Groundwater Potential Mapping Using a Novel Statistical-Data Mining Ensemble Model

Mojtaba Dolat Kordestani[1], Seyed Amir Naghibi[2*], Hossein Hashemi[2], Kourosh Ahmadi[3], Bahareh Kalantar[4], Biswajeet Pradhan[5, 6]

1. Department of Range and Watershed Management, Faculty of Agriculture and Natural Resources Sciences, University of Hormozgan, Bandar Abbas, Iran
2. Center for Middle Eastern Studies & Department of Water Resources Engineering, Lund University, Lund, Sweden
3. Department of Forestry, College of Natural Resources, Tarbiat Modares University, Noor, Mazandaran, Iran
4. Disaster Resilience Science Team, RIKEN Center for Advanced Intelligence Project, Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
5. School of Systems, Management, and Leadership, Faculty of Engineering and IT, University of Technology Sydney, Australia
6. Department of Energy and Mineral Resources Engineering, Choongmu-gwan, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea

*Corresponding author: Amirnaghibi2010@yahoo.com; Amir.naghibi@cme.lu.se

Abstract

Freshwater scarcity is an ever-increasing problem throughout the arid and semi-arid countries, which results in poverty. Thus, it is necessary to enhance our insights into the freshwater resources availability, particularly groundwater, and to be able to implement functional water resources plans. This study introduces a novel statistical approach-data mining ensemble model, through implementing Evidential Belief Function and Boosted Regression Tree (EBF-BRT) algorithms for groundwater potential mapping of the Lordegan aquifer in central Iran. To do so, spring locations are determined and partitioned into two groups for training and validating the individual and ensemble methods. In the next step, twelve groundwater conditioning factors (GCFs) including topographical and hydrogeological factors are prepared for the modeling process. The mentioned factors are employed in the application of EBF model. Then, the EBF values of GCFs are implemented as input to the BRT algorithm. The results of the modeling process are then plotted to produce groundwater spring potential maps. To verify the results, the Receiver Operating Characteristics (ROC) test is applied to the model's output. The findings of the ROC test indicated that the areas under the curves are 75 and 82% for EBF and EBF-BRT models, respectively. Therefore, it can be inferred that the combination of the two techniques could increase the efficacy of them in the groundwater potential mapping.

Keywords: Geographic information system (GIS), Groundwater, Water resources management, Data mining, Iran

1. Introduction

Groundwater could be regarded as the water identified in the saturated parts of the Earth, which fills the pore section of geologic formations and soil beneath the water table (Freeze and cherry 1979). Groundwater has broader advantages over surface water including its capability to be utilized when needed, and it is less vulnerable to catastrophic incidents (Naghibi and Pourghasemi 2015). Furthermore, groundwater contributes the most in supplying freshwater demands in arid and semi-arid areas such as the Middle East (Chezgi et al. 2015). Groundwater potential mapping is one of the well-studied subjects in the literature and has attracted many researchers over the years.

Many researchers have used statistical and data mining algorithms to map groundwater potential. Some of them have used spring locations as groundwater indicator, while others used qanat and well locations. According to the literature, frequency ratio (Naghibi et al. 2015), weights-of-evidence (Ozdemir 2011a; Corsini et al. 2009; Razandi et al. 2015; Tahmassebipoor et al. 2016), and index of entropy (Naghibi et al. 2015) are among the most popular methods used by the scholars. Moreover, other data mining methods such as classification and regression tree, random forest, and boosted regression tree (BRT) are widely used techniques to assess the potential of groundwater (e.g. Naghibi and Pourghasemi 2015; Naghibi et al. 2016; Zabihi et al. 2016; Rahmati et al. 2016; Mousavi et al. 2017; Golkarian et al. 2018). Although data mining techniques have proved to be liable in working with nonlinear and complex data (Naghibi et al. 2016), one of the drawbacks is overfitting, which impacts the models' estimation quality and prediction validity. In two recent papers by Naghibi and Moradi Dashtpagerdi (2016) and Naghibi et al. (2018), various data mining algorithms including random forest, BRT, support vector machine, artificial neural network, quadratic discriminant analysis, linear discriminant analysis, flexible discriminant analysis, penalized discriminant analysis, k-nearest neighbors, and multivariate adaptive regression splines were employed for groundwater assessment taking into account spring and qanat locations. Other techniques include evidential belief function (EBF) method to map the potentiality of groundwater (Nampak et al. 2014; Rahmati and Melesse 2016). Nampak et al. (2014) used EBF to map groundwater potential and compared its performance with a logistic regression model. The results indicated the superior performance of the EBF model to logistic regression. In another research, Naghibi and Pourghasemi (2015) examined the efficacy of the EBF model and compared the results with classification and regression tree, random forest, BRT, and generalized linear model. Their findings also yielded in an acceptable performance of the EBF model.

The above-mentioned studies mostly used single models in the groundwater-related research however, the ensemble models have been used in other fields of study including landslides (Lee et al. 2012; Umar et al. 2014) and flood susceptibility modelling (Tehrany et al. 2013, 2014). Very recently, Naghibi et al. (2017b) introduced a novel ensemble model, which was constructed based on four data mining models and frequency ratio in a groundwater related study. The findings of their research indicated that the produced ensemble model showed a better performance than a single application of the models. Similarly, Pourghasemi and Kerle (2016) combined EBF and random forest models to achieve better model performance and their results indicated a higher efficacy of the ensemble method.

BRT as a data mining technique was selected for this purpose as it has the ability for feature selection (Naghibi et al. 2016) as well as implementing the stochastic gradient boosting to diminish variance and bias (Abeare, 2009). BRT model also defines the importance of the impacting factors in the modelling procedure. Considering the aforementioned strong features of the BRT model, this model was chosen to be combined with EBF model to improve its prediction accuracy. In this research, the proposed ensemble method (EBF-BRT) improves on the weak points of each method and combines their advantages by analyzing the relationships of groundwater with each independent layer and with each class of independent layers. Furthermore, groundwater-related independent variables can be assessed. Since this combined approach is almost new in groundwater potential assessment, through this research its efficiency and capability can be examined. This research aims to improve the performance of statistical techniques through the extension of statistical-data mining ensemble model in a groundwater potential mapping. Thus, the aims of this study are: (i) evaluating the performance of the EBF-BRT model in groundwater potentiality assessment, (ii) ranking the importance of Groundwater Conditioning Factors (GCFs) and the relationship between groundwater potential and the GCFs, and (iii) providing spatial information and guideline to support decision making process concerning groundwater management in the Lordegan aquifer.

