.; Clement, M.; Thomas, O. Impacts of Climate Change on Surface Water Quality in Relation to Drinking Water Production. *Environ. Int.* **2009**, *35*, 1225–1233. [CrossRef] [PubMed]
11. Rivett, M.O.; Buss, S.R.; Morgan, P.; Smith, J.W.N.; Bemment, C.D. Nitrate Attenuation in Groundwater: A Review of Biogeochemical Controlling Processes. *Water Res.* **2008**, *42*, 4215–4232. [CrossRef] [PubMed]
12. European Commission Council Directive. Concerning the Protection of Waters against Pollution Caused by Nitrates from Agricultural Sources. *Off. J.* **1991**, *375*, 1–8.

13. Hem, J.; Survey, U.S.G. *Study and Interpretation of the Chemical Characteristics of Natural Water*, 3rd ed.; US Government Printing Office: Washington, DC, USA, 1985.
14. Okyere, E.; Podgorski, J.; Berg, M.; Abass, G.; Preko, L.; Ganyaglo, S. Groundwater Quality and Fluoride Enrichment: Hydrochemical and Lithological Controls in a Crystalline Basement Aquifer of Northern Ghana. *Water* **2026**, *18*, 2153. [CrossRef]
15. Rheaume, S.; Survey, U.S.G. *Geohydrology and Water Quality of Kalamazoo County, Michigan, 1986–1988*; US Geological Survey: Reston, VA, USA, 1990.
16. Kozar, M.; McCoy, K.; Weary, D.; Field, M.; Pierce, H.A.; Schill, W.; Young, J.; Survey, U.S.G. *Hydrogeology and Water Quality of the Leetown Area, West Virginia*; US Geological Survey: Reston, VA, USA, 2008.
17. Page, R.; LeBlanc, R.A.; Survey, U.S.G. *Geology, Hydrology, and Water Quality in the Fresno Area, California*; US Geological Survey: Reston, VA, USA, 1969.
18. Gowd, S.S. Assessment of Groundwater Quality for Drinking and Irrigation Purposes: A Case Study of Peddavanka Watershed, Anantapur District, Andhra Pradesh, India. *Environ. Geol.* **2005**, *48*, 702–712. [CrossRef]
19. European Commission. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000. Establishing a Framework for Community Action in the Field of Water Policy. *Off. J. Eur. Commun.* **2000**, *22*, 2000.
20. BRGM. *Référentiel des Masses d'eau Souterraine de France Métropolitaine BRGM RP-52118-FR*; Bureau de Recherches Géologiques et Minières: Orléans, France, 2003.
21. Thabit, F.; Gouissem, A.; Rangarajan, R.; Esmaeili, A.; Benlamri, R. Artificial Intelligence in Groundwater and Water Quality Assessment: A Survey. In Proceedings of the 2026 International Wireless Communications and Mobile Computing (IWCMC), Wuzhou, China, 1–6 June 2026; pp. 1370–1378. [CrossRef]
22. Huang, Y.; Wang, C.; Wang, Y.; Lyu, G.; Lin, S.; Liu, W.; Niu, H.; Hu, Q. Application of machine learning models in groundwater quality assessment and prediction: Progress and challenges. *Front. Environ. Sci. Eng.* **2024**, *18*, 23. [CrossRef]
23. Haggerty, R.; Sun, J.; Yu, H.; Li, Y. Application of machine learning in groundwater quality modeling—A comprehensive review. *Water Res.* **2023**, *233*, 119745. [CrossRef] [PubMed]
24. Zhu, M.; Wang, J.; Yang, X.; Zhang, Y.; Zhang, L.; Ren, H.; Wu, B.; Ye, L. A review of the application of machine learning in water quality evaluation. *Eco-Environ. Health* **2022**, *1*, 107–116. [CrossRef] [PubMed]
