Assessment of Groundwater Potential in a Mountainous Area Using Machine Learning and GIS, Rherhaya Basin, High Atlas, Morocco

Mourad Jadoud1,*, Abderrahim El Achheb1, Noureddine Laftouhi2, Abdellah Khouz3,4,5,6, Hasna Eloudi7, Fatima El Bchari8, Said Rachidi3, Blaid Bougadir1, and Tarik Saouab10

1 Department of Geology, College of sciences, Chouaib Doukkali University, El Jadida, Morocco
2 Department of Geology, College of Sciences Semlalia, Cadi Ayyad University, Marrakesh, Morocco
3 Higher School of Technology Essaouira, Cadi Ayyad University, Essaouira, Morocco
4 Department of Sciences and Technology, Aberta University, Lisbon, Portugal
5 Centre of Geographical Studies, Institute of Geography and Spatial Planning, Lisboa University, Lisbon, Portugal
6 Associated Laboratory Terra, Lisboa University, Lisbon, Portugal
7 Department of Geology, College of Sciences, Ibn Zohr University, Agadir, Morocco
8 Polydisciplinary Faculty of Safi, Department of Earth Sciences, Cadi Ayyad University, Safi, Morocco.
9 High School of Technology-Salé, Mohammadia School of Engineers (EMI), Mohamed V University, Rabat, Morocco
10 Department of Earth Sciences, College of Sciences and Technology, Cadi Ayyad University, Marrakesh, Morocco

* Correspondence: mourad_jadoud@yahoo.fr

Abstract

Mountainous regions are vital for recharging aquifers in plain areas downstream, but understanding the geological, hydrogeological, and climatic factors is crucial to comprehend groundwater processes in these regions. Several parameters, including lithology, topography, secondary porosity, geological structures, and climatic conditions, affect the potential of groundwater in mountainous aquifers. Traditional groundwater modeling tools face several challenges in handling large amounts of real-time data, such as extracting useful features, quantifying uncertainty, and identifying links between different variables. Recent technological advances in artificial intelligence, particularly machine learning, provide solutions for hydrogeological research and applications. This paper focuses on modeling potential zones of groundwater sources using various methodologies based on GIS, spatial remote sensing, and machine learning. The study evaluated three models, Random Forest, Support Vector Machine, and Logistic Regression, in identifying potential groundwater zones in the Rherhaya watershed. More than 200 localized spring points were needed to ensure efficient model learning. The Support Vector Machine model demonstrated the highest performance during the 70/30% split, with a ROC-AUC of 84.4% for the test data. The study identified four critical conditioning factors of groundwater potentiality, including Topographic position index, River Distance, Valley Depth, and Plane Curvature. The models also highlighted the distance to rivers as a significant factor, particularly in the upstream portion of the watershed. The very low potentiality class occupied the largest area (over 32%), followed by low (between 24 and 29%), moderate (12–19%), high (10–14%), and very high (only 9–12%) classes. Only the Support Vector Machine model predicted that 12% of the catchment area had high potential for groundwater resources, indicating its superior performance in identifying high-potential zones. The results offer valuable insights that can aid decision-makers in effectively managing water resources in vulnerable areas.

Keywords: Groundwater Potentiality; Random Forest; Support Vector Machine; Logistic Regression; Semi-arid zone; Morocco

DOI: 10.46717/igi.56.2D.22ms-2023-10-28
1. Introduction

Mountainous regions are renowned for their ability to provide 50% of available freshwater (Kohler et al., 2010). These areas play a crucial role in recharging numerous continuous aquifers in downstream plain regions (Bouchaou et al., 2008 and 2009). However, in order to gain a deeper understanding of groundwater dynamics in these mountains, it is essential to thoroughly comprehend the climatic concept of those mountainous regions (Bouimouass et al., 2020).

Additionally, direct access to groundwater outcrops is limited in mountainous regions where groundwater dynamics are deep and complex, making hydrogeological exploration more difficult. Furthermore, these regions are often fractured (such as the Rherhaya watershed in the Atlas Mountains) and contain discontinuous aquifers, which further complicates the evaluation of groundwater potential. The parameters governing this potential in mountainous aquifers are numerous and interdependent, including lithology, geomorphology, topography, secondary porosity, geological structures, fracture density, permeability, drainage patterns and densities, groundwater recharge, water table depth, slope, land use/cover, and climate conditions (Rathay et al., 2018).

Despite significant efforts, it remains a challenge to process hydrological and water-demand data and forecast the effects of various management strategies under the impacts of climate change. To aid water resource planners, integrated water resource models have been developed (Al-Omari et al., 2009). To ensure the sustainable management of water resources in the Rherhaya watershed, it is crucial to create a Water Management Support System that consolidates all current and future resources and demands into a single tool (Rochdane et al., 2012).