2. Material and methods

Spring can be defined as places where groundwater flows from an aquifer to the surface. Based on the physiographical and hydrological characteristics of the study area, this study assumes that the natural spring occurrences and their discharge rates can be related to the potential of groundwater resources in the studied basin. To quantify this relationship, Groundwater Potential Map (GPM) is proposed as a tool for providing spatial information and determining the relationship between the spring occurrence and effective factors, here is called conditioning factors.

For modelling of groundwater potential, two datasets were prepared including spring locations inventory and the GCFs. Using the mentioned datasets, EBF model was conducted, and the resultant GPM was plotted using ArcGIS 10.4. In the next step, EBF values were extracted and then used as an input to the BRT model, and the ensemble EBF-BRT model was trained. Finally, by implementing ROC plot, the efficacy of the EBF and EBF-BRT methods were validated. Figure 1 shows the methodology flowchart implemented in this research.

2.1. Study area and preparation of the conditioning factors

2.1.1. Study area

The Lordegan Basin covers the areas between 31°19'09" and 31°38'06" North latitudes and 50°28'02" and 51°13'13" East longitudes and is located in Chaharmahal-e-Bakhtiari Province, Iran. Lordegan Basin covers an area of 1,486 km^2. Altitude in Lordegan Basin ranges between 850 and 3,640 m above mean sea level (amsl) with a mean altitude of 2,044 m amsl. The lithology of the Lordegan Basin is mainly composed of sedimentary and tertiary rocks and quaternary deposits, and about 33.3% of its area is classified under group 5 including low-level piedmont fan and valley terraces deposits (Table 1). The dominated land use is rangeland, which covers approximately 44%

of the basin floor. Other types of land use encompass forest, agriculture, orchard, and residential area. Spring occurrence is not limited to the plain areas and it can be seen on different slopes and altitudes hence, the study was carried out at the basin scale.

2.1.2. Data preparation

In this study, a spring inventory dataset including 94 springs (2014) was prepared based on the field surveys (Fig. 2). The dataset was then split into two subsets for training (70% of the dataset: 66 springs) and validating (30% of the dataset: 28 springs) the models (Pourghasemi and Beheshtirad 2015). It should be noted that the division of the spring dataset into two subsets was conducted on the basis of a random algorithm in ArcGIS 10.4.

Based on the literature (Ozdemir 2011a, b) and availability of data, twelve GCFs were selected for the modelling process. GCFs are composed of eight topographical factors, two river-related factors, and two physical factors including land use and lithology. It should be noted that as EBF works with classified factors, GCFs were classified based on the literature (Ozdemir 2011a, b; Naghibi et al. 2018).

In the first step, a 20 m resolution Digital Elevation Model (DEM) of the studied basin was derived from a 1:50,000-scale topographic map. The slope angle derived from DEM was split into four ranges of 0-5, 5-15, 15-30, and >30 degree (Fig. 3a). Slope aspect was also derived from DEM data and then classified into nine classes (Fig. 3b). Altitude is another important GCF (Ozdemir 2011a, b) that was employed in this investigation (Fig. 3c). The altitude of the studied basin was partitioned into five equal classes.

Plan curvature is a topographical-based variable, which shows the direction of flow (Ozdemir 2011a) (Fig. 3d). Profile curvature clarifies at which rate the slope changes in the maximum slope direction (Ozdemir 2011b) (Fig. 3e). Slope-length (LS) is considered as a mixture of the two variables of slope steepness and slope length (Naghibi et al. 2016) and is calculated as follows (Moore et al. 1991) (Fig. 3f):

$$LS = \left(\frac{A_s}{22.13}\right)^{0.6} \left(\frac{\sin\alpha}{0.0896}\right)^{1.3} \tag{1}$$

where, A_s depicts the specific watershed area and α is the estimated slope gradient (degree).

Stream power index (SPI) could be implemented to show potential flow erosion at a specific location of the basin (Moore et al. 1986) (Fig. 3g). Further, Topographic Wetness Index (TWI) was taken into account in this investigation. TWI denotes the spatial changes of soil moisture (Moore et al. 1986) (Fig. 3h).

Distance from rivers and river density are two crucial GCFs that affect the groundwater potentiality (Naghibi et al. 2015). These two layers were calculated in ArcGIS 10.4 using Euclidean distance and line density functions. Concerning the distance from rivers, 100 m-intervals were regarded, which was then classified into five groups (Fig. 3i). Rivers density map was partitioned into four categories by natural break classification method (Fig. 3j).

Land use map was produced by implementing Landsat 8/ enhance thematic mapper plus (ETM+) images for the year 2015 based on a likelihood algorithm. The land use map contained five different land use classes of the orchard, residential area, rangeland, agriculture, and forest (Fig. 3k).

Geology is composed of three GCFs including lithological classes, and fault-related factors such as distance and density maps (Naghibi et al. 2016). After investigating the fault layer of the studied region, it was found that only a tiny portion of the studied region is affected by fault; therefore, fault-related factors were not considered in the current research. Based on a 1:100,000-scale geological map, the geological units were partitioned into thirteen units including groups 1 to 13 (Table 1) (Fig. 3l).

2.2. Modelling process

In this section, a description of the models is presented and then, the process of applying a novel statistical- data mining model (EBF-BRT) is explained.

2.2.1. Evidential Belief Function (EBF)

The EBF model is developed based on the Dempster–Shafer approach of evidence (Dempster 1967; Shafer 1976), which includes uncertainty (Unc), belief (Bel), plausibility (Pls), and disbelief (Dis) that change from 0 to 1 (Carranza and Hale 2003). This model has a relative flexibility and is able to work with uncertain conditions (Nampak et al. 2014). In the Dempster–Shafer theory, Bel and Pls define the lower and upper probabilities of generalized Bayesian, respectively (Nampak et al. 2014). Therefore, it can be inferred that Bel is greater than or equal to Pls. Unc could be calculated by differentiating Pls and Bel values (Naghibi and Pourghasemi 2015). Based on the evidential data, disbelief depicts the belief in the false proposition. For calculating the Bel value, first, a frame of discernment could be calculated (Dempster 1967; Shafer 1976; Pourghasemi and Beheshtirad 2015):

$$m: 2^{\Theta} = \{\phi, T_P, \overline{T_P}, \Theta\} \quad with\ \Theta = \{S_P, \overline{S_P}\} \tag{2}$$

where, T_P shows the pixels that include springs, $\overline{T_P}$ shows the pixels that do not include springs, and ϕ represents empty set.