25. Naz, I.; Ahmad, I.; Aslam, R.W.; Quddoos, A.; Yaseen, A. Integrated Assessment and Geostatistical Evaluation of Groundwater Quality through Water Quality Indices. *Water* **2024**, *16*, 63. [CrossRef]
26. Thivya, C.; Chidambaram, S.; Keesari, T.; Prasanna, M.V.; Thilagavathi, R.; Adithya, V.S.; Singaraja, C. Lithological and hydrochemical controls on distribution and speciation of uranium in groundwaters of hard-rock granitic aquifers of Madurai District, Tamil Nadu (India). *Environ. Geochem. Health* **2016**, *38*, 497–509. [CrossRef] [PubMed]
27. Sarath, K.V.; Shaji, E.; Adarshkrishna, K.R.; Muthukumar, P.; Renjith, P.N.; Nandakumar, V. Hydrogeochemical assessment and multivariate statistical analyses for the characterization of trace metals in groundwater of Chandragiri River Basin, Kerala, southwestern India. *Phys. Chem. Earth* **2026**, *143*, 104397. [CrossRef]
28. Keesstra, S.D.; Geissen, V.; Mosse, K.; Piiranen, S.; Scudiero, E.; Leistra, M.; van Schaik, L. Soil as a filter for groundwater quality. *Curr. Opin. Environ. Sustain.* **2012**, *4*, 507–516. [CrossRef]
29. Krasovskaia, I.; Gottschalk, L.; Leblois, E. Signature d'un Climat en Évolution dans les Régimes Hydrologiques de la Région Rhône-Méditerranée-Corse. *Houille Blanche* **2002**, *88*, 25–30. [CrossRef]
30. Henocque, Y.; Andral, B. The French Approach to Managing Water Resources in the Mediterranean and the New European Water Framework Directive. *Mar. Pollut. Bull.* **2003**, *47*, 155–161. [CrossRef] [PubMed]
31. Botrel, I.; Mercier, L.; Leger, Y.; Dugleux, E. Restaurer la Qualité de l'eau: Les Outils Fonciers. L'accompagnement des collectivités par les Safer et l'agence de l'eau Rhône-Méditerranée et Corse. *Tech. Sci. Méthodes* **2009**, *12*, 77–84. [CrossRef]
32. Mocochain, L.; Clauzon, G.; Bigot, J.-Y.; Brunet, P. Geodynamic Evolution of the Peri-Mediterranean Karst during the Messinian and the Pliocene: Evidence from the Ardèche and Rhône Valley Systems Canyons, Southern France. *Sediment. Geol.* **2006**, *188–189*, 219–233. [CrossRef]
33. Bernon, N.; Chartier, R.; Herniot, P. *Atlas du Potentiel Très Basse Énergie des Aquifères de la Région Rhône-Alpes. Rapport Final, BRGM RP-60856-FR*; Bureau de Recherches Géologiques et Minières: Orléans, France, 2012.
34. Sassine, L.; Khaska, M.; Ressouche, S.; Simler, R.; Lancelot, J.; Verdoux, P.; Le Gal La Salle, C. Coupling Geochemical Tracers and Pesticides to Determine Recharge Origins of a Shallow Alluvial Aquifer: Case Study of the Vistrenque Hydrogeosystem (SE France). *Appl. Geochem.* **2015**, *56*, 11–22. [CrossRef]
35. Chery, L.; Laurent, A.; Vincent, B.; Tracol, R. *Echanges SISE-Eaux/ADES: Identification des Protocoles Compatibles avec les Scénarios d'Echange SANDRE*; ONEMA: Vincennes, France; Orléans, France, 2011.
36. *ISO/IEC 17025:2017*; General Requirements for the Competence of Testing and Calibration Laboratories. International Organization for Standardization: Vienna, Austria, 2017.