In hydrogeology, aquifer simulation is essential for a better understanding of groundwater flow and transport, as well as to guide effective in situ groundwater resource management and remediation. Recent technological advances have provided hydrogeologists with access to significant amounts of real-time data. However, the use of traditional modeling tools to process this data poses certain problems, such as extracting useful features, quantifying uncertainty, or identifying links between different variables. The use of artificial intelligence, particularly machine learning, could be a solution for hydrogeological research and applications. However, the adoption of these methods is hindered by several aspects, such as the “black box” nature of most models, their often limited generalization ability, hypothetical convergence, and uncertain transferability (Adombi et al., 2021).

Proper mapping of groundwater potential requires reliable methods and appropriate conditioning factors. The accuracy of the mapping is heavily influenced by the quality of the dataset, the statistical method or machine learning algorithm (MLA) used, and the size of the study area, as demonstrated by Razandi et al. (2015) and Chen et al. (2022). To this end, a variety of MLA, including tree-based and ensemble methods that combine two or more MLA, have been widely used for GPM. Prasad et al. (2020) and Chen et al. (2022) utilized random forest (RF) and support vector machine (SVM) algorithms to assess groundwater potential zones, while Nguyen et al. (2020) employed ensemble models based on logistic regression (LR) to map groundwater potential.

Understanding hydrogeological properties with inherent spatial heterogeneity, such as hydraulic conductivity and specific storage coefficients, is necessary for simulating aquifers. Traditional methods of groundwater exploration, involving geological, hydrogeological, and geophysical techniques, can be expensive and time-consuming due to drilling costs and the length of investigations. Additionally, these approaches may not fully consider the various factors that influence the presence and flow of groundwater. An alternative approach that is gaining popularity is remote sensing and GIS (Fatah et al. 2020; Al-Gburi et al. 2020; Al-Gburi et al. 2022a; Al-Gburi et al. 2022b; Ramzi et al. 2022), which provides spatial and spectral data over large areas, including inaccessible regions, with frequent revisits. This fast and cost-effective method combines data on geology, geomorphology, lineaments, slope,
morphometry, and other factors, which can aid in the evaluation of potential groundwater zones, particularly when machine-learning algorithms are utilized (Naghibi et al., 2016; Oh et al., 2011).

The proposed research approaches and computational algorithms can also be applied to the inversion of random fields of other physical properties in various scientific disciplines, including hydrology, geology, geophysics, earth sciences, environmental engineering, etc.

The primary aims of this research are to: (1) compare the performance, stability, and reliability of the models using various statistical and validation methods; (2) evaluate the effectiveness of individual and ensemble models in predicting groundwater potential in mountainous regions of the Rherhaya watershed; (3) assess the significance of using a maximum number of groundwater influencing factors; and (4) create dependable maps that illustrate the geographical distribution of groundwater potential in the research area.

2. Study Area

The Rherhaya basin covers 225 km² and originates from the highest point in North Africa, the Toubkal summit (4167 m above mean sea level), and flows through the hydrometric station of Tahanaut (31.3°N, 7.9°W), located at the outlet of Oued Rherhaya, which is formed by the confluence of two oueds, Imenane and Imlil (Fig. 1). It is bounded by the Haouz plain to the north, the High Atlas range to the south, the Ourika basin to the east, and the N’fis basin to the west. The Rherhaya basin has a semi-temperate to sub-humid climate, with sub-humid conditions prevailing in the northern part of the watershed, resulting in significant winter rainfall. The higher elevations of the basin experience temperate temperatures and heavy winter snowfall. The area is characterized by a sharp contrast in altitude and soil cover, with mountainous regions having temporary snow cover, degraded soil, or forested areas. Additionally, the steep slope of the basin results in a shorter time of concentration (El Fels et al., 2018).

![Fig.1. The Rherhaya basin, western High Atlas, Morocco](image)

The study area is part of the Haouz-Tensift basin, which generally experiences low and irregular rainfall and high evaporation rate, with an average annual potential evapotranspiration of 1600 mm yr⁻¹
in the plain. The surface and groundwater resources are overexploited for agriculture and domestic use, leading to a decline in groundwater levels since the 1980s. In the Rherhaya watershed, 60% of the basin surface is occupied by low-permeability formations, while the downstream part is characterized by permeable or impermeable sandstone-clay and limestone sedimentary formations. The vegetation cover is sparse on the slopes, with areas of irrigated crops located near the talwegs. The upstream part of the basin has steeper slopes, while the average rainfall at the Tahanaout outlet station and the Oukaimeden station is 356 mm.yr\(^{-1}\) and 445 mm.yr\(^{-1}\), respectively. The average flow recorded at the Tahanaout hydrometric station is about 1.15 m\(^3\).s\(^{-1}\) according to the data provided by the Tensift Hydraulic Bassin Agency.

