Scientific Research Publishing

Open Journal of Applied Sciences, 2026, 16(9), 3558-3576
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Geospatial Modeling and Hierarchical Clustering of Hydrogeological Borehole Parameters: Application to the Centre-Ouest Area, Burkina Faso

Fabrice Ouoba[1,2*], Vini Yves Bernadin Loyara[1,3], Hay Yoba Talkibing[1,4], Diakarya Barro[1,5]

[1]Laboratoire d'Analyse Numerique d'Informatique et de Biomathématique (LANIBIO), Universite Joseph KI-ZERBO, Ouagadougou, Burkina Faso
[2]Ecole Supérieure d'Ingénierie (ESI), Université Yembila Abdoulaye TOGUYENI, Fada N'Gourma, Burkina Faso
[3]Centre Universitaire Polytechnique de Kaya, Kaya, Burkina Faso
[4]Faculte des Sciences Tehcniques et Technoligiques (FSTT), Universite de Pala (UDP), Pala, Tchad
[5]UFR-SEG, Université Thomas Sankara, Ouagadougou, Burkina Faso
Email: *didifab@yahoo.fr

How to cite this paper: Ouoba, F., Loyara, V.Y.B., Talkibing, H.Y. and Barro, D. (2026) Geospatial Modeling and Hierarchical Clustering of Hydrogeological Borehole Parameters: Application to the Centre-Ouest Area, Burkina Faso. *Open Journal of Applied Sciences*, **16**, 3558-3576.
https://doi.org/10.4236/ojapps.2026.169197

Received: August 19, 2026
Accepted: September 20, 2026
Published: September 23, 2026

Abstract

This study presents an integrated geostatistical and multivariate clustering framework to characterize hydrogeological variability across 261 boreholes in the crystalline basement aquifers of Burkina Faso. Spatial autocorrelation analysis based on Moran's I indicates significant spatial dependence for borehole depth ($I = 0.088$, $p = 0.0059$) and static water level ($I = 0.063$, $p = 0.0336$), whereas discharge exhibits no significant spatial autocorrelation ($I = -0.032$, $p = 0.7847$). Hierarchical agglomerative clustering identifies two complementary organizational scales: a functional hydrogeological classification into three clusters ($k = 3$, $s = 0.3045$), corresponding to regolith reservoirs, hydrodynamically constrained media, and high-yielding structural fractures, and a broader macro-spatial regionalization into two clusters ($k = 2$, $s = 0.4978$). This scale-dependent organization highlights the dominant role of localized litho-structural heterogeneities in controlling borehole productivity. Variogram-based ordinary kriging captures broad spatial trends, with root mean square errors of 11.658 m for depth and 6.503 m for static water level. However, the low coefficients of determination ($R^2 = 0.068$ for depth and $R^2 = 0.007$ for static water level) indicate limited local predictive performance and substantial short-range variability. Accordingly, kriging is more appropriate for delineating regional spatial patterns than for precise point-scale prediction. Overall, the integrated framework provides a useful basis for regional hydrogeological characterization and supports groundwater management and bore-

hole siting, while highlighting the need to incorporate auxiliary litho-structural, geophysical, and topographic information to improve local-scale prediction.

Keywords

Boreholes, Spatial Autocorrelation, Geostatistics, Hierarchical Clustering, Hydrogeology

1. Introduction

Access to potable water remains a critical challenge in sub-Saharan Africa, and particularly in Burkina Faso, a Sahelian nation where groundwater serves as the primary water source for rural populations [1]. Annual precipitation, ranging from 300 to 1200 mm along a north-south gradient, is insufficient to meet demand, a shortfall further exacerbated by demographic pressure and anthropogenic climate change [2] [3]. Crystalline and crystalline-phyllitic basement aquifers, covering approximately 70% of the country, exhibit strong spatial heterogeneity stemming from differential weathering and bedrock fracturing [4] [5]. This geological complexity hampers the predictive accuracy of borehole hydrodynamic characteristics, leading to non-negligible failure rates (15% - 20%) during rural water supply campaigns and discharge rates frequently below expected thresholds [6] [7].

Given these observations, it becomes imperative to employ quantitative methods capable of characterizing and modeling the spatial variability of hydrogeological parameters. Geostatistics, grounded in the theory of regionalized variables pioneered by Matheron [8], provides a rigorous framework for analyzing spatial dependence and generating predictions at un sampled locations. This approach has been used to model CO_2 storage in classified forests of Burkina Faso [9]. Its central tool is the variogram, which measures the semi-variance of increments:

$$\gamma(\boldsymbol{h}) = \frac{1}{2}\mathbb{E}\left[\left(Z(\boldsymbol{x}+\boldsymbol{h}) - Z(\boldsymbol{x})\right)^2\right], \tag{1}$$

where $Z(\boldsymbol{x})$ designates the regionalized variable (depth, flow rate, or static level) at the point $\boldsymbol{x}$, and $\boldsymbol{h}$ a displacement vector. The experimental estimation of the variogram is carried out by the method of moments:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{(i,j)\in N(h)}\left(Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_j)\right)^2, \tag{2}$$

$N(h)$ is the set of pairs of points about h. Quantifying spatial structure can be achieved by choosing a theoretical model (spherical, exponential, Gaussian, Matérn) and adjusting its parameters (C_0—nugget effect, C—psill, a—range).

The variographic analysis requires an essential complement in the form of a study on global spatial autocorrelation, as this offers insights into the presence of clusters of values that are similar or dissimilar. The Moran index [10] is often used:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}, \tag{3}$$

with w_{ij} a spatial weight. The expectation is $\mathbb{E}[I] = -1/(n-1)$ in the absence of spatial autocorrelation. A significantly positive I denotes spatial clustering of similar values, whereas a negative I suggests spatial dispersion.

In parallel to the continuous spatial analysis, it is advantageous to adopt a multivariate discrete classification framework to identify distinct hydrogeological profiles among individual boreholes [11]. Hierarchical Agglomerative Clustering (HAC) segments the sample of N boreholes based on their dissimilarity in the M-dimensional hydrogeological feature space. Prior to clustering, variables are Z-score standardized, and the pairwise dissimilarity between boreholes p and q is measured using the Euclidean distance:

$$d(p,q) = \sqrt{\sum_{j=1}^{M}\left(Z_{p,j} - Z_{q,j}\right)^2}, \tag{4}$$

where $Z_{p,j}$ represents the standardized value of the j-th hydrogeological parameter for borehole p. The cluster aggregation is performed using Ward's minimum variance criterion [12], which minimizes the increase in within-cluster variance at each merging step:

$$\Delta I(A,B) = \frac{n_A n_B}{n_A + n_B} d^2\left(G_A, G_B\right), \tag{5}$$

where n_A and n_B represent the number of boreholes in clusters A and B, while G_A and G_B designate their respective centroids. The optimal partition quality is validated using the mean Silhouette width [13]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{6}$$

where $a(i)$ denotes the average distance between borehole i and all other elements within its cluster, and $b(i)$ is the minimum average distance from borehole i to elements of the nearest neighboring cluster. A Silhouette value close to 1 indicates a distinct and robust clustering structure, whereas values near 0 or negative suggest overlapping or misclassified instances.