From Equation (1), the bel function could be computed as follows (Park 2011; Pourghasemi and Beheshtirad 2015):

$$\left[\lambda(S_P)_{A_{ij}}\right] = \left[\frac{N(S \cap A_{ij})}{N(S)}\right] / \left[\left(N\left(A_{ij} - N\left(S \cap A_{ij}\right)\right)\right) / [N(P) - N(S)]\right] \tag{3}$$

$$Bel = \left[\frac{\lambda(S_P)_{A_{ij}}}{\sum \lambda(S_P)_{A_{ij}}}\right] \tag{4}$$

where, $N(S \cap A_{ij})$ denotes density of spring pixels incidence in A_{ij}, $N(S)$ denotes the total density of all springs in the studied basin, $N(A_{ij})$ represents the density of pixels in A_{ij}, and N(P) is the density of pixels in the whole studied basin. More descriptions and information about EBF algorithm could be found in Carranza and Hale (2003).

2.2.2. The novel statistical- data mining ensemble model

The BRT is a data mining/machine learning approach, which comprises of both decision trees and boosting techniques and could be employed for both regression and classification issues (Youssef et al. 2015). It aims to increase the efficacy as well as prediction capability of a single methods by combining several fitted models (Naghibi et al. 2016). Boosting is applied in order to combine the results of the decision trees, which is similar to model averaging. There are some parameters that require optimizing in this model such as a number of trees, shrinkage (or learning rate), and interaction depth. Shrinkage or learning rate defines the importance of trees in the built model (Naghibi et al. 2016). Interaction depth or complexity determines the number of nodes in trees.

The BRT model can be explained as follows (Elith et al. 2008; Naghibi et al. 2016):

Starting weights to be equal to f_i =1/n

For m=1 to iteration classifier $C_{m)}$:

1. Run classifier C_m to the weighted data,
2. Calculate misclassification rate r_m,
3. Consider the classifier weight $\alpha_m \log\left(\frac{(1-r_m)}{r_m}\right)$,
4. Recalculation of weights $w_i = w_i exp\left(\alpha_m I(y_i \neq C_m)\right)$,

Finally, the majority vote can be obtained by: $sign = [\sum_{m-1}^{M} \alpha_m C_m(X)]$

It is noted that the best set of parameters in BRT were selected by using accuracy index and Cohen's kappa index, which can be calculated as below:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (5)$$

$$\text{Kappa} = \frac{\text{P}_{\text{obs}}-\text{P}_{\text{exp}}}{1-\text{P}_{\text{obs}}} \quad (6)$$

$$\text{P}_{\text{obs}} = \text{TP} + \text{TN}/\text{n} \quad (7)$$

$$\text{P}_{\text{exp}} = (\text{TP} + \text{FN})(\text{TP} + \text{FP}) + (\text{FP} + \text{TN})(\text{FN} + \text{TN})/\sqrt{\text{N}} \quad (8)$$

where, n is the ratio of cells, which is correctly categorized, and N shows the number of total training cells. TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively (Naghibi and Moradi Dashtpagerdi 2016).

To apply a novel statistical- data mining ensemble model, first, EBF model was applied and belief values were assigned to different classes of the GCFs. Then, new maps of each factor were produced by lookup function in ArcGIS 10.4. A new dataset was provided for training of the data mining model (i.e. BRT). In this dataset, 1 was assigned to spring and 0 was assigned to non-spring locations. It is noted that the non-spring locations were randomly defined using ArcGIS 10.4. Using the new training dataset and new GCFs' layers with Bel values, BRT model was conducted using R open source software by the gbm package (Ridgeway, 2015). The BRT model was run using a 10-fold cross-validation deemed to be a sufficient number of the run for optimization of the assigned parameters. It needs to be clarified that the GPMs by EBF and BBF-BRT are classified into four classes of low, moderate, high, and very high by natural break classification method (Naghibi et al. 2018).

3. Results and Discussion

3.1. Evidential belief function

The results of the EBF model are presented in Table 2 where the values of the Bel, Dis, and Unc are reported. As it was mentioned in the methodology section, a class with high Bel value has a high potential for the occurrence of the event, which in this case is the existence of the spring (Nampak et al. 2014; Pourghasemi and Beheshtirad 2015). Based on the results, it can be observed that there is an inverse relationship between slope angle and the Bel value, which means that the groundwater potential decreases with the increase in slope angle. Regarding the results of slope aspect, flat and north-east classes show the highest Bel values. On the contrary, south-east and south-west classes have Bel value of zero, which indicates their low potential of spring incidence. This finding can be related to the less sunshine duration over the north slope aspects in the northern hemisphere. In the case of altitude, the results indicated that an inverse relationship exists between GCF and spring incidence. In lower altitudes, water has concentrated near the rivers and therefore, wetness index is higher in these areas that can result in the higher potential of groundwater. The flat characteristic of the plan curvature had the highest Bel value (Bel=0.54). The highest amount of Bel was observed in (-0.001) - (0.001) category of profile curvature. An inverse relationship was observed between the slope length and spring incidence. In the case of SPI, the results indicated that < 200 and 400-600 categories have the highest Bel value of 0.34 and 0.24, respectively. The findings of TWI signified a direct relationship between TWI and spring incidence. Regarding the distance from rivers, an inverse relationship between the distance from river and the spring occurrence was observed. Regarding river density, 0.86-1.46 class has the highest Bel value of 0.40 followed by >1.46, 0.31-0.86, and <0.31 classes. The modeling results with respect to land use showed that agriculture has the highest Bel value, followed by forest and rangeland. Regarding lithology, the highest values of Bel were observed for Group 2 and Group 10 with values of 0.22 and 0.17, respectively.

Overall, these findings signified that a direct relationship exists between spring incidence and TWI factor. On the contrary, an inverse relationship was observed between the groundwater potentiality and three GCFs including altitude, slope length, and distance from rivers. Naghibi and Pourghasemi (2015) obtained the same relationship between altitude, TWI, and distance from rivers and spring occurrence. However, in some other factors such as LS, our findings differ from theirs. These differences can be due to the different properties of the studied regions (i.e. topographical and hydrological characteristics). Furthermore, the results of the EBF-BRT model revealed that the distance from rivers, lithology, river density, and plan curvature had the highest importance in the groundwater potential mapping of the studied basin.

GPM produced by the EBF model in the current study is presented in Figure 4a and Table 3. It should be noted that the final EBF map was obtained by summing all the Bel values. Based on the findings, the value of GPM in this model ranges from 0.88 to 5.29. Low, moderate, high, and very high potential categories composed 34, 28, 20, and 18% of the studied basin, respectively.