The Rherhaya watershed is characterized by steep slopes at higher elevations in the south and more moderate slopes towards the north. Morphometric analysis, as calculated by Sajadi et al., 2022, has identified various factors, such as size, shape, slope, stream order, stream length, and drainage density, that influence the hydrological and geomorphological behavior of the watershed. Pareta (2011) and Sajadi et al. (2022) suggest that these factors regulate the rate of surface runoff, inflow, and erosion. The hypsometric curve for the Rherhaya basin indicates that the area is comprised of mature and youthful landforms at the higher and lower elevation points, respectively.

From a geological perspective, the geology and geomorphology of the High Atlas have been significantly influenced by various orogenetic events. According to Dresch and Proust, several longitudinal zones have been identified in the High Atlas, including the sub-atlantic zone, the Horst of Rherhaya, and the axial zone. The northernmost part of the basin is known as the zone sub-atlasique, and it is made up of a hard, Precambrian and Paleozoic gréso-schistous and calcium-rich core. The reddish Permo-triassic cliffs surrounding this area consist of claystones made of sand and clay. The Asni Basin is dominated by tendril-shaped rocks, while the Kik and Tihallatine plateaus to the north and south of the Asni basin are made up of harder and more carbonated rocks.

To the south, the Rherhaya horn slopes eastward in relation to the Sub-atlantic zone and is primarily made up of rock exposures belong to the Precambrian, Primaeval, and Permo-Triassic periods. The ante-permo-triasic surface became fossilized under a thick layer of detritus deposits. The axial zone, which forms the highest block at the top of the basin, consists mainly of Precambrian outcrops characterized by andesite, rhyolite, and granite rocks. These formations produce the tall peaks culminating in Mount Toubkal.

The Rherhaya Basin's lithology includes magmatic rocks in its southernmost regions., Liassic limestones along the northern boundary, and Triassic sediments in the center and north. The region is distinguished by significant faults in the N 70 and horst directions, leading to a dynamic reworking of litho-units over time, as demonstrated by the exposed and faulted rocks.

2. Materials and Methods

According to Kordestani et al. (2019), a spring is a channel through which groundwater flows from an aquifer to the Earth's surface. This feature reflects the potentiality of groundwater. To evaluate the relationship between the occurrence of springs and factors that control groundwater flow, the researchers used a groundwater potential mapping (GPM) tool to provide spatial information.

The study's methodology is outlined in the flowchart (Fig. 2). The main steps involve preparing the data for modeling, including creating an inventory map of springs and collecting datasets on conditioning factors (Table 1). The researchers then applied a frequency ratio method to identify the spatial relationships between spring occurrence and predisposing factors. They also used RF, LR, and SVM models to map groundwater potential and tested different ensemble models to find the most accurate prediction rate. This resulted in various groundwater potentiality maps.
To test the model results, the researchers applied several statistical parameters and compared them using a correlation matrix and ROC-Curve AUC. The study utilized Geographic Information System environments and statistics software for database preparation and groundwater potential mapping. R packages for modeling machine learning algorithms were also utilized.

![Fig.2. The methodology employed in this research](image)

**Table 1.** Data sources and datasets used to generate the conditioning factors maps

<table>
<thead>
<tr>
<th>Data type</th>
<th>Data denomination</th>
<th>Source</th>
<th>Scale/ resolution/ Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topographic maps</td>
<td>Jbel Toubkal</td>
<td>National Agency of Land Conservation. Cadastre and Cartography (ANCFCC)</td>
<td>1/50 000</td>
</tr>
<tr>
<td></td>
<td>Tahannaout</td>
<td>(Proust, 1973)</td>
<td>1/100 000</td>
</tr>
<tr>
<td>Geological map</td>
<td>Proust map</td>
<td>(Proust, 1973)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amenzal</td>
<td>ABHT</td>
<td>04/1997-12/2022</td>
</tr>
<tr>
<td></td>
<td>Tioudriou</td>
<td>ABHT</td>
<td>07/1996-12/2022</td>
</tr>
<tr>
<td></td>
<td>Agouns</td>
<td>ABHT</td>
<td>09/1996-12/2022</td>
</tr>
<tr>
<td></td>
<td>Nkouris</td>
<td>ABHT</td>
<td>03/1974-12/2022</td>
</tr>
<tr>
<td>Hydrological stations</td>
<td>Yaacoub Al Mansour</td>
<td>Tensift Hydraulic basin agency (ABHT)</td>
<td>09/2008-12/2022</td>
</tr>
<tr>
<td></td>
<td>Améd</td>
<td>ABHT</td>
<td>04/1999-12/2022</td>
</tr>
<tr>
<td></td>
<td>Imin hamam</td>
<td>ABHT</td>
<td>03/1969-12/2022</td>
</tr>
<tr>
<td></td>
<td>Tahnaout</td>
<td>ABHT</td>
<td>03/1969-12/2022</td>
</tr>
<tr>
<td></td>
<td>Agbalou</td>
<td>ABHT</td>
<td>04/1969-12/2022</td>
</tr>
<tr>
<td></td>
<td>Takerkoutee</td>
<td>ABHT</td>
<td>10/1962-12/2022</td>
</tr>
<tr>
<td>Landsat ETM+</td>
<td><a href="https://earthexplorer.usgs.gov/">https://earthexplorer.usgs.gov/</a> (USGS-EROS 2018)</td>
<td>15 m</td>
<td></td>
</tr>
<tr>
<td>Satellite images</td>
<td>High-resolution Ortho-imagery</td>
<td><a href="https://earthexplorer.usgs.gov/">https://earthexplorer.usgs.gov/</a> (USGS-EROS 2018)</td>
<td>0.3 m</td>
</tr>
<tr>
<td>Digital elevation model</td>
<td>Digital elevation model</td>
<td><a href="https://search.asf.alaska.edu/">https://search.asf.alaska.edu/</a> (JAXA/METI 2020)</td>
<td>12.5 m</td>
</tr>
</tbody>
</table>
3.1. The Water Sources Inventory