Utilizing both geostatistical methods and classification techniques allows for the understanding of hydrogeological parameter variability at various scales. Continuous spatial analysis enables the creation of predictive maps (kriging) and associated uncertainties. In discrete classification, homogeneous domains are highlighted that help with the regionalization and orientation of field campaigns.

In this work, we apply this integrated approach to data from 261 boreholes located in four provinces of Burkina Faso (Boulkiemdé, Sanguié, Sissili and Ziro). After a description of the data and the geological context, we present the theoretical foundations of the tools mobilized. The results include spatial autocorrelation analysis by the Moran index, hierarchical classification of provinces validated by the silhouette index, variogram modeling and kriging maps. The discussion confronts these results with regional hydrogeological knowledge and previous work, before concluding on the operational and scientific perspectives opened by this study.

2. Materials and Methods

2.1. Study Area and Data

2.1.1. Hydrogeological Context

The study area covers four provinces in the Central-Western region of Burkina Faso: Boulkiemdé, Sanguié, Sissili, and Ziro (Figure 1). These provinces are located between 11˚20' and 12˚40' of north latitude and 1˚50' and 2˚50' of west longitude, covering an area of about 15,000 km².

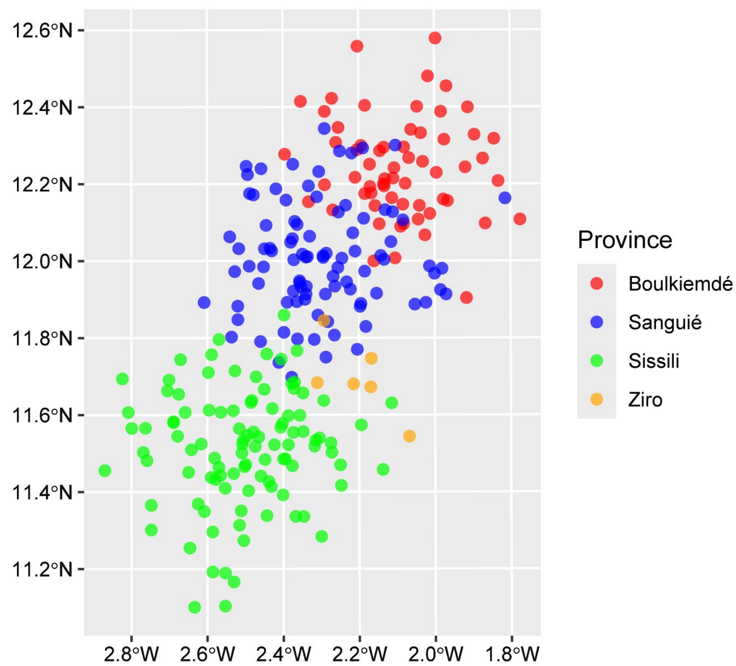


Figure 1. Location of the study area and distribution of drilling.

The geological setting comprises crystalline basement units (granites, migmatites) and sedimentary sequences of the Taoudéni Basin. Basement aquifers, which underlie approximately 70% of the national territory, are typically composed of two vertically superimposed horizons [14]:

- Alterites (alteration zone): thickness of 5 to 40 m, permeability from 10^{-6} to 10^{-4} m/s;
- The fractured horizon: thickness of 10 - 50 m, exhibiting permeability values spanning 10^{-7} to 10^{-5} m/s.

2.1.2. Characterization of the Hydrogeological Database

The hydrogeological dataset was obtained in 2022 from the national borehole database maintained by the Ministry of Water and Sanitation of Burkina Faso. The initial database comprised $N = 281$ borehole records distributed across the four investigated provinces. To ensure the reliability and consistency of the subsequent statistical and geostatistical analyses, a systematic data quality-control procedure was applied. This procedure included verification of the completeness and consistency of spatial coordinates, cross-checking of hydrogeological attributes, identification of physically implausible values, and removal of records containing missing or unreliable measurements. A total of 20 boreholes were excluded because of incomplete records, including missing spatial coordinates (e.g., unidentified villages) or unrecorded hydrogeological attributes. Consequently, a final dataset of $N = 261$ validated boreholes was retained for subsequent analyses.

Table 1. Descriptive statistics of borehole parameters.

Province	*n*	Depth (m)	Flow rate (m^3/h)	Static water level (NS) (m)
Boulkiemdé	61	59.2 ± 13.3	3.03 ± 2.83	12.9 ± 6.77
Sanguié	93	59.9 ± 11.5	3.58 ± 3.36	9.93 ± 5.51
Sissili	101	52.0 ± 10.4	2.96 ± 2.39	13.2 ± 6.5
Ziro	6	57.4 ± 8.12	3.57 ± 2.23	16.0 ± 9.78
Data set	**261**	**56.5 ± 11.99**	**3.21 ± 2.95**	**12.03 ± 6.48**

The retained boreholes were distributed as follows: 61 in Boulkiemdé Province, 93 in Sanguié Province, 101 in Sissili Province, and 6 in Ziro Province (Table 1). Three hydrogeological parameters were considered for each borehole: total drilled depth, expressed in meters (m); discharge rate, expressed in cubic meters per hour (m^3/h) and determined from pumping tests; and static water level (SWL), expressed in meters (m) and measured as the vertical distance between the borehole reference point at the wellhead and the groundwater surface using a water-level sounding probe. Static water-level measurements were conducted during the dry season, specifically between March and May, providing a relatively consistent hydrological reference period for comparison among boreholes. The pumping-test discharge measurements provide an indicator of borehole hydraulic performance, while restricting SWL measurements to the dry season limits the influence of seasonal groundwater-level fluctuations on the spatial analysis.

Descriptive statistics indicate that the borehole depth is 56.5 ± 11.99 m, while the flow rates and static water levels exhibit mean values of 3.21 ± 2.95 m^3/h and 12.03 ± 6.48 m, respectively.