3.2. The novel statistical- data mining ensemble model

The findings of the application of BRT algorithm are presented in Figure 5. The final BRT model was applied with minimum terminal node size of 10, shrinkage value of 0.1, 50 number of trees, and interaction depth of 1 (Accuracy index = 0.66 and Cohen's Kappa index = 0.33). The contribution of the GCFs to the modelling process is presented in Figure 6. The results indicated that the distance from rivers, lithology, river density, and plan curvature have the highest contribution to groundwater potential estimated by EBF-BRT model (Fig. 6). The land use and profile curvature showed the lowest contribution and SPI showed no effect on groundwater potential. The GPM obtained from EBF-BRT method is presented in Figure 4b and Table 3. The GPM produced by EBF-BRT model resulted in low, moderate, high, and very high potential categories, which composed 32, 28, 25, and 15% of the studied basin, respectively.

3.3. Validation and verification of the GPMs

This section includes two steps: (i) validation of the maps using the validation dataset and ROC curve and (ii) verifying the results by taking the observed spring discharges into account.

Chung and Fabbri (2003) stated that the validation is regarded as a very necessary stage in the modeling procedure. To do so, the ROC curve was implemented to define the accuracy of the GPMs produced by EBF and EBF-BRT models. The GPMs were verified employing training and validation datasets. The area under the curve of ROC varies between 0.5 and 1 (Sangchini et al. 2016; Hong et al. 2017; Kalantar et al. 2018). A larger area under the curve of ROC denotes higher efficacy of the models in spatial modeling (Jaafari and Gholami 2017; Pham et al. 2018) such as groundwater potential mapping. Figure 7 presents the prediction performance of the produced GPMs by EBF and EBF-BRT models implementing ROC curve. Accordingly, the area under the curve of ROC for validation dataset was defined as 75.5 and 82.1% for EBF and EBF-BRT models, respectively. Further, area under ROC curve for training dataset was calculated as 77.2 and 83% for EBF and EBF-BRT, respectively. It was assumed that the values of more than 70% indicate an acceptable performance of the model (Naghibi et al. 2016).

To verify the resulted groundwater potential map of the basin, the spring discharge record was used. For this, the observed discharge values higher than the median discharge, 0.75 L/s, were selected for models' verification. Distribution of the selected springs in different potential zones produced by EBF and EBF-BRT is presented in Table 4. As can be seen in the table, among 47 high-discharge springs, 15 and 16 springs were located in the very high potential zone produced by EBF and EBF-BRT, respectively. According to the modeling results, very few springs with high-discharge were located in the low potential zone (Table 4). The distribution of the high-discharge springs in the identified groundwater potential zones, as well as the computed area under ROC curve, confirm the satisfying performance of the models in this study.

3.4. Performance comparison

The findings of this study indicated superior performance of the EBF-BRT to EBF in producing groundwater potential maps. Therefore, it can be observed that making the ensemble EBF-BRT model increased the efficacy of the GPM in this research. The validation results also indicated an acceptable capability of the EBF model in producing GPM. Naghibi and Pourghasemi (2015) and Nampak et al. (2014) employed EBF model for producing GPMs. Their results depicted acceptable performance of the EBF, which is in agreement with the findings of this study. Other researchers have employed different methods to improve the performance of the EBF model. Tien Bui et al. (2015) employed an EBF-fuzzy logic hybrid method for modelling landslide. Their findings showed the higher efficacy of the hybrid method relative to EBF model. In another research, Pourghasemi and Kerle (2016) employed an EBF-random forest model to map landslide susceptibility, and their findings depicted a better performance of the EBF-random forest model than EBF model. In a related work, Naghibi et al. (2017a) used and ensemble model comprised of four data mining models and frequency ratio. Their results indicated a better performance of the ensemble model by the reduction of overfitting. Moreover, Naghibi et al. (2017b) used a genetic algorithm to optimize random forest as an ensemble model, and this combination yielded a better performance. In the current research, the more accurate results of the EBF-BRT model could be due to the strong features of the single BRT and EBF models. The BRT model is capable of coping with nonlinear relationships (Naghibi et al. 2016). BRT applies a combination of boosting and regression techniques, which results in a better performance (Elith et al. 2008). The EBF, on the other hand, is proved to be a robust model for managing uncertainties in spatial modelling and can deal with missing values (Tangestani and Moore 2002).

4. Conclusions

Groundwater potential mapping has been considered as an important aspect of groundwater-related studies and has attracted many scholars worldwide. In this study, a novel ensemble EBF-BRT model was introduced, and its performance was assessed in groundwater potential mapping. EBF-BRT model was applied using a training dataset of the belief values extracted from EBF model results. Using the ROC curve, performance of the EBF and EBF-BRT models was evaluated. The findings indicated that EBF-BRT model yielded better performance than simple EBF model. Therefore, it can be concluded that application of the BRT model can enhance the prediction strength of the EBF model. However, both of the models had acceptable performance in this study. The better performance of EBF-BRT model could be due to stronger features of the BRT model such as its capability to cope with phenomena in which there are nonlinear relationships. Regarding the conditioning factors, it was observed that the distance from rivers, lithology, rivers density, and plan curvature have the highest importance in the GPMs by EBF-BRT model. Considering the findings of this study, the implemented methodology can be recommended for other areas with similar geological and hydrological setting. GPMs can be regarded as a guiding tool for freshwater professionals to properly manage land and water resources. GPMs would also provide superior insight of groundwater condition in various parts of a basin that would subsequently lead to efficient exploitation of groundwater.

The GPMs can be employed for functional water resources management especially through land use planning. Those activities with high water requirements, i.e. irrigated agriculture, can be located in areas with higher groundwater potential. However, the rate of exploitation should be monitored and controlled. The GPMs can also support decision making processes in the land use and water resources planning that ultimately leads to environmental sustainability, which is very crucial in the Middle Eastern countries such as Iran. It is evident that overexploitation issue causes many problems for people and the government in most of the aquifers in Iran. The outputs of this study could be channeled to the relevant agencies/organizations and result in a better aquifer management strategy through defining the places where are more groundwater productive. A better land use planning could lead to lower pressure on aquifers. However, it is the first step and there need to more remediation steps such as artificial recharge through water harvesting, and flood spreading.