The water source inventory was compiled using numerous field data sets, resulting in the identification of a total of 217 sources. To ensure the stability of the model, three Training/Validation subdivisions will be tested: 30/70%, 50/50%, and 70/30% (refer to Fig. 2). Furthermore, in order to improve model learning, it is recommended to use equal proportions of source and non-source points. Consequently, 217 non-source points were also identified and randomly mapped. Fig. 3 shows the inventory of sources and non-sources utilized in this study.

![Inventory map](image)

**Fig. 3.** Inventory of sources and non-sources utilized in this study

3.2. Factors Influencing Groundwater (GIFs)

Selecting factors that influence the potential of groundwater is a challenging task due to the complexity of aquifer functioning and the study area. For this study, we combined as many factors as possible that could potentially influence the potential of groundwater by preparing a total of 19 factors in geological, hydrological, climatic, topographic, and land use order (Fig. 4).

Climatic factors, climate is a key factor that has a direct impact on the availability of groundwater, particularly precipitation (Kumar, 2012), which directly allows and promotes aquifer recharge. For this study, annual precipitation data were obtained from the ABHT hydrological measuring stations.

Based on the precipitation map, the average annual precipitation varies across the basin. The southeast regions have averages of more than 574 mm, while the southwest regions have low values of 322 mm. The north and center of the basin record annual precipitation averages of approximately 400 mm (Fig. 4c).
Geological factors, geological factors are essential for understanding the formation, availability, and recharge of groundwater. Among these factors, lithology plays a major role in determining aquifer formation and controlling water permeability and circulation (Muchingami et al., 2022). In the Rherhaya watershed, the lithological map (Fig. 4a) reveals the diversity of rock types, which vary from shale, sandstone, limestone, basalt, granite, clays, etc., and which have a direct influence on the potential of groundwater.

In this tectonically active mountainous context, faults may also play an important role in the emergence of springs and the availability of groundwater. The Marrakech High Atlas, where the Rherhaya watershed is located, is characterized by a prevailing fault system oriented towards the northeast-southwest direction. Thus, the distance to faults has been mapped (Fig. 4b) to better understand the relationship between faults and the potential of groundwater.

In summary, lithology and faults are the key geological factors that have a direct influence on the potential of groundwater in the Rherhaya watershed. Knowledge of their spatial distribution is essential for sustainable groundwater resource management.

Hydrological factors, the selected hydrological factors are the distance to rivers and drainage density. The distance to rivers was calculated using the Euclidean distance method in the ArcGIS Pro environment to determine the distance of sources from the drainage system (Fig. 4j), while drainage density was used to determine the spatial distribution of streams in the study area (Fig. 4e). The maps reveal that river distances range from 0 to 763 m and that river systems are more focused in the northwest portion of the basin, which is a plain zone, than in the southeast part, which is a mountainous zone.

Topographic factors, are critical in regulating hydrological and hydrogeological conditions. For this study, twelve topographic factors were considered (Fig. 4g,h,l,k,l,m,n,o,p,q,r and s): elevation, slope aspect, slope angle, profile curvature, plan curvature, topographic wetness index (TWI), stream power index (SPI), terrain roughness index (TRI), topographic position index (TPI), geomorphic units, valley depth, and slope length (LS).

Slope factors generally have a negative impact on the potential of groundwater, while in flat areas, whether high or low, rainwater has much more time to infiltrate and recharge groundwater (Botzen et al., 2013). For the aspect factor, the exposure of slopes to the sun plays a role in the spatial distribution of vegetation, which in turn impacts groundwater recharge.