2.2. Spatial Statistical Analysis

Analysis of Spatial Dependence Using Moran's I

Spatial autocorrelation of hydrogeological parameters across the $N = 261$ retained boreholes was evaluated using the global Moran's I index [10]:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \cdot \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \tag{7}$$

where x_i and x_j denote observed values at locations i and j, $\bar{x}$ is the sample mean, and w_{ij} represents spatial proximity weights.

The spatial weight matrix W was constructed using a row-standardized k-nearest neighbors approach ($k = 5$), with non-zero weights defined by inverse Euclidean distance ($w_{ij} \propto d_{ij}^{-1}$ for $j \in N_k(i)$). Statistical significance was assessed non-parametrically via Monte Carlo permutation tests ($n_{\text{sim}} = 999$). To ensure topological robustness, a sensitivity analysis evaluated alternative neighborhood sizes ($k \in \{4,5,6,8,10\}$).

Under the null hypothesis of spatial randomness ($\mathbb{E}[I] = -(N-1)^{-1}$), calculated values $I > \mathbb{E}[I]$, $I \approx \mathbb{E}[I]$, and $I < \mathbb{E}[I]$ indicate positive spatial autocorrelation (clustering), randomness, and spatial dispersion, respectively. Theoretical variance is given by:

$$\mathbb{V}[I] = \frac{N^2 S_1 - N S_2 + 3 S_0^2}{S_0^2 \left(N^2 - 1\right)} - \left(-\frac{1}{N-1}\right)^2 \tag{8}$$

where $S_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$, $S_1 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(w_{ij} + w_{ji}\right)^2$, and

$$S_2 = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} w_{ij} + \sum_{j=1}^{N} w_{ji} \right)^2.$$

2.3. Hierarchical Agglomerative Clustering (HAC)

To identify distinct hydrogeological profiles and spatial patterns, a Hierarchical Agglomerative Clustering (HAC) analysis was performed on the individual borehole observations ($N = 261$). Two independent clustering procedures were implemented:

- Attribute-based clustering: This analysis was based on three hydrogeological variables: borehole depth (m), discharge (m^3/h), and static water level (m). Spatial coordinates were explicitly excluded from this attribute-based segmentation in order to prevent geographical location from influencing the hydrogeological classification.
- Spatial clustering: This analysis was based exclusively on projected metric spatial coordinates (X, Y) in UTM Zone 30N (EPSG:32630), with the objective of characterizing the geographical clustering structure independently of hydrogeological attributes.

Prior to attribute-based clustering, all continuous variables were standardized using Z-score normalization to eliminate scale effects, according to:

$$Z_{i,j} = \frac{x_{i,j} - \bar{x}_j}{\sigma_j} \tag{9}$$

where $x_{i,j}$ denotes the value of variable j for borehole i, while $\bar{x}_j$ and σ_j represent, respectively, the sample mean and sample standard deviation of variable j. For both clustering approaches, the dissimilarity between two observations, p and q, was quantified using the Euclidean distance:

$$d(p,q) = \sqrt{\sum_{j=1}^{M} \left(Z_{p,j} - Z_{q,j}\right)^2} \tag{10}$$

To derive internally coherent clusters, Ward's minimum-variance method was employed. At each agglomerative step, the algorithm merges the pair of clusters, A and B, that produces the smallest increase in the total within-cluster sum of squares, expressed as:

$$\Delta I(A,B) = \frac{n_A n_B}{n_A + n_B} d^2\left(G_A, G_B\right) \tag{11}$$

where n_A and n_B denote the numbers of boreholes assigned to clusters A and B, respectively, and G_A and G_B represent their corresponding centroids. The optimal number of clusters ($k \in 2,3,4,5$) was formally assessed using the mean Silhouette width [13]. For a given observation i, the Silhouette coefficient $s(i)$ is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \tag{12}$$

where $a(i)$ denotes the mean intra-cluster Euclidean distance between observation i and all other observations belonging to the same cluster, whereas $b(i)$ represents the minimum mean distance between observation i and the observations belonging to its nearest neighboring cluster. The overall clustering quality for a given value of k was quantified by the mean Silhouette coefficient across all N boreholes, with values approaching 1 indicating a highly distinct and well-separated cluster structure.

2.4. Geostatistical Modeling

2.4.1. Covariance Function and Stationarity Hypothesis

Kriging is based on the second-order stationarity assumption of the random function $Z(x)$ modeling the studied phenomenon:

$$\begin{cases} \mathbb{E}\left[Z(x)\right] = m \ \text{(constant)} \\ \mathbb{C}\text{ov}\left[Z(x), Z(x+h)\right] = C(h) \ \left(\text{covariance that depends solely on the distance } h = \|\boldsymbol{h}\|\right) \end{cases}$$

In this context, the variogram $\gamma(h)$ is defined as the half-variance of the increases:

$$\gamma(h) = \frac{1}{2} \mathbb{V}\text{ar}\left[Z(x+h) - Z(x)\right] = C(0) - C(h) \tag{13}$$

2.4.2. Variographic Analysis and Spatial Prediction

For a dataset $\left\{Z(x_i)\right\}_{i=1}^{n}$, the experimental variogram is estimated by the method of moments as

$$\hat{\gamma}(h) = \frac{1}{2\left|N(h)\right|} \sum_{(i,j)\in N(h)} \left[Z(x_i) - Z(x_j)\right]^2 \tag{14}$$

where $N(h) = \left\{(i,j) : \left\|x_i - x_j\right\| \in \left[h - \varepsilon, h + \varepsilon\right]\right\}$ is the set of pairs of points about

h, and $|N(h)|$ is the number of such pairs.

To evaluate spatial anisotropy, directional experimental variograms were computed along four principal directions (0˚, 45˚, 90˚, and 135˚ with an angular tolerance of 22.5˚). No significant directional dependence or structural anisotropy was observed, thereby justifying the adoption of isotropic variogram models for subsequent spatial interpolation.

The number of distance classes was determined using an adapted Sturges rule

$$n_{\text{classes}} = \min\left(25, \max\left(15, \left\lceil 1 + \log_2\left(\frac{n(n-1)}{200}\right)\right\rceil\right)\right) \quad (15)$$

This experimental variogram must be adjusted by a theoretical variogram model. We present here five theoretical models (see Table 2) with their analytical expressions and characteristics.

Table 2. Some theoretical models of variogram.