References

Abeare SM (2009) Comparisons of boosted regression tree, glm and gam performance in the standardization of yellowfin tuna catch-rate data from the gulf of mexico lonline fishery. LSU Master's Theses. 2880

Carranza JEM, Hale M (2003) Evidential belief functions for data-driven geologically constrained mapping of gold potential, Baguio district, Philippines. Ore Geology Reviews 22: 117–132

Chezgi J, Pourghasemi HR, Naghibi SA, Moradi HR, Kheirkhah Zarkesh M (2015) Assessment of a spatial multi-criteria evaluation to site selection underground dams in the Alborz Province, Iran. Geocarto Int 31: 1–19. doi:10.1080/10106049.2015.1073366

Chung JF, Fabbri AG (2003) Validation of spatial prediction models for landslide hazard mapping. Nat Hazards 30(3): 451-472

Corsini A, Cervi F, Ronchetti F (2009) Weight of evidence and artificial neural networks for potential groundwater spring mapping: an application to the Mt. Modino area (Northern Apennines, Italy). Geomorphology 111: 79–87

Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping: Ann. Math. Stat 38: 325–339

Dempster AP (1968) Generalization of Bayesian inference. J. Roy. Stat. Soc., Ser. B 30: 205–247

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813. doi: 10.1111/j.1365-2656.2008.01390.x

Elmahdy SI, Mohamed MM (2014) Probabilistic frequency ratio model for groundwater potential mapping in Al Jaww plain, UAE. Arab. J. Geosci. doi:10.1007/s12517-014-1327-9

Fitts CR (2002) Groundwater Science. Academic Press

Freeze RA, Cherry JA (1979) Groundwater, Prentice-Hall, Engle- wood Cliffs, N. J., XVI, 604 pp

Geology Survey of Iran (GSI) (1997) http://www.gsi.ir/Main/Lang_en/index.html.

Golkarian A, Naghibi SA, Kalantar B, Pradhan B (2018) Groundwater potential mapping using C5. 0, random forest, and multivariate adaptive regression spline models in GIS. Environmental monitoring and assessment, 190(3), p.149.

Hong, H., Naghibi, S.A., Dashtpagerdi, M.M., Pourghasemi, H.R. and Chen, W., 2017. A comparative assessment between linear and quadratic discriminant analyses (LDA-QDA) with frequency ratio and weights-of-evidence models for forest fire susceptibility mapping in China. Arabian Journal of Geosciences, 10(7), p.167.

Jaafari, A. and Gholami, D.M., 2017. Wildfire hazard mapping using an ensemble method of frequency ratio with Shannon's entropy. Iranian Journal of Forest and Poplar Research, 25(2).

Kalantar B, Pradhan B, Naghibi SA, Motevalli A Mansor S (2018) Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomatics, Natural Hazards and Risk, 9(1), pp.49-69.

Lee M-J, Choi J-W, Oh H-J, Won J-S, Park I, Lee S (2012) Ensemble-based landslide susceptibility maps in jinbu area, Korea. Environ. Earth Sci. 67: 23–37. doi:10.1007/s12665-011-1477-y

Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol. Processes 5(1): 3- 30

Moore ID, Burch GJ (1986) Sediment Transport Capacity of Sheet and Rill Flow: Application of Unit Stream Power Theory. Water Resour. Res. 22: 1350–1360. doi:10.1029/WR022i008p01350

Mousavi SM, Golkarian A, Naghibi SA, Kalantar B, Pradhan B (2017) GIS-based groundwater spring potential mapping using data mining boosted regression tree and probabilistic frequency ratio models in Iran. AIMS Geosciences, 3(1): 91-115.

Nampak H, Pradhan B, Manap MA (2014) Application of GIS based data driven evidential belief function model to predict groundwater potential zonation. J. Hydrol. 513: 283-300, doi:10.1016/j.jhydrol.2014.02.053

Naghibi SA, Pourghasemi HR, Pourtaghi ZS, Rezaei A, (2015) Groundwater qanat potential mapping using frequency ratio and Shannon's entropy models in the Moghan watershed, Iran. Earth Sci. Informatics 8: 1–16. doi:10.1007/s12145-014-0145-7

Naghibi SA, Pourghasemi HR (2015) A Comparative Assessment Between Three Machine Learning Models and Their Performance Comparison by Bivariate and Multivariate Statistical Methods in Groundwater Potential Mapping. Water Resour. Manag. 29(14): 5217-5236.

Naghibi SA, Pourghasemi HR, Dixon B (2016) GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ. Monit. Assess. 188: 44. doi:10.1007/s10661-015-5049-6

Naghibi SA, Moradi Dashtpagerdi M (2016) Evaluation of four supervised learning methods for groundwater spring potential mapping in Khalkhal region (Iran) using GIS-based features. Hydrogeol. J. 25(1):169–189.

Naghibi SA, Moghaddam DD, Kalantar B, Pradhan B, Kisi O (2017a) A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping. J. Hydrol. 548: 471–483. doi:10.1016/j.jhydrol.2017.03.020

Naghibi SA, Ahmadi K, Daneshi A (2017b) Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. Water Resour. Manag. 31(9): 2761–2775

Naghibi, S.A., Pourghasemi, H.R. and Abbaspour, K., 2018. A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS. Theoretical and Applied Climatology, 131(3-4), pp.967-984.

Oh H-J, Kim Y-S, Choi J-K, Park E, Lee S (2011) GIS mapping of regional probabilistic groundwater potential in the area of Pohang City, Korea. J. Hydrol. 399:158–172. doi:10.1016/j.jhydrol.2010.12.027

Ozdemir A (2011a) GIS-based groundwater spring potential mapping in the Sultan Mountains (Konya, Turkey) using frequency ratio, weights of evidence and logistic regression methods and their comparison. J. Hydrol. 411: 290–308. doi:10.1016/j.jhydrol.2011.10.010

Ozdemir A (2011b) Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). J. Hydrol. 405: 123–136. doi:10.1016/j.jhydrol.2011.05.015

Pham, B.T., Jaafari, A., Prakash, I. and Bui, D.T., 2018. A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling. Bulletin of Engineering Geology and the Environment, pp.1-22.

Pourghasemi HR, Beheshtirad M (2015) Assessment of a data-driven evidential belief function model and GIS for groundwater potential mapping in the Koohrang Watershed, Iran. Geocarto Int 30:662–685. doi: 10.1080/10106049.2014.966161

Pourghasemi HR, Kerle N (2016) Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran. Environ. Earth Sci. 75: 185. doi:10.1007/s12665-015-4950-1

Pourtaghi ZS, Pourghasemi HR (2014) GIS-based groundwater spring potential assessment and mapping in the Birjand Township, southern Khorasan Province, Iran. Hydrogeology Journal, 22(3): 643-662. http://doi.org/10.1007/s10040-013-1089-6

Rahmati O, Melesse AM (2016) Application of Dempster– Shafer theory, spatial analysis and remote sensing for groundwater potentiality and nitrate pollution analysis in the semi-arid region of Khuzestan, Iran. Science of The Total Environment, 568(15): 1110-1123. doi:10.1016/j.scitotenv.2016.06.176

Rahmati O, Pourghasemi HR, Melesse AM (2016) Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran. Catena 137: 360–372. doi:10.1016/j.catena.2015.10.010

Razandi Y, Pourghasemi HR, Neisani NS, Rahmati O (2015) Application of analytical hierarchy process, frequency ratio, and certainty factor models for groundwater potential mapping using GIS. Earth Sci Informatics 8:867–883. doi: 10.1007/s12145-015-0220-8

Ridgeway, G., 2006. gbm: Generalized boosted regression models. R package version, 1(3), p.55.