### 3.2. Analysis of Groundwater Influencing Factors

#### 3.3.1. Analysis of Multicollinearity and Confusion Matrix

Multicollinearity analysis is a statistical technique used to identify information redundancy across factors that may have a detrimental effect on the performance of the model and to identify and quantify linearity between conditioning factors of a given event. The term "multicollinearity" describes the potential non-independence of conditioning factors in data sets. It is frequently used to foresee a variety
Fig. 4. Spatial distribution of groundwater influencing factors (GIFs): (a) Lithology; (b) Distance to faults; (c) Rainfall; (d) Land use land cover (LULC); (e) Drainage density; (f) NDVI; (g) Elevation; (h) stream power index (SPI); (i) Topographic wetness index (TWI); (j) Distance to rivers; (k) Aspect; (l) Valley depth; (m) Topographic position index (TPI); (n) Terrain roughness index (TRI); (o) Slope; (p) Geomorphons; (q) Curvature profile; (r) Curvature plan; (s) slope length (LS)

of occurrences, including landslides, gully erosion, and the potential for groundwater. Equations (1) and (2) were used in this study to determine multicollinearity for groundwater affecting factors utilizing confusion matrix methods and the variable inflation factor (VIF):

\[
\text{Tolerance} = 1 - R_j^2 \quad (1)
\]
Where
\[ R_j^2 \] is the coefficient of determination

There are linear correlations between conditioning factors when VIF is greater than 10.

### 3.3.2. Spatial correlation between the inventory and groundwater influencing factors

Before moving on to the modeling phase, it is important to assign weights to each class of factors. Several researchers recommend using the frequency ratio (FR) method, which determines the spatial link between predisposing factors and the dependent factor (Juliev et al., 2019). In the FR method, each factor is divided into multiple classes, and equation (3) is used to determine the FR index for each class of factors:

\[
FR = \frac{PS_i}{PD_i} = \left( \frac{NS_i}{NST} \right) * 100 \left( \frac{NA_i}{NAt} \right) * 100
\]

Where:
- \( PS_i \) denotes the percentage of source pixels for each class \( i \) of influencing factors, relative to the total number of source pixels in the study area;
- \( PD_i \) is the percentage of each class \( i \) of influencing factors, relative to the total area;
- \( NS_i \) is the number of source pixels in a thematic class \( i \);
- \( NST \) is the number of pixels of all sources;
- \( NA_i \) is the total number of pixels in a thematic class \( i \);
- \( NAt \) is the total number of all pixels.

If the results show a correlation between each class of influencing factors and groundwater source zones, the last step is to normalize the weighting factor (FR) to give equal importance to different factors. The method used is to rank FR values between 0.01 and 0.99 using the max-min normalization method according to equation (4):

\[
FRN = \frac{FR - Max (FR)}{Max (FR) - Min (FR)} * (0.99 - 0.01) + 0.01
\]

Where:
- \( FRN \) is the normalized FR matrix;
- \( FR \) is the original data matrix.

### 3.3.3. Machine Learning Methods

Random Forest (RF) Model, based on classification and regression trees (CART) (Breiman, 2001) that evaluates the relationships between dependent factors such as groundwater sources and independent factors to identify the most suitable model for building the groundwater potential map and determining the weight of each factor. In order to improve prediction, this method involves creating multiple decision trees using a random selection of a set of predictive factors at each node (Knoll et al., 2019). Consequently, the prediction results of all trees are averaged to build the final set of predictions of the model (Kuhn and Johnson, 2013).

Mtry and ntree are the fundamental variables that make up the RF model. mtry represents the number of factors to consider when constructing each tree, and ntree represents the number of trees. The advantage of this method is that mtry and ntree can be varied to test different possibilities and select
changed the best performance pathways while reducing error. The RF technique additionally enables factor ranking based on relevance. Measuring the average decline in forecast accuracy is used to calculate weight.

LR Model, a machine learning technique created to address classification issues. Without assuming a normal distribution, the LR predictive analysis technique allows continuous or discrete data as model input (Bourenane et al., 2016). The method relies on utilizing probability to establish the correlation between independent factors and the dependent factors. In order to accomplish this goal, the dependent factor is coded in binary form, where the high potential of groundwater is represented by 1 and the low potential of groundwater by 0. Independent factors may be categorical or continuous. Using the frequency ratio method, we categorised all of the factors into numerical values that represented their weights.

\[ P = \frac{1}{1 + e^{-z}} \]

\[ Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \]

Where:
- \( P \) is the probability;
- \( Z \) is the linear combination of independent variables;
- \( \beta_0 \) is the intercept of the model;
- \( \beta_1, \beta_2 \ldots \beta_n \) are the coefficients of the LR model;
- \( x_1, x_2 \ldots x_n \) are the independent variables;
- \( n \) is the number of independent variables.

SVM Model. By using SVM to determine potential groundwater zones, their operation is based on the same classification and regression method as any other machine learning problem. SVM seeks to separate data into classes using a decision boundary called the "hyperplane," which maximizes the distance between these classes. This distance is called the "margin." To determine the optimal decision hyperplane, SVM moves the data into a higher-dimensional vector space, where it is easier to separate them into distinct classes.