Models	Expression $\gamma(h)$	Parameters	Behavior at the origin
Spherical	$\begin{cases} C_0 + C\left[\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right] & h \le a \\ C_0 + C & h > a \end{cases}$	C_0 (nugget), C (psill), a (range)	Linear
Exponential	$C_0 + C\left[1 - \exp\left(-\frac{3h}{a}\right)\right]$	C_0, C, a (practical range)	Linear
Gaussian	$C_0 + C\left[1 - \exp\left(-\frac{3h^2}{a^2}\right)\right]$	C_0, C, a	parabolic
Matern	$C_0 + C\left[1 - \frac{1}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{h}{a}\right)^{\kappa} \mathcal{K}_\kappa\left(\frac{h}{a}\right)\right]$	C_0, C, a, κ	Variable [a]
Stein	$C_0 + C\left[1 - \frac{h}{a}\mathcal{K}_1\left(\frac{h}{a}\right)\right]$	C_0, C, a	linear

The parameter C_0 (nugget effect) quantifies the variability of the phenomenon at the origin. The parameter a (range) represents the maximum distance of inter-site dependence. The parameter C (sill of the model) measures the large-scale variability. The parameter $\sigma^2 = C_0 + C$ quantifies the total variability. The values of these parameters are obtained via the weighted least squares minimization technique [15].

$$\min_{\theta} \sum_{k=1}^{K} |N(h_k)| \left[\frac{\hat{\gamma}(h_k) - \gamma(h_k;\theta)}{\gamma(h_k;\theta)}\right]^2 \quad (16)$$

where θ denotes the vector of model parameters (C_0, C, a, κ), and $|N(h_k)|$ the number of pairs in the k-th lag class.

The optimal model is the one that minimizes the sum of squared errors (SSE), defined as

$$\mathrm{SSE}=\sum_{k=1}^{K}\left|N\left(h_k\right)\right|\left(\hat{\gamma}\left(h_k\right)-\gamma\left(h_k;\hat{\theta}\right)\right)^2 \tag{17}$$

and which maximizes the variogram coefficient of determination, defined as

$$R_{\text{vario}}^2=1-\frac{\sum_{k=1}^{K}\left|N\left(h_k\right)\right|\left(\hat{\gamma}\left(h_k\right)-\gamma\left(h_k;\hat{\theta}\right)\right)^2}{\sum_{k=1}^{K}\left|N\left(h_k\right)\right|\left(\hat{\gamma}\left(h_k\right)-\bar{\gamma}\right)^2} \tag{18}$$

where $\bar{\gamma}=\frac{1}{\sum_k\left|N\left(h_k\right)\right|}\sum_k\left|N\left(h_k\right)\right|\hat{\gamma}\left(h_k\right)$.

For the subsequent mapping procedure, we resort to ordinary kriging, which is a linear unbiased estimator characterized by minimal variance [8]. Under this framework, the estimate $Z^*\left(x_0\right)$ at an unobserved point x_0 is given by a weighted linear combination of the n measurements:

$$Z^*\left(x_0\right)=\sum_{i=1}^{n}\lambda_i Z\left(x_i\right) \tag{19}$$

with the unbiasedness constraint imposing $\sum_{i=1}^{n}\lambda_i=1$.

Minimization of the estimation variance subject to this unbiasedness condition yields the optimal weights λ_i, which are obtained by solving the ordinary kriging system:

$$\begin{cases}\sum_{j=1}^{n}\lambda_j\gamma\left(x_i-x_j\right)+\mu=\gamma\left(x_i-x_0\right) & \forall i=1,\cdots,n\\ \sum_{j=1}^{n}\lambda_j=1\end{cases} \tag{20}$$

in which μ represents the Lagrange multiplier arising from the unbiasedness constraint.

The resulting estimation variance, commonly referred to as the kriging variance, is given such that:

$$\sigma_K^2\left(x_0\right)=\sum_{i=1}^{n}\lambda_i\gamma\left(x_i-x_0\right)+\mu \tag{21}$$

This variance quantifies the uncertainty associated with the prediction at each point.

2.5. Cross-Validation Assessment of the Kriging Model

The identification of the best-fitting model was achieved through a k-fold cross-validation strategy. In each iteration, the observations belonging to a given fold were estimated on the basis of the data from the remaining $k-1$ folds. The following performance metrics were then derived:

The optimal model is the one that minimizes the RMSE, exhibits a ME close to zero, achieves an R^2 near unity, and minimizes the $\mathrm{RMSE}_{\text{norm}}$. The mathematical formulations of these metrics are provided in **Table 3**.

Table 3. Cross-validation evaluation metrics used for model performance assessment.

Metrics	Formula	Interpretation
RMSE	$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Z(x_i)-Z^*(x_i)\right)^2}$	Prediction error (same units as the variable)
ME	$\mathrm{ME}=\frac{1}{n}\sum_{i=1}^{n}\left(Z(x_i)-Z^*(x_i)\right)$	Estimator bias (≈0 ideally)
R^2	$R^2=1-\frac{\sum_{i=1}^{n}\left(Z(x_i)-Z^*(x_i)\right)^2}{\sum_{i=1}^{n}\left(Z(x_i)-\bar{Z}\right)^2}$	Proportion of variance explained ($0\le R^2\le 1$)
$\mathrm{RMSE}_{\mathrm{norm}}$	$\mathrm{RMSE}_{\mathrm{norm}}=\frac{\mathrm{RMSE}}{\sigma_Z}$	Normalized Error (Scale-Independent Metric)

3. Results and Discussion

3.1. Descriptive Statistics

The analysis is based on 261 boreholes across four provinces (**Table 4**). Depths range from 19.2 to 105.0 m, flow rates from 0.5 to 24.0 m^3/h, and static water levels from 0.27 to 35.0 m. The minimum and maximum borehole depths are observed in Sanguié (19.2 m) and Boulkiemdé (105.0 m), respectively. The minimum flow rate occurs in Sissili (0.5 m^3/h), whereas the maximum occurs in Sanguié (24.0 m^3/h). The lowest and highest static water levels are recorded in Sanguié (0.27 m) and Ziro (35.0 m), respectively. The overall mean values are 56.5 m for depth, 3.21 m^3/h for flow rate, and 12.03 m for static water level.

Table 4. Summary statistics of borehole parameters by province.

Province	Variable	Min	Max	Mean	Standard deviation
Boulkiemdé	Depth	35.5	105.0	59.2	13.3
	Flow rate	0.7	12.0	3.03	2.83
	Static water level	2.5	31.9	12.9	6.77
Sanguié	Depth	19.2	86.2	59.9	11.5
	Flow rate	0.7	24.0	3.58	3.36
	Static water level	0.27	32.0	9.93	5.51
Sissili	Depth	28.0	90.0	52.0	10.4
	Flow rate	0.5	14.4	2.96	2.39
	Static water level	2.17	29.0	13.2	6.50
Ziro	Depth	47.1	69.8	57.4	8.12
	Flow rate	0.9	7.2	3.57	2.23
	Static water level	6.52	35.0	16.0	9.78
Overall	Depth	19.2	105.0	56.5	11.99
	Flow rate	0.5	24.0	3.21	2.95
	Static water level	0.27	35.0	12.03	6.48

Mean values are 56.5 m (depth), 3.21 m^3/h (flow rate), and 12.03 m (static water level) (Table 4). Sanguié has the highest mean depth and flow rate; Ziro has the highest mean static water level. Variability is greatest in Boulkiemdé for depth, in Sanguié for flow rate, and in Ziro for static water level.