37. Pachepsky, Y.A.; Shelton, D.R. *Escherichia coli* and Fecal Coliforms in Freshwater and Estuarine Sediments. *Crit. Rev. Environ. Sci. Technol.* **2011**, *41*, 1067–1110. [CrossRef]

38. Abbas, A.; Baek, S.; Silvera, N.; Soulileuth, B.; Pachepsky, Y.; Ribolzi, O.; Boithias, L.; Cho, K.H. In-Stream Escherichia Coli Modeling Using High-Temporal-Resolution Data with Deep Learning and Process-Based Models. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 6185–6202. [CrossRef]
39. Abou Zakhem, B.; Al-Charideh, A.; Kattaa, B. Using Principal Component Analysis in the Investigation of Groundwater Hydrochemistry of Upper Jezireh Basin, Syria. *Hydrol. Sci. J.* **2017**, *62*, 2266–2279. [CrossRef]
40. Helena, B.; Pardo, R.; Vega, M.; Barrado, E.; Fernandez, J.M.; Fernandez, L. Temporal Evolution of Groundwater Composition in an Alluvial Aquifer (Pisuerga River, Spain) by Principal Component Analysis. *Water Res.* **2000**, *34*, 807–816. [CrossRef]
41. Lazar, H.; Ayach, M.; Bousouis, A.; Huneau, F.; Mori, C.; Garel, E.; Kacimi, I.; Valles, V.; Barbiero, L. Multivariate and Spatial Study and Monitoring Strategies of Groundwater Quality for Human Consumption in Corsica. *Hydrology* **2024**, *11*, 197. [CrossRef]
42. Mohsine, I.; Kacimi, I.; Valles, V.; Leblanc, M.; El Mahrad, B.; Dassonville, F.; Kassou, N.; Bouramtane, T.; Abraham, S.; Touiouine, A.; et al. Differentiation of Multi-Parametric Groups of Groundwater Bodies through Discriminant Analysis and Machine Learning. *Hydrology* **2023**, *10*, 230. [CrossRef]
43. Ayach, M.; Lazar, H.; Lamat, C.; Bousouis, A.; Touzani, M.; El Jarjini, Y.; Kacimi, I.; Valles, V.; Barbiero, L.; Morarech, M. Groundwaters in the Auvergne-Rhône-Alpes Region, France: Grouping Homogeneous Groundwater Bodies for Optimized Monitoring and Protection. *Water* **2024**, *16*, 869. [CrossRef]
44. Jabrane, M.; Touiouine, A.; Bouabdli, A.; Chakiri, S.; Mohsine, I.; Valles, V.; Barbiero, L. Data Conditioning Modes for the Study of Groundwater Resource Quality Using a Large Physico-Chemical and Bacteriological Database, Occitanie Region, France. *Water* **2023**, *15*, 84. [CrossRef]
45. European Commission. *Guidance Document No. 22. Guidance on Implementing the Geographical Information System (GIS) Elements of the EU Water Policy. Tools and Services for Reporting Under RBMP Within WISE. Guidance on Reporting of Spatial Data for the WFD (RBMP)*; Portuguese Environment Agency: Amadora, Portugal, 2009.
46. European Commission. *Guidance on Groundwater Status and Trend Assessment, Guidance Document #18*; European Communities: Luxembourg, 2009; ISBN 978-92-79-11374-1.
47. Bouguettaya, A.; Yu, Q.; Liu, X.; Zhou, X.; Song, A. Efficient Agglomerative Hierarchical Clustering. *Expert Syst. Appl.* **2015**, *42*, 2785–2797. [CrossRef]
48. Madhulatha, T.S. An Overview on Clustering Methods. *IOSR J. Eng.* **2012**, *2*, 719–725. [CrossRef]
49. Miles, J.R. Squared, Adjusted R Squared. In *Wiley StatsRef: Statistics Reference Online*; Wiley: Hoboken, NJ, USA, 2014; ISBN 9781118445112.