In the field of potential groundwater zone determination, this means that SVM uses geological, topographical, climatic, and other environmental data to create a high-dimensional representation of the data space. SVM can then use this representation to identify potential zones where groundwater may be found by maximizing the margin between zones where water is present and those where it is absent.

3.3.4. Measures and Performance Comparison

Validating the results of the modeling is an essential step to confirm the validity of the results and the performance of the models. However, it is important to verify the stability of the calculations when the partition of the database is modified. To do this, several sample divisions, such as 30/70%, 50/50%, and 70/30%, were examined to evaluate the gains in terms of success rate and prediction of the models. Then, statistical measures were used to evaluate the performance of LR, Random Forests (RF), and SVM models on the test dataset.

The statistical measures, which is essential for evaluating the performance of the models, is based on calculating four parameters: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These parameters are determined by calculating the number of correctly or incorrectly classified source pixels in the training datasets of the models and in their test data. Sensitivity is the proportion of source pixels correctly classified as potential groundwater zones, while specificity is the proportion of non-source pixels correctly classified (Song et al., 2018).
Other parameters were computed to enhance the comparison among the models. Accuracy measures the proportion of pixels correctly classified; precision measures the proportion of pixels classified as sources that are actually sources; FP rate measures the proportion of pixels classified as sources that are not real sources; the Matthews correlation coefficient (MCC) measures the correlation between the model’s predictions and the actual observations; and the Kappa index measures the agreement between the model’s predictions and the actual observations, taking into account the possibility of chance agreement.

A model’s performance is better when its sensitivity, MCC, specificity, accuracy, precision, FP rate, and Kappa index values are higher. A Kappa index value of 1 means the model is perfect, while -1 indicates that the model is unreliable. The following equations were used to calculate these parameters:

\[
\text{Sensitivity} = \frac{TP}{TP - FN} \quad (7)
\]
\[
\text{Specificity} = \frac{TN}{FP - TN} \quad (8)
\]
\[
\text{Accuracy} = \frac{TN + TP}{TP + FP + TN + TP} \quad (9)
\]
\[
\text{Kappa} = \frac{\text{Exactitude} - B}{1 - B} \quad (10)
\]

Where

\[
B = \frac{(TP + FN)(TP + FP) + (TN + FP)(TN + FN)}{\sqrt{TP + TN + FN + FP}} \quad (11)
\]

Receiver Operating Characteristic (ROC) Curve: The ROC curve is a very useful validation method to evaluate the excellence and performance of machine learning models (Song et al., 2018). This method compares the potential groundwater map to the source inventory using the ROC curve. This curve graphically represents the percentage of true positives on the y-axis and the cumulative percentage of false positives on the x-axis (Demir et al., 2015). Then, the area under the ROC curve (AUC) is calculated and used to evaluate the accuracy of the model. The value of the area under the ROC curve ranged from 0 to 1, and could be classified as poor (0.5–0.6), fair (0.6–0.7), good (0.7–0.8), very good (0.8–0.9), and excellent (0.9–1.0) (Fawcett, 2006; Yesilnacar and Topal, 2005).

4. Results

4.1. Selection and Analysis of GIFs

A study was carried out to determine the factors that had an impact on the selection of the most relevant GIFs and to remove those exhibiting multicollinearity. First, the multicollinearity analysis of the 19 groundwater influencing factors showed that the tolerance values ranged from 0.349 for the "elevation" factor to 0.953 for the "TRI" factor. Similarly, the VIF values fluctuated between 1.049 for the "TRI" factor and a maximum value of 2.867 for the "elevation" factor (Table 2). The outcomes are satisfactory as the tolerance values exceed 0.1 and the VIF values are below 10, indicating that there is no multicollinearity among the chosen factors. Moreover, the results of the confusion matrix diagram indicate a low linear correlation between all variables (Fig. 5). These results suggest that the data redundancy of some factors has a minor impact on performance, indicating that all factors must be considered in this analysis.
Table 2. Multicollinearity diagnostic, VIF for groundwater influencing factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>VIF</th>
<th>TOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td>1.124</td>
<td>0.890</td>
</tr>
<tr>
<td>Plan curvature</td>
<td>1.508</td>
<td>0.663</td>
</tr>
<tr>
<td>Profile curvature</td>
<td>1.207</td>
<td>0.829</td>
</tr>
<tr>
<td>Distance to faults</td>
<td>1.235</td>
<td>0.810</td>
</tr>
<tr>
<td>Distance to rivers</td>
<td>1.660</td>
<td>0.602</td>
</tr>
<tr>
<td>Drainage density</td>
<td>1.232</td>
<td>0.812</td>
</tr>
<tr>
<td>Elevation</td>
<td>2.867</td>
<td>0.349</td>
</tr>
<tr>
<td>Geomorphological units</td>
<td>1.081</td>
<td>0.925</td>
</tr>
<tr>
<td>Land use</td>
<td>1.561</td>
<td>0.641</td>
</tr>
<tr>
<td>Lithology</td>
<td>2.645</td>
<td>0.378</td>
</tr>
<tr>
<td>NDVI</td>
<td>1.252</td>
<td>0.799</td>
</tr>
<tr>
<td>LS</td>
<td>1.376</td>
<td>0.727</td>
</tr>
<tr>
<td>Precipitation</td>
<td>2.406</td>
<td>0.416</td>
</tr>
<tr>
<td>Slope</td>
<td>1.136</td>
<td>0.880</td>
</tr>
<tr>
<td>SPI</td>
<td>1.943</td>
<td>0.515</td>
</tr>
<tr>
<td>TPI</td>
<td>2.131</td>
<td>0.469</td>
</tr>
<tr>
<td>TRI</td>
<td>1.049</td>
<td>0.953</td>
</tr>
<tr>
<td>Valley depth</td>
<td>1.742</td>
<td>0.574</td>
</tr>
</tbody>
</table>