The analysis of the stochastic dependence among borehole hydrodynamic parameters relies on accurate estimation of the marginal probability distributions for each variable. Figure 2 shows that these three parameters are not normally distributed.

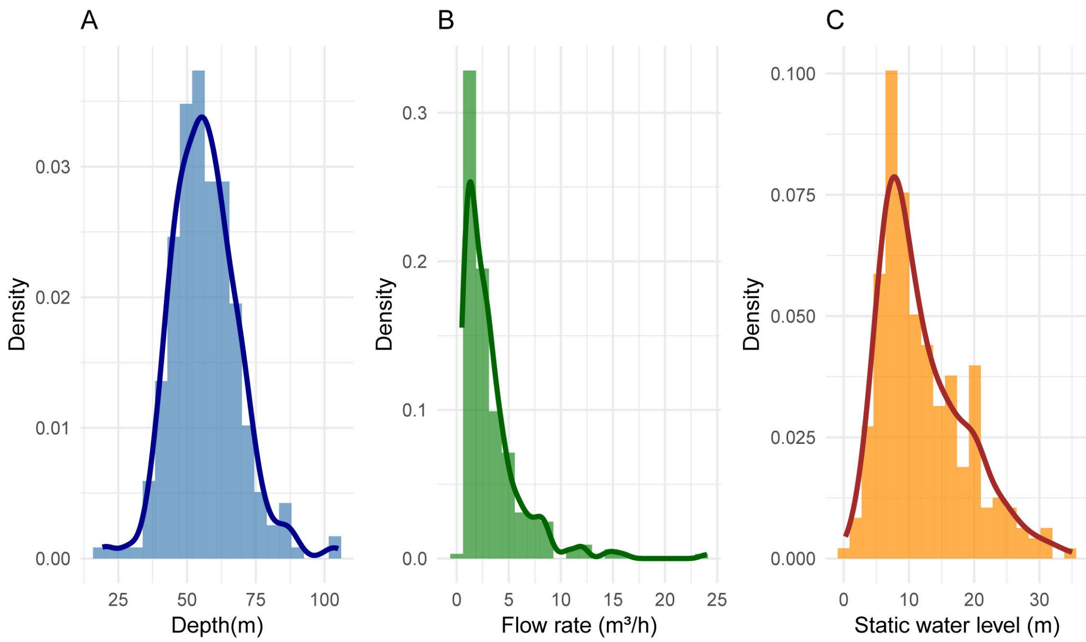


Figure 2. Histograms of the three hydrodynamic variables.

To confirm these departures from normality, we applied the Shapiro-Wilk test (see Table 5). This test reveals that the p-value for each parameter is below (5×10^{-2}).

Table 5. Summary of normality test results.

Test	Variable	Statistic	p-value	Decision
Shapiro	Depth	0.98	2.70×10^{-4}	Normality rejected
	Flow rate	0.74	5.90×10^{-20}	Normality rejected
	Static water level	0.92	2.19×10^{-10}	Normality rejected

Subsequently, four theoretical distribution models were fitted, and the smallest Akaike Information Criterion (AIC) value (see Table 6) was used to select the most appropriate distribution for each parameter. Based on this fitting procedure, we can confirm that depth and static water level follow a Gamma distribution, while flow rate follows a Log-normal distribution.

Table 6. AIC values for theoretical distribution fits.

Distribution	Depth	Flow rate	Static water level
Gamma	2042.71	1090.46	1665.18
Log-normal	2051.19	1053.74	1683.06
Weibull	2070.10	1108.16	1676.17
Exponential	2630.79	2630.79	2630.79

3.2. Analysis of Spatial Autocorrelation

The computed Moran's I indices reveal differentiated spatial structures among the three hydrogeological parameters (Table 7).

Flow rate exhibits a random spatial distribution, as evidenced by a Moran's I value not significantly different from zero. From a statistical standpoint, this finding indicates the absence of spatial structure in borehole productivity; values vary randomly from one location to another, independently of the distance separating them. From a hydrogeological perspective, this randomness suggests that flow rate is primarily controlled by local factors that do not exhibit any form of regional spatial continuity. Consequently, spatial interpolation and mapping of flow rate across the study area are not warranted, as no predictable spatial trend exists.

Table 7. Results of spatial autocorrelation tests.

Variable	**Moran's I**	**p-value**	**Structure**
Depth	0.088	0.0059	Significative positive
Flow rate	−0.032	0.7847	Aléatoire
Static water level (NS)	0.063	0.0336	Significative positive

Sensitivity analysis across neighborhood configurations ($k \in \{4,5,6,8,10\}$) demonstrates the spatial invariance of both borehole depth ($0.0798 \leq I \leq 0.088$, $p < 0.05$) and static water level ($0.0384 \leq I \leq 0.063$, $p < 0.05$), with spatial dependence peaking at $k = 5$ (Table 7). The persistent positive autocorrelation highlights a regionalized spatial structure characterized by spatial clustering of similar values (e.g., adjacent deep boreholes or shallow water tables). However, the modest magnitude of Moran's I reveals that this regional pattern is weak and superimposed on substantial local-scale variability, such as fracture networks and localized aquifer heterogeneity [16]. Despite this local noise, mapping these regional trends via kriging remains essential to identify potential recharge/discharge zones and inform future borehole siting.

3.3. Hydrogeological and Spatial Hierarchical Clustering of Boreholes

Hierarchical cluster analysis (HCA) performed on the entire borehole dataset ($N = 261$) using standardized hydrogeological variables (Depth, Flow Rate, Static

Water Level) and spatial coordinates (UTM) provides critical insights into the structural organization of crystalline basement aquifers within the study area. The evaluation of mean silhouette coefficients across different cluster partitions (Table 8) highlights two distinct scales of spatial and physical organization:

- Hydrogeological Clustering (Physical Variables): The optimal partition is achieved at $k = 3$ clusters with a mean silhouette score of $s = 0.3045$, reflecting a coherent and physically meaningful grouping based on operational borehole characteristics (Figure 3 left).
- Spatial Clustering (UTM Coordinates): The optimal partition occurs at $k = 2$ clusters with a high silhouette score ($s = 0.4978$), confirming a pronounced binary geographic regionalization across the study domain (Figure 3 right).