50. Stahle, L.; Wold, S. Analysis of Variance (ANOVA). *Chemom. Intell. Lab. Syst.* **1989**, *6*, 259–272. [CrossRef]
51. Bousouis, A.; Bouabdli, A.; Ayach, M.; Lazar, H.; Ravung, L.; Valles, V.; Barbiero, L. Discrimination of Spatial and Temporal Variabilities in the Analysis of Groundwater Databases: Application to the Bourgogne-Franche-Comté Region, France. *Water* **2025**, *17*, 384. [CrossRef]
52. Huberty, C.J. Discriminant Analysis. *Rev. Educ. Res.* **1975**, *45*, 543–598. [CrossRef]
53. Wilson, S.R.; Close, M.E.; Abraham, P. Applying Linear Discriminant Analysis to Predict Groundwater Redox Conditions Conducive to Denitrification. *J. Hydrol.* **2018**, *556*, 611–624. [CrossRef]
54. Amiri, V.; Nakagawa, K. Using a Linear Discriminant Analysis (LDA)-Based Nomenclature System and Self-Organizing Maps (SOM) for Spatiotemporal Assessment of Groundwater Quality in a Coastal Aquifer. *J. Hydrol.* **2021**, *603*, 127082. [CrossRef]
55. Ho, T.K. Nearest Neighbors in Random Subspaces. In *Proceedings of the Advances in Pattern Recognition*; Amin, A., Dori, D., Pudil, P., Freeman, H., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; pp. 640–648.
56. Day, W.H.E.; Edelsbrunner, H. Efficient Algorithms for Agglomerative Hierarchical Clustering Methods. *J. Classif.* **1984**, *1*, 7–24. [CrossRef]
57. Heung, B.; Ho, H.C.; Zhang, J.; Knudby, A.; Bulmer, C.E.; Schmidt, M.G. An Overview and Comparison of Machine-Learning Techniques for Classification Purposes in Digital Soil Mapping. *Geoderma* **2016**, *265*, 62–77. [CrossRef]
58. Sun, D.; Wen, H.; Wang, D.; Xu, J. A Random Forest Model of Landslide Susceptibility Mapping Based on Hyperparameter Optimization Using Bayes Algorithm. *Geomorphology* **2020**, *362*, 107201. [CrossRef]
59. Arabameri, A.; Asadi Nalivan, O.; Saha, S.; Roy, J.; Pradhan, B.; Tiefenbacher, J.P.; Thi Ngo, P.T. Novel Ensemble Approaches of Machine Learning Techniques in Modeling the Gully Erosion Susceptibility. *Remote Sens.* **2020**, *12*, 1890. [CrossRef]
60. Freitas, A.A. Automated Machine Learning for Studying the Trade-Off Between Predictive Accuracy and Interpretability. In *Proceedings of the Machine Learning and Knowledge Extraction*; Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 48–66.
61. Masoud, A.A. Groundwater Quality Assessment of the Shallow Aquifers West of the Nile Delta (Egypt) Using Multivariate Statistical and Geostatistical Techniques. *J. Afr. Earth Sci.* **2014**, *95*, 123–137. [CrossRef]
62. Ha, D.H.; Nguyen, P.T.; Costache, R.; Al-Ansari, N.; Van Phong, T.; Nguyen, H.D.; Amiri, M.; Sharma, R.; Prakash, I.; Van Le, H.; et al. Quadratic Discriminant Analysis Based Ensemble Machine Learning Models for Groundwater Potential Modeling and Mapping. *Water Resour. Manag.* **2021**, *35*, 4415–4433. [CrossRef]

63. Bousouis, A.; Bouabdli, A.; Ayach, M.; Ravung, L.; Valles, V.; Barbiero, L. The Multi-Parameter Mapping of Groundwater Quality in the Bourgogne-Franche-Comté Region (France) for Spatially Based Monitoring Management. *Sustainability* **2024**, *16*, 8503. [CrossRef]
64. Lazar, H.; Ayach, M.; Barry, A.A.; Mohsine, I.; Touiouine, A.; Huneau, F.; Mori, C.; Garel, E.; Kacimi, I.; Valles, V.; et al. Groundwater bodies in Corsica: A critical approach to GWBs subdivision based on multivariate water quality criteria. *Hydrology* **2023**, *10*, 213. [CrossRef]