Fig. 5 displays the linear correlation results among all the conditioning factors. According to the FR analysis, the weights of different classes of factors indicate an upward correlation between the potentiality of springs and high values of TWI. This is evidenced by the highest FR weight of classes 12–16, which is 7.738. The class 1 of TPI follows closely with a value of 7.457 and is followed by the class 4 of geomorphological units (FR=5.36) and the class 5 of LS (4.82).

Additionally, the significance of factors was evaluated using three Machine Learning methods (RF, SVM, and LR) (Fig. 6). According to these three methods, the main factors contributing to groundwater
potential are valley depth, TPI, distance to rivers, and planar curvature. Conversely, slope and distance to faults are considered less significant.

![Fig. 6. Importance of conditioning factors for the three models](image)

### 4.2. Groundwater Potential Maps

The analysis conducted using the three methods resulted in three maps of groundwater potential (Fig. 7a, b and c). The RF, SVM, and LR models revealed that the low potential class represented the largest part of the watershed area, accounting for 32%, 34%, and 36% of the total area, respectively. On the other hand, areas with excellent potential accounted for only 9%, 12%, and 10% according to the RF, SVM, and LR models, respectively (Table 3). From a spatial perspective, high potential areas were generally located upstream of the watershed, particularly in the slopes adjacent to the watercourses, suggesting a significant contribution of surface water to the hydrological system.

**Table 3.** Percentage of the different classes of the realized maps

<table>
<thead>
<tr>
<th>Classes</th>
<th>RF%</th>
<th>SVM%</th>
<th>LR%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Low</td>
<td>32.52</td>
<td>34.29</td>
<td>36.25</td>
</tr>
<tr>
<td>Low</td>
<td>24.07</td>
<td>29.66</td>
<td>24.3</td>
</tr>
<tr>
<td>Medium</td>
<td>19.21</td>
<td>12.83</td>
<td>16.36</td>
</tr>
<tr>
<td>High</td>
<td>14.61</td>
<td>10.22</td>
<td>12.95</td>
</tr>
<tr>
<td>Excellent</td>
<td>9.60</td>
<td>13</td>
<td>10.14</td>
</tr>
</tbody>
</table>

### 4.3. Results Validation

The validation of the results demonstrated the excellent performance of the three models in detecting potential groundwater zones in this mountainous system (Figs. 8, 9, 10, and Table 4). The SVM model was found to be the most performant among the three models during the 70/30% split, with an AUC of 84.4% for the test data. However, the RF model showed the lowest performance during the 50/50% split. Overall, the three models demonstrated stability with respect to variations in the percentages of data reserved for training and validation.
Fig. 7. Groundwater potential maps are produced by the (a) RF model; (b) LR model and (c) SVM model.
Fig. 8. Comparison of the performances of the RF, SVM, and LR models using ROC curves and AUC values for the 30/70% subdivision

Fig. 9. Performance comparison of RF, SVM, and LR models using ROC curves and AUC values for the 50/50% subdivision

Fig. 10. Comparison of the performances of the RF, SVM and LR models using ROC curves and AUC values for the 70/30% subdivision
Table 4. Results of model validation

<table>
<thead>
<tr>
<th>Indexes</th>
<th>RF</th>
<th>SVM</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.78</td>
<td>0.84</td>
<td>0.78</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.75</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.77</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.53</td>
<td>0.51</td>
<td>0.49</td>
</tr>
</tbody>
</table>

5. Discussion

Based on the results of the investigation, this section will examine the outcomes in three parts: (i) Contributing factors to GWP; (ii) Precision and relevance of generated GWP maps; and (iii) Methods applicability and limitations.