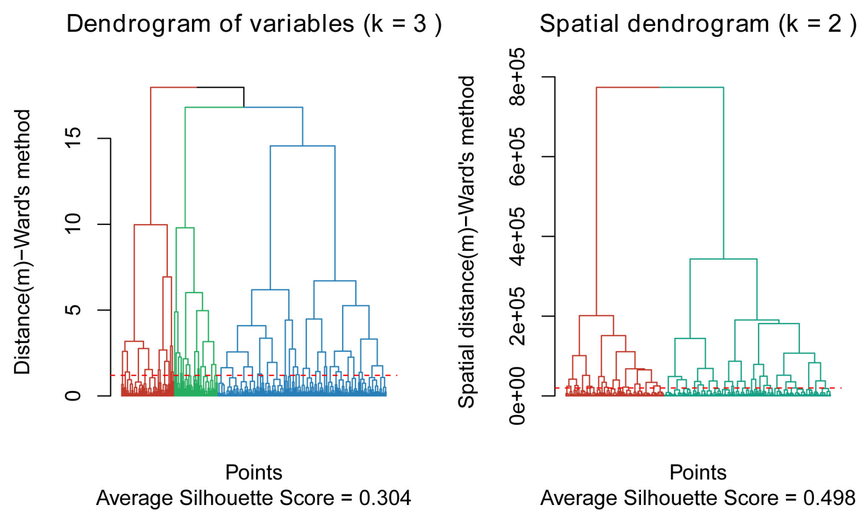


Figure 3. Hierarchical clustering dendrograms (Ward.D2 method): Hydrogeological variables on the left ($k = 3$) and spatial UTM coordinates on the right ($k = 2$).

This decoupling between spatial partitioning ($k = 2$) and hydrogeological partitioning ($k = 3$) underscores the intrinsic heterogeneity characteristic of crystalline basement aquifers [17]. Although geographic distribution exhibits a macro-scale binary trend, the physical behavior and hydraulic performance of individual boreholes are primarily dictated by localized structural and lithological factors [18].

Table 8. Mean silhouette coefficients for standard (hydrogeological) and spatial hierarchical clustering.

Clustering Type	Number of Clusters (k)	Silhouette Index (s)
Standard (Variables)	$k = 2$	0.2882
	$\boldsymbol{k = 3}$	**0.3045**
	$k = 4$	0.2775
	$k = 5$	0.2789

Continued

Spatial (UTM)	$\boldsymbol{k = 2}$	**0.4978**
	$k = 3$	0.3392
	$k = 4$	0.3165
	$k = 5$	0.2827

Analyzing the empirical characteristics of the three identified hydrogeological clusters ($k = 3$) enables the establishment of a functional classification of the boreholes (**Table 9**):

Table 9. Mean hydrogeological characteristics by cluster ($k = 3$).

Hydro_Cluster	Sample Size (N)	Depth (m)	Flow Rate (m^3/h)	Static Water Level (m)
1	166	56.15 ± 12.14	2.40 ± 1.38	8.86 ± 3.54
2	52	55.87 ± 14.27	1.71 ± 1.07	21.79 ± 4.80
3	43	59.21 ± 8.52	8.11 ± 3.84	12.50 ± 5.43

- Cluster 1 (Regolith Aquifers/Weathered Mantle Reservoirs): Comprising 63.6% of the sampled boreholes, this group consists of moderate-depth wells (56.15 ± 12.14 m) characterized by shallow static water levels (8.86 ± 3.54 m) and low to moderate discharge rates (2.40 ± 1.38 m^3/h). This profile corresponds to typical tapping within thick regolith (saprolite) horizons, providing sustained storage capacity but limited specific yield.
- Cluster 2 (Hydrodynamically Constrained/Depleted Zones): Exhibiting a comparable mean depth (55.87 ± 14.27 m), this cluster is distinguished by significantly deeper static water levels (21.79 ± 4.80 m) and low discharge rates (1.71 ± 1.07m^3/h). These restricted hydraulic performances reflect either poorly developed weathered horizons, high local abstraction stress, or low recharge rates.
- Cluster 3 (High-Discharge Hydrogeological Profile): Despite a slightly greater average depth (59.21 ± 8.52 m), this cluster stands out due to substantially higher discharge rates (8.11 ± 3.84 m^3/h) coupled with intermediate static water levels (12.50 ± 5.43 m). This hydrogeological profile is characterized by substantially higher discharge rates and may be compatible with locally enhanced hydraulic connectivity; however, the available dataset does not permit a direct attribution of this cluster to major faults, open shear zones, or specific geological structures.

These findings demonstrate that borehole productivity in this crystalline basement environment is not merely a function of drilling depth, but depends predominantly on intercepting permeable fractured zones (Cluster 3). Consequently, future borehole siting strategies should incorporate targeted geophysical surveys to maximize the success rate of tapping high-yielding aquifer units.

The following section presents the geostatistical modeling approach implemented to generate spatial predictions and maps of these borehole hydrogeological parameters across the study area.

3.4. Variographic Modeling and Spatial Prediction

The spatial autocorrelation analysis, conducted using Moran's I, demonstrated that depth and static water level display significant positive spatial autocorrelation, whereas flow rate shows a random spatial distribution (Table 7). In light of these findings, we proceeded to model the spatial structure of depth and static water level by computing their experimental variograms.For variogram modeling, five candidate models were evaluated (Table 10).

Table 10. Cross-validation performance metrics for variogram models.

Variable	Model	RMSE	MAE	ME	R^2
Depth	Spherical	11.683	8.781	−0.00400	0.064
	Exponential	11.673	8.735	−0.00341	0.066
	Gaussian	11.876	8.986	−0.00737	0.033
	Circular	11.682	8.778	−0.00404	0.065
	Pentaspherical	11.658	8.729	−0.00435	0.068
Static water level	Spherical	6.516	5.259	0.010	0.003
	Exponential	NaN	NaN	NaN	NaN
	Gaussian	6.558	5.292	0.021	0.009
	Circular	6.503	5.249	0.008	0.007
	Pentaspherical	6.5247	5.261	0.013	0.0005

To determine the most suitable theoretical structure for each experimental variogram, a comparative cross-validation procedure was conducted (Table 10). The cross-validation results indicate limited local predictive performance, although the fitted variogram models adequately capture the broad-scale spatial structure of the variables.