Sangchini, E.K., Emami, S.N., Tahmasebipour, N., Pourghasemi, H.R., Naghibi, S.A., Arami, S.A. and Pradhan, B., 2016. Assessment and comparison of combined bivariate and AHP models with logistic regression for landslide susceptibility mapping in the Chaharmahal-e-Bakhtiari Province, Iran. Arabian Journal of Geosciences, 9(3), p.201.

Shafer G (1976) A mathematical Theory of Evidence. Princeton Univ. Press, Priceton, NJ

Tehrany MS, Pradhan B, Jebur MN (2013) Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J. Hydrol. 504: 69–79. doi:10.1016/j.jhydrol.2013.09.034

Tehrany MS, Pradhan B, Jebur MN (2014) Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. J. Hydrol. 512: 332–343. doi:10.1016/j.jhydrol.2014.03.008

Tahmassebipoor N, Rahmati O, Noormohamadi F, Lee S (2016) Spatial analysis of groundwater potential using weights-of-evidence and evidential belief function models and remote sensing. Arab. J. Geosci. 9: 79. doi:10.1007/s12517-015-2166-z

Tien Bui D, Pradhan B, Revhaug I, et al (2015) A novel hybrid evidential belief function-based fuzzy logic model in spatial prediction of rainfall-induced shallow landslides in the Lang Son city area (Vietnam). Geomatics, Nat Hazards Risk 5705:1–30. doi: 10.1080/19475705.2013.843206

Umar Z, Pradhan B, Ahmad A, Jebur MN, Tehrany MS (2014) Earthquake induced landslide susceptibility mapping using an integrated ensemble frequency ratio andlogistic regressionmodels inWestSumatera Province, Indonesia. Catena 118: 124–135. http://dx.doi.org/10.1016/j.catena.2014.02.005.

Youssef AM, Pourghasemi HR, Pourtaghi ZS, Al-Katheeri MM (2015) Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides. doi:10.1007/s10346-015-0614-1

Zabihi M, Pourghasemi HR, Pourtaghi ZS, Behzadfar M (2016) GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran. Environ. Earth Sci. 75: 665. doi:10.1007/s12665-016-5424-9

Table 1. Lithology characteristics of Lordegan Basin, Iran.

Class	Lithology characteristics
Class 1	Anhydrite, salt, grey, and red marl alternating with anhydrite, argillaceous limestone and limestone
Class 2	Blue and purple shale and marl inter bedded with the argillaceous limestone
Class 3	Bluish grey marl and shale with subordinate thin- bedded argillaceous-limestone
Class 4	Brown to grey, calcareous, feature- forming sandstone and low weathering, gypsum- veined, red marl and siltstone
Class 5	Low level piedmont fan and valley terraces deposit
Class 6	Low weathering grey marls alternating with bands of more resistant shelly limestone
Class 7	Pale red marl, marlstone, limestone, gypsum and dolomite
Class 8	Cream to brown- weathering, feature- forming, well- jointed limestone with intercalations of shale
Class 9	Dark red, medium- grained arkosic to subarkosic sandstone and micaceous siltstone
Class 10	Limestone, dolomite, dolomitic limestone and thick layers of anhydrite in alternation with dolomite in middle part
Class 11	Massive, shelly, cliff- forming partly anhydrite limestone
Class 12	Undivided Bangestan group, mainly limestone and shale, albian to companian
Class 13	Undivided Eocene rock

Table 2 Spatial relationship between GCFs and springs using EBF model.

Factor	Class	% of pixels in domain	No. of Springs	Bel	Dis	Unc
Slope Angle (Degree)	0-5	29.46	38	0.54	0.15	0.31
	5-15	22.58	20	0.37	0.23	0.41
	15-30	35.25	8	0.09	0.34	0.57
	>30	12.71	0	0.00	0.29	0.71
Slope Aspect	Flat	8.70	10	0.22	0.19	0.59
	North	13.59	8	0.11	0.21	0.68
	Northeast	14.69	13	0.17	0.19	0.64
	East	8.65	4	0.09	0.21	0.70
	Southeast	8.66	6	0.00	0.00	1.00
	South	10.47	4	0.07	0.21	0.72
	Southwest	13.60	10	0.00	0.00	1.00
	West	11.17	8	0.14	0.00	0.86
	Northwest	10.47	3	0.06	0.00	0.94
Altitude (m)	<1400	1.63	4	0.61	0.24	0.15
	1400-1900	40.15	36	0.22	0.19	0.58
	1900-2500	45.22	25	0.14	0.29	0.57
	2500-3000	9.22	1	0.03	0.28	0.70
	>3000	3.79	0	0.00	0.00	1.00
Plan Curvature (100/m)	Concave	29.54	16	0.28	0.36	0.36
	Flat	37.60	39	0.54	0.22	0.24
	Convex	32.86	11	0.18	0.42	0.41
Profile curvature (100\m)	< (-0.001)	35.30	23	0.33	0.34	0.33
	(-0.001)-(0.001)	32.79	30	0.46	0.27	0.27
	> (0.001)	31.91	13	0.21	0.39	0.40
Slope Length (m)	<20	38.46	40	0.41	0.16	0.43
	20-40	16.73	12	0.29	0.25	0.47
	40-60	14.23	8	0.22	0.26	0.52
	>60	30.58	6	0.08	0.33	0.59
Stream Power Index	<200	30.62	27	0.34	0.21	0.45
	200-400	12.96	7	0.21	0.26	0.54
	400-600	9.55	6	0.24	0.25	0.51
	>600	46.87	26	0.21	0.28	0.50
Topographic Wetness Index	<8	19.44	2	0.05	0.39	0.56
	8-12	56.23	32	0.29	0.38	0.33
	>12	24.33	32	0.66	0.22	0.12
Distance from Rivers (m)	<100	4.69	27	0.71	0.17	0.12
	100-200	4.15	5	0.15	0.27	0.58
	200-300	4.10	2	0.06	0.28	0.66
	300-400	4.03	1	0.03	0.28	0.69
	>400	83.04	31	0.00	0.00	1.00
River Density (Km/Km^2)	<0.31	60.74	18	0.08	0.42	0.50
	0.31-0.86	11.82	8	0.18	0.23	0.60
	0.86-1.46	21.94	33	0.40	0.14	0.45
	>1.46	5.50	7	0.34	0.21	0.45
Land use	Agriculture	24.58	33	0.61	0.16	0.23
	Forest	30.83	11	0.16	0.30	0.54
	Orchard	0.04	0	0.00	0.25	0.75
	Rangeland	43.99	22	0.23	0.29	0.48
	Residential area	0.57	0	0.00	0.00	1.00
Lithology	Group 1	3.25	4	0.16	0.07	0.76
	Group 2	4.22	7	0.22	0.07	0.71