65. Raiber, M.; White, P.A.; Daughney, C.J.; Tschritter, C.; Davidson, P.; Bainbridge, S.E. Three-Dimensional Geological Modelling and Multivariate Statistical Analysis of Water Chemistry Data to Analyse and Visualise Aquifer Structure and Groundwater Composition in the Wairau Plain, Marlborough District, New Zealand. *J. Hydrol.* **2012**, *436–437*, 13–34. [CrossRef]
66. El Jirari, M.; Barry, A.A.; Bousouis, A.; Zeiki, Z.; Ayach, M.; Sadiki, M.; Bouabdli, A.; Touzani, M.; Guiraud, M.; Valles, V.; et al. Deciphering Spatial Patterns in Groundwater Quality Across Nouvelle-Aquitaine, France: A Multivariate Analysis of Two Decades of Monitoring Data. *Hydrology* **2026**, *13*, 72. [CrossRef]
67. Ayach, M.; Lazar, H.; Bousouis, A.; Touiouine, A.; Kacimi, I.; Valles, V.; Barbiero, L. Multi-Parameter Analysis of Groundwater Resources Quality in the Auvergne-Rhône-Alpes Region (France) Using a Large Database. *Resources* **2023**, *12*, 143. [CrossRef]
68. Durand, S.; Chabaux, F.; Rihs, S.; Duringer, P.; Elsass, P. U isotope ratios as tracers of groundwater inputs into surface waters: Example of the Upper Rhine hydrosystem. *Chem. Geol.* **2005**, *220*, 1–19. [CrossRef]
69. Amiotte Suchet, P.; Probst, J.-L. CO_2 flux consumed by chemical weathering of continents: Influences of drainage and lithology. *Comptes-Rendus l'Acad. Sci. Paris Sér. II* **1993**, *317*, 615–622.
70. Meybeck, M. Composition chimique des ruisseaux non pollués en France. *Sci. Géol. Bull.* **1986**, *39*, 3–77. [CrossRef]
71. Kulabako, N.R.; Nalubega, M.; Thunvik, R. Study of the Impact of Land Use and Hydrogeological Settings on the Shallow Groundwater Quality in a Peri-Urban Area of Kampala, Uganda. *Sci. Total Environ.* **2007**, *381*, 180–199. [CrossRef] [PubMed]
72. Minnesota Department of Health. Pathogen Project: Drinking Water Protection. 2024. Available online: https://www.health.mn.gov/communities/environment/water/docs/swp/pathogproj.Pdf#1#1#1 (accessed on 20 March 2024).
73. Epting, J.; Huggenberger, P.; Radny, D.; Hammes, F.; Hollender, J.; Page, R.M.; Weber, S.; Bänninger, D.; Auckenthaler, A. Spatiotemporal scales of river-groundwater interaction—The role of local interaction processes and regional groundwater regimes. *Sci. Total Environ.* **2018**, *618*, 1224–1243. [CrossRef] [PubMed]
74. Walsh, J.F.; Hunt, R.J.; Anderson, A.C.; Owens, D.W.; Rice, N. Temporally dense monitoring of pathogen occurrence at four drinking-water well sites—Insights and Implications. *Water Res.* **2024**, *259*, 121809. [CrossRef] [PubMed]
75. Kumar, A.; Singh, P.K.; Kumar, V.; Kumar, R.P.; Patel, R.; Kumari, S.; Saw, S. Comparative Analysis of Machine Learning Models in Groundwater Quality Assessment: A Systematic Review. *Water Environ. Res.* **2026**, *98*, e70401. [CrossRef] [PubMed]
76. Amiri, V.; Li, P.; Torabi-Kaveh, M. A Comprehensive Risk Assessment Framework for Groundwater Quality: Integrating Natural Background Levels, Threshold Values, and Machine Learning-Based Hydrochemical Facies Classification. *Water Environ. Res.* **2026**, *98*, e70521. [CrossRef] [PubMed]
77. Mogaraju, J.K. Application of machine learning algorithms in the investigation of groundwater quality parameters over YSR district, India. *Turk. J. Eng.* **2023**, *7*, 64–72. [CrossRef]
78. Taloor, A.K.; Sambyal, S.; Sharma, R.; Dev, S.; Shastri, S.; Kumar, R. Advanced hydrogeochemical facies classification: A comparative analysis of Machine Learning models with SMOTE in the Tawi basin. *Phys. Chem. Earth* **2025**, *137*, 103785. [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.