For the depth variable, the pentaspherical model proved to be the most performant, yielding the lowest root mean square error (RMSE = 11.658) and mean absolute error (MAE = 8.729), coupled with the highest coefficient of determination ($R^2 = 0.068$). Mathematically, this model can be expressed as follows,

$$\gamma(h) = \begin{cases} 139.76 + 61.4\left[\dfrac{15}{8}\left(\dfrac{h}{476.72}\right) - \dfrac{5}{4}\left(\dfrac{h}{476.72}\right)^3 + \dfrac{3}{8}\left(\dfrac{h}{476.72}\right)^5\right] & \text{if } h \le 476.72 \\ 201.16 & \text{if } h > 476.72 \end{cases} \quad (22)$$

This model highlights a substantial nugget effect ($C_0 = 139.76$), reflecting pronounced small scale spatial heterogeneity, combined with a structural sill of $C = 61.39$ (yielding a total sill of 201.16). The spatially structured component accounts for approximately 30.5% of the total variance. The high spatial range ($a = 476.72$) indicates a large scale regional structural continuity. From a hydrogeological perspective, this combination reflects the joint influence of local micro-fracturing responsible for short distance variability and major regional geological formations controlling the basement morphology.

Regarding the static water level variable, the circular model was selected as it outperformed all alternative structures by minimizing the root mean square error (RMSE = 6.503) and MAE (5.249), while exhibiting a near-zero mean error (ME = 0.008). Figure 4 illustrates the empirical variograms overlaid with their respective fitted theoretical models. Mathematically, this model can be expressed as follows,

$$\gamma(h) = \begin{cases} 33.19 + \dfrac{18.98}{\pi}\left[\arcsin\left(\dfrac{h}{6.58}\right) + \dfrac{h}{6.58}\sqrt{1-\left(\dfrac{h}{6.58}\right)^2}\right] & \text{if } h \le 6.58 \\ 42.68 & \text{if } h > 6.58 \end{cases} \tag{23}$$

This model is characterized by a nugget effect of $C_0 = 33.19$, a structural sill of $C = 9.489$, and a total sill reaching 42.68. The practical range set at $a = 6.58$ defines the spatial correlation limit of the piezometric level. Beyond this distance, groundwater fluctuations become spatially independent, suggesting piezometric dynamics governed by local hydrogeological sub-catchments.

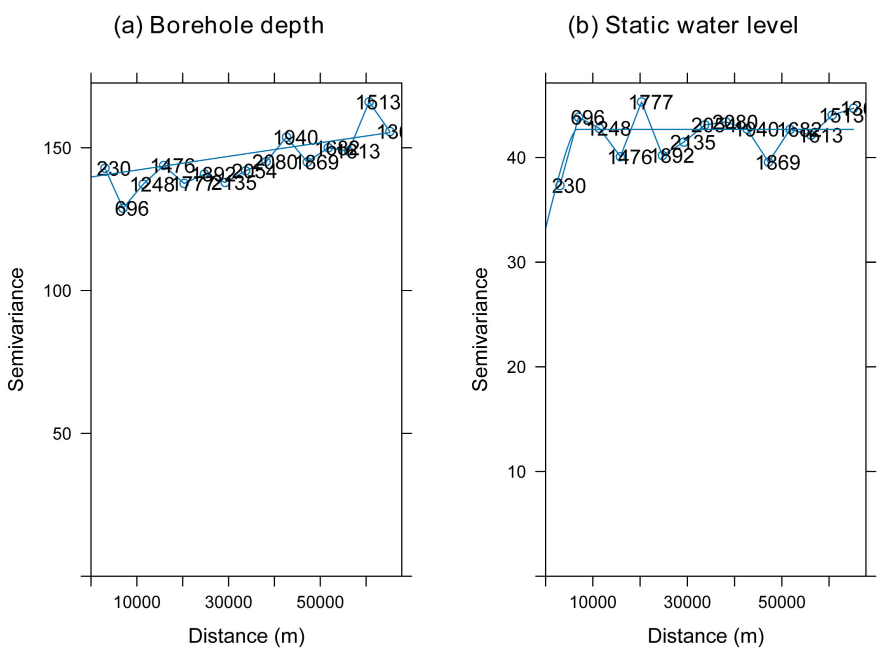
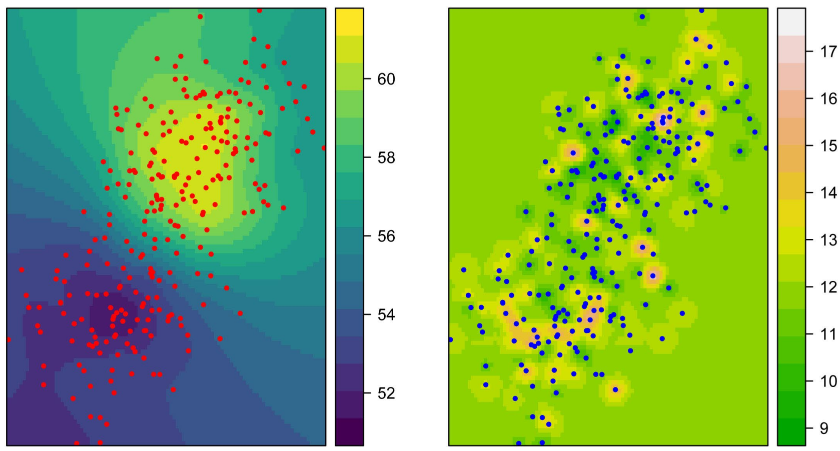


Figure 4. Empirical variograms fitted with the pentaspherical model (left panel) and the circular model (right panel).

Following the validation of the optimal variogram structures, spatial interpolation of the variables was performed using ordinary kriging. Figure 5(a) displays the spatial prediction map for borehole depth, with values ranging from 51 to 61

m. A continuous spatial gradient is clearly evident, extending from the southwestern area (shallower depths, 51 - 53 m, depicted in blue/purple hues) toward the north-central region characterized by the greatest depths (58 - 61 m, depicted in yellow hues).

The spatial prediction map for static water level, presented in **Figure 5(b)**, reveals values ranging from 9 to 17.5 m. The shallowest static water levels (9 - 11 m, represented in dark green to green) are predominantly concentrated along the instrumented central belt surrounding the sampling points. Conversely, intermediate static water levels transition through yellow (12 - 14 m), while localized hotspots and peripheral pockets exhibit the deepest static water levels, ranging from 15 to 17.5 m (represented in salmon to light pink and white).



(a) Borehole depth (b) Static water level in a borehole

Figure 5. Kriging of hydrogeological parameters.