Group 3	0.22	0	0.00	0.08	0.92
Group 4	4.44	5	0.15	0.07	0.78
Group 5	33.32	26	0.10	0.07	0.82
Group 6	8.23	2	0.03	0.08	0.89
Group 7	1.53	0	0.00	0.08	0.92
Group 8	28.52	17	0.08	0.08	0.84
Group 9	2.39	1	0.06	0.08	0.87
Group 10	1.60	2	0.17	0.08	0.76
Group 11	0.02	0	0.00	0.08	0.92
Group 12	1.40	0	0.00	0.08	0.92
Group 13	10.86	2	0.03	0.08	0.89

Table 3. Range and area of different classes of the groundwater potential map (GPM) produced by EBF model.

	EBF		**EBF-BRT**	
Class	Range of the values	Area %	Range of the values	Area %
Low	0.88-1.91	34	0-0.23	32
Moderate	1.91-2.60	28	0.23-0.41	28
High	2.60-3.41	20	0.41-0.61	25
Very high	3.41-5.29	18	0.61-0.96	15

Table 4. Distribution of the high-discharge springs in the identified groundwater potential zones.

Potential Zones	EBF		BRT	
	No. Spring	Spring (%)	No. Spring	Spring (%)
Low	8	17.02	4	8.52
Moderate	10	21.28	12	25.53
High	14	29.79	15	31.91
Very high	15	31.91	16	34.04

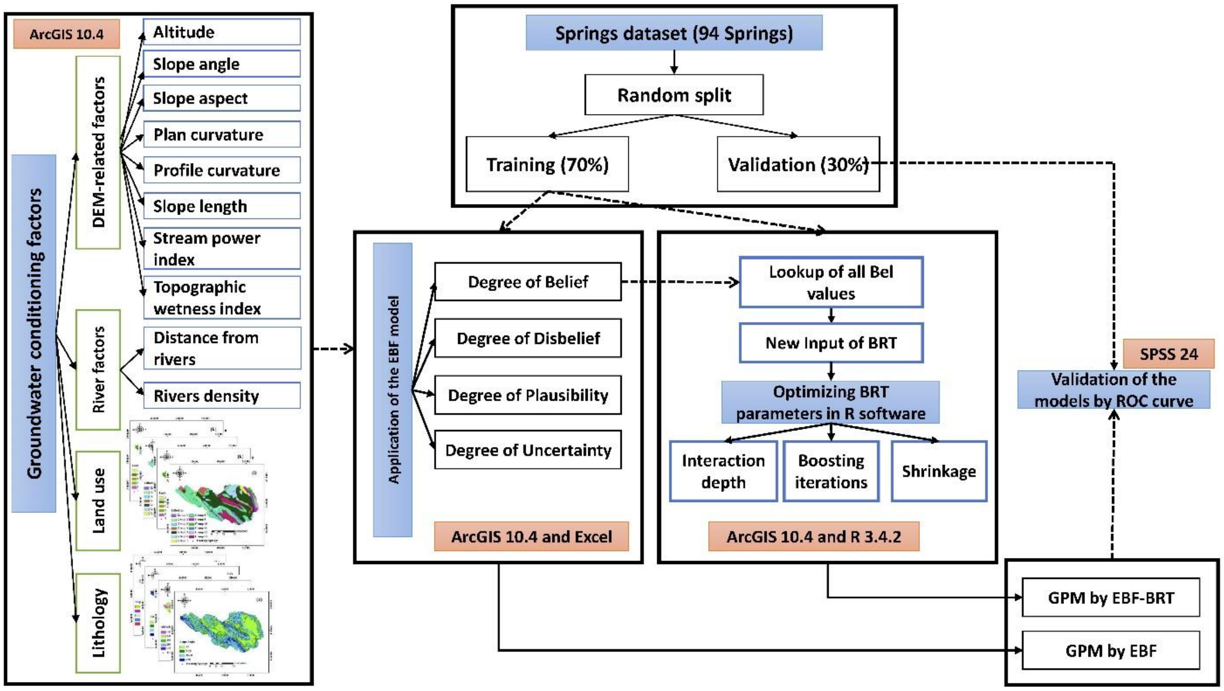


Figure 1. Flowchart of the methodology implemented in this study.