5.1. Contributing Factors to GWP

In mountainous environments, the availability and accessibility of groundwater are influenced by a variety of factors. In this study, we have identified 19 factors based on our familiarity with the studied area and the availability of data. These components fall into four main groups: Topographic factors (Elevation, Aspect, Slope, Plan Curvature, Profile Curvature, SPI, TWI, Valley Depth, TPI, TRI, LS, and Geomorphons); Hydrological Factors (Distance to rivers and Drainage density); Geological factors (Lithology and Distance to faults); Anthropogenic factors (NDVI and LULC); and Climatic factors (Rainfall). The results indicate the fundamental importance of four major factors: TPI, River Distance, Valley Depth, and Plane Curvature. This is due to the importance of topographic control on groundwater and the development of the aquifer system in high-altitude regions, which favor snowfall, and since the periods of snow presence sometimes exceed five months (Boudhar et al., 2009), leads to the high infiltration and recharge of aquifers. Also, the research region is distinguished by the excavation of deep valleys through which rivers flow and springs emerge. Mosavi et al., 2021 in the province of Fars, Iran, and Pradhan et al. (2021) in the Nepal Himalayas, both reached the same finding about mountainous regions. From a different perspective, the three utilized models identify the distance to rivers as having a significant role in the groundwater potentiality, particularly in the upstream portion of the watershed. This may be addressed by the interaction of surface water and groundwater in this location, which is consistent with studies indicating that the aquifers are mostly recharged by surface water in mountainous regions (Ma et al., 2005).

5.2. Precision and Relevance of Generated GWP Maps

The methodologies adopted allowed for the development of three maps depicting the groundwater’s potential spatial variability. The three models reveal that the upstream sector of the watershed has a larger amount of potential compared to the downstream regions. This makes perfect sense given that the upstream sections get the most precipitation, particularly snowfall (Boudhar et al., 2009; Ouassanouan et al., 2022). Moreover, when considering the area occupied by each potentiality class, it is clearly apparent that the three models operate in a manner that is highly comparable, indicating that the very low potentiality class has the highest percentage (over 32%), followed by the low (between 24 and 29%), moderate (12-19%), high (10-14%), and very high (only 9-12%) classes. And therefore, only the SVM model, which predicts that 12 percent of the catchment area has a high potential for groundwater resources, is regarded as having a high potential. Comparable results were found by Thanh et al. (2022), using an RF model, and Ouali et al., 2023 using machine and deep learning methods.
5.3. Methods Applicability and Limitations

This paper proposes a low-cost, easy-to-use, and effective method for predicting groundwater potential in mountainous and semi-arid areas. The challenge here is to implement the models properly and to test their performance before their application. In addition, the representativeness of the inventory data is the most crucial step. In this instance, Moghaddam et al. (2020) imply that certain models are susceptible to sample size reduction; consequently, we used a database that included 217 source points that were spatially distributed over the watershed. Moreover, the stability of the models was evaluated under various Training/Validation proportions, as suggested by numerous authors (Namous et al., 2021), and the models’ performance was stable throughout the subdivision of the three databases (30/70%, 50/50%, and 70/30%), with SVM and LR proving to be the most stable models. In terms of model performance, the used approaches performed well, with SVM providing the most accurate prediction. Similar findings have been identified in other parts of Morocco and worldwide (Farzin et al., 2021; Naghibi et al., 2017; Namous et al., 2021). However, it is very important to take into consideration some limitations of the methods applied. Firstly, the inventory of springs must be representative in both number and water supply, as springs may dry up periodically during the dry season or even completely in response to periods of drought linked to climate change. For this reason, it is recommended that the database of springs be updated, especially in terms of productivity. Secondly, and depending on the characteristics of the target area, the maximum possible number of conditioning factors should be integrated into the analysis. Thirdly, ML models are very sensitive to the hyperparameters specific to each model, so the solution is to conduct an optimization for each specific use.

6. Conclusions

In summary, the study successfully applied various GIS and machine learning techniques to assess the groundwater potential of the Rherhaya watershed in Morocco. The researchers utilized several conditioning factors and localized source points to prepare the data for modeling and applied the frequency ratio method to identify the spatial relationships between spring occurrence and predisposing factors. They also employed three different models (RF, SVM, and LR) to map groundwater potential and tested different ensemble models to find the most accurate prediction rate.

The results of the analysis showed that the low-potential class represented the largest part of the watershed area, while areas with excellent potential accounted for only a small percentage of the total area. High-potential areas were generally located upstream of the watershed, particularly on the slopes adjacent to the watercourses, suggesting a significant contribution of surface water to the hydrological system. The study’s findings can serve as an important database for decision-makers to contribute to effective water resource management in vulnerable areas.

Overall, this study showcases the effectiveness of using GIS, spatial remote sensing, and machine learning techniques to evaluate groundwater potential in a mountainous context. The results of this study have the potential to guide water resource management decisions in the Rherhaya watershed and beyond. With the increasing demand for water resources globally, it is crucial to have accurate and up-to-date information on groundwater potential, and studies like this one can contribute significantly to achieving that goal.

References