Areas with shallow static water levels (9 - 11 m) represent favorable operational targets due to lower pumping head costs; however, high borehole yield remains conditioned by local fracture transmissivity rather than SWL alone. Conversely, areas with deeper static water levels (15 - 17.5 m) warrant careful consideration from a resource management perspective: the increased depth of the hydraulic head renders the aquifer vulnerable to overexploitation. [19] demonstrated that in such crystalline basement environments, excessive abstraction can induce irreversible drawdown, with long-term ramifications for groundwater availability.

The kriging-derived maps reveal substantial spatial variability in the hydrogeological parameters. These observations support the interpretation of a structured groundwater flow system, primarily governed by the paleotopography of the weathered basement, consistent with the conceptual model proposed by [16]. Within this framework, paleoweathered valleys serve as preferential groundwater flow pathways, combining high well productivity with intrinsic vulnerability to contamination and overexploitation.

Collectively, these spatial patterns provide a regional exploratory basis for groundwater management and preliminary borehole-siting considerations, but should not be interpreted as precise point-scale predictions.

4. Conclusion

This study demonstrates that integrating geostatistical modeling with Hierarchical Agglomerative Clustering provides a useful integrated framework for characterizing heterogeneity in crystalline basement aquifers. Multivariate clustering clearly distinguishes local hydrogeological behavior ($k = 3$, $s = 0.3045$) from regional spatial organization ($k = 2$, $s = 0.4978$), highlighting the dominant influence of localized litho-structural controls on borehole productivity. Ordinary kriging captures broad regional spatial trends; however, its limited predictive performance ($R^2 \leq 0.068$) reveals substantial short-range variability. These findings indicate that spatial interpolation based solely on geographic proximity is insufficient for reliable local-scale prediction. Integrating geophysical and topographic covariates within hybrid geostatistical frameworks, particularly kriging with external drift, therefore represents a critical pathway toward improving local estimation accuracy and supporting sustainable groundwater management in hard-rock environments.

Acknowledgements

The authors express their sincere gratitude to the Ministry of Water and Sanitation of Burkina Faso for making the borehole data available and for their logistical support throughout this research.

Author Contributions

Conceptualization, F.O. and V.Y.B.L.; methodology, F.O.; software, F.O.; validation, F.O., H.Y.T., and V.Y.B.L.; formal analysis, F.O.; investigation, F.O.; resources, F.O.; data curation, F.O.; writing original draft preparation, F.O.; writing review and editing, F.O.; visualization, F.O.; supervision, D.B.; All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] UNESCO (2020) Rapport mondial sur la mise en valeur des ressources en eau, Paris.

[2] Taylor, R.G., Todd, M.C., Kongola, L., Maurice, L., Nahozya, E., Sanga, H., *et al.* (2013) Evidence of the Dependence of Groundwater Resources on Extreme Rainfall in East Africa. *Nature Climate Change*, **3**, 374-378. https://doi.org/10.1038/nclimate1731

[3] GIEC (2021) Changements Climatiques 2021. Cambridge University Press.

[4] Ouandaogo-Yameogo, S., Blavoux, B., Nikiema, J. and Savadogo, A.N. (2013) Caractérisation du fonctionnement des aquifères de socle dans la région de Ouagadougou à partir d'une étude de la qualité chimique des eaux. *Revue des Sciences de l'eau*, **26**,

173-191. https://doi.org/10.7202/1018784ar

[5] Castany, G. (1982) Principes et méthodes de l'hydrogéologie. Dunod.

[6] Direction Générale de l'Eau Potable (2019) Bilan annuel de l'hydraulique villageoise et des forages au Burkina Faso, Direction Générale de l'Eau Potable. Ministère de l'Eau et de l'Assainissement, Ouagadougou.

[7] MacDonald, A.M., Davies, J., Calow, R.C. and Chilton, P.J. (2005) Developing Groundwater. Practical Action, ITDG Publishing, 384.

[8] Matheron, G. (1965) Les variables régionalisées et leur estimation. Masson.

[9] Ouoba, F., Balima, L.H., Talkibing, H.Y. and Barro, D. (2023) Geospatial Modelling of Forest Carbon Stocks in Burkina Faso, West Africa. *Advances and Applications in Statistics*, **89**, 203-225. https://doi.org/10.17654/0972361723058

[10] Moran, P.A.P. (1950) Notes on Continuous Stochastic Phenomena. *Biometrika*, **37**, Article 17. https://doi.org/10.2307/2332142

[11] Kaufman, L. and Rousseeuw, P.J. (1990) Finding Groups in Data. Wiley. https://doi.org/10.1002/9780470316801

[12] Ward, J.H. (1963) Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, **58**, 236-244. https://doi.org/10.1080/01621459.1963.10500845

[13] Rousseeuw, P.J. (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[14] Dewandel, B., Lachassagne, P., Wyns, R., Maréchal, J.C. and Krishnamurthy, N.S. (2006) A Generalized 3-D Geological and Hydrogeological Conceptual Model of Granite Aquifers Controlled by Single or Multiphase Weathering. *Journal of Hydrology*, **330**, 260-284. https://doi.org/10.1016/j.jhydrol.2006.03.026

[15] Cressie, N.A.C. (1993) Statistics for Spatial Data. Wiley. https://doi.org/10.1002/9781119115151

[16] Maréchal, J.C., Dewandel, B. and Subrahmanyam, K. (2004) Use of Hydraulic Tests at Different Scales to Characterize Fracture Network Properties in the Weathered-fractured Layer of a Hard Rock Aquifer. *Water Resources Research*, **40**, 1-17. https://doi.org/10.1029/2004wr003137

[17] Macdonald, A.M., Calow, R.C., Macdonald, D.M.J., Darling, W.G. and Dochartaigh, B.É.Ó. (2009) What Impact Will Climate Change Have on Rural Groundwater Supplies in Africa? *Hydrological Sciences Journal*, **54**, 690-703. https://doi.org/10.1623/hysj.54.4.690

[18] Lachassagne, P., Wyns, R. and Dewandel, B. (2011) The Fracture Permeability of Hard Rock Aquifers Is Due Neither to Tectonics, Nor to Unloading, but to Weathering Processes. *Terra Nova*, **23**, 145-161. https://doi.org/10.1111/j.1365-3121.2011.00998.x

[19] Maréchal, J., Varma, M.R.R., Riotte, J., Vouillamoz, J., Kumar, M.S.M., Ruiz, L., *et al.* (2009) Indirect and Direct Recharges in a Tropical Forested Watershed: Mule Hole, India. *Journal of Hydrology*, **364**, 272-284. https://doi.org/10.1016/j.jhydrol.2008.11.006