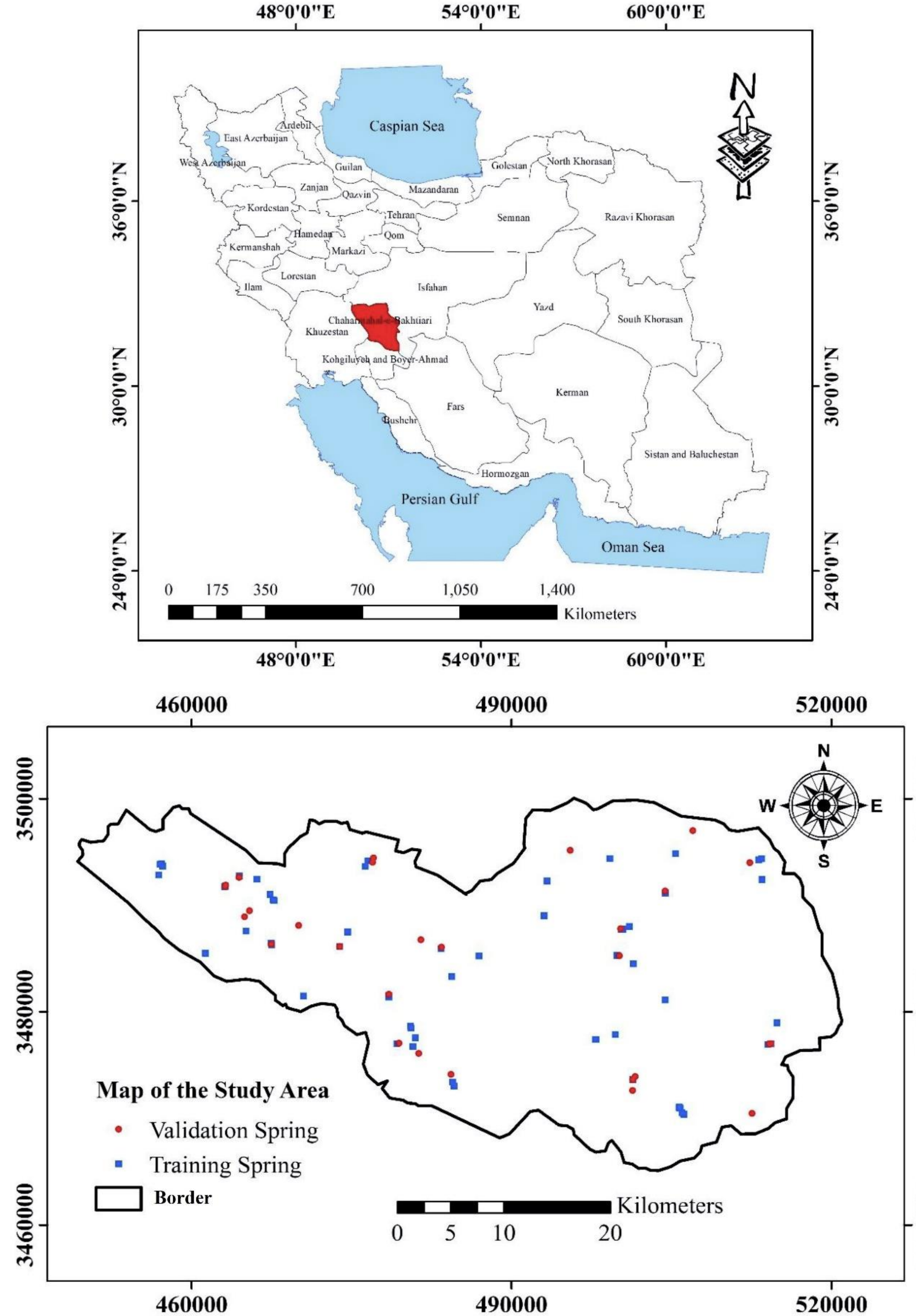
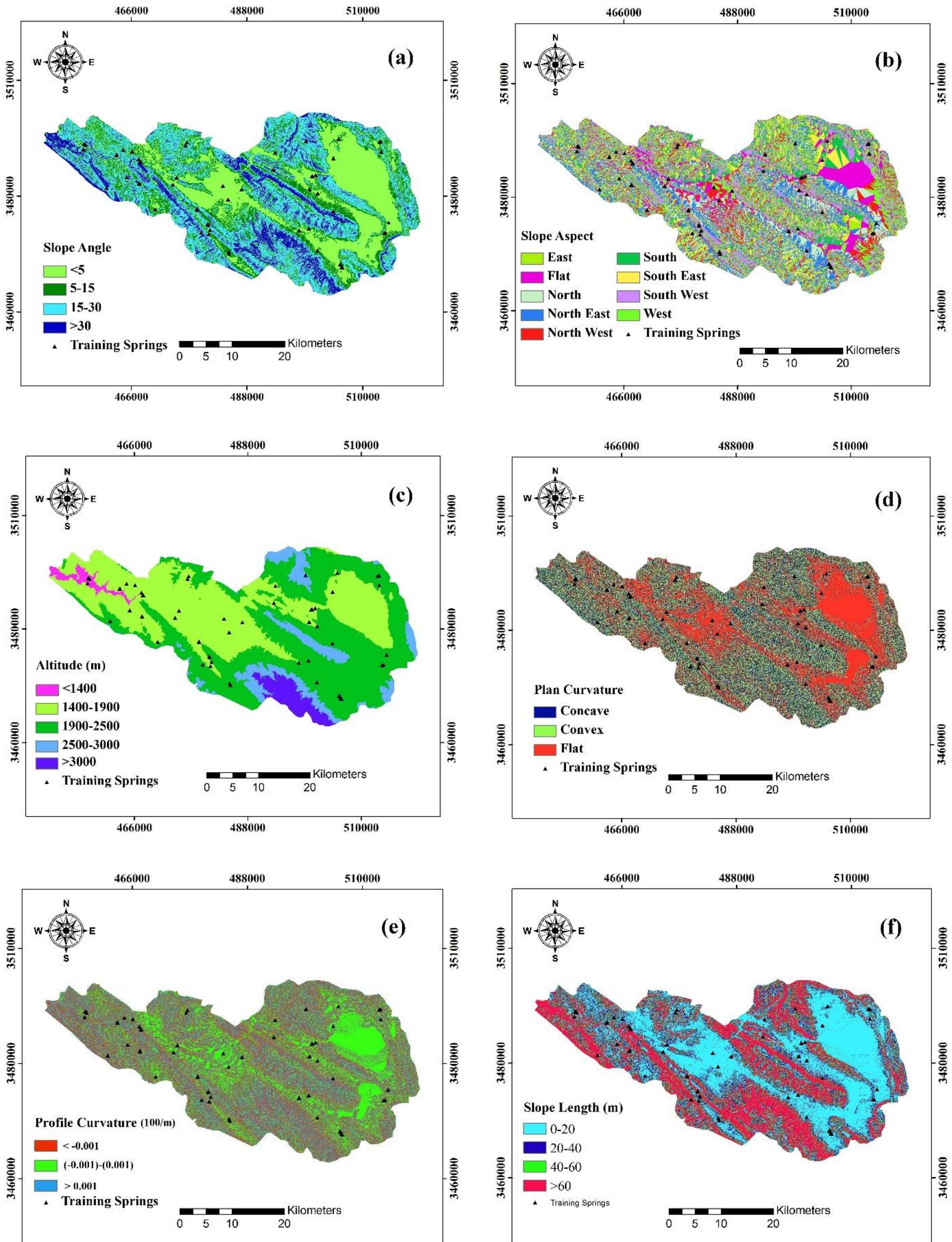


Figure 2. Location of the study area in Iran, training, and validation spring.


(a)
Slope Angle
<5
5-15
15-30
>30
Training Springs
(b)
Slope Aspect
East
Flat
North
North East
North West
South
South East
South West
West
Training Springs
(c)
Altitude (m)
<1400
1400-1900
1900-2500
2500-3000
>3000
Training Springs
(d)
Plan Curvature
Concave
Convex
Flat
Training Springs
(e)
Profile Curvature (100/m)
< -0.001
(-0.001)-(0.001)
> 0.001
Training Springs
(f)
Slope Length (m)
0-20
20-40
40-60
>60
Training Springs
Kilometers
0 5 10 20
466000
488000
510000
3460000
3480000
3510000
N
S
E
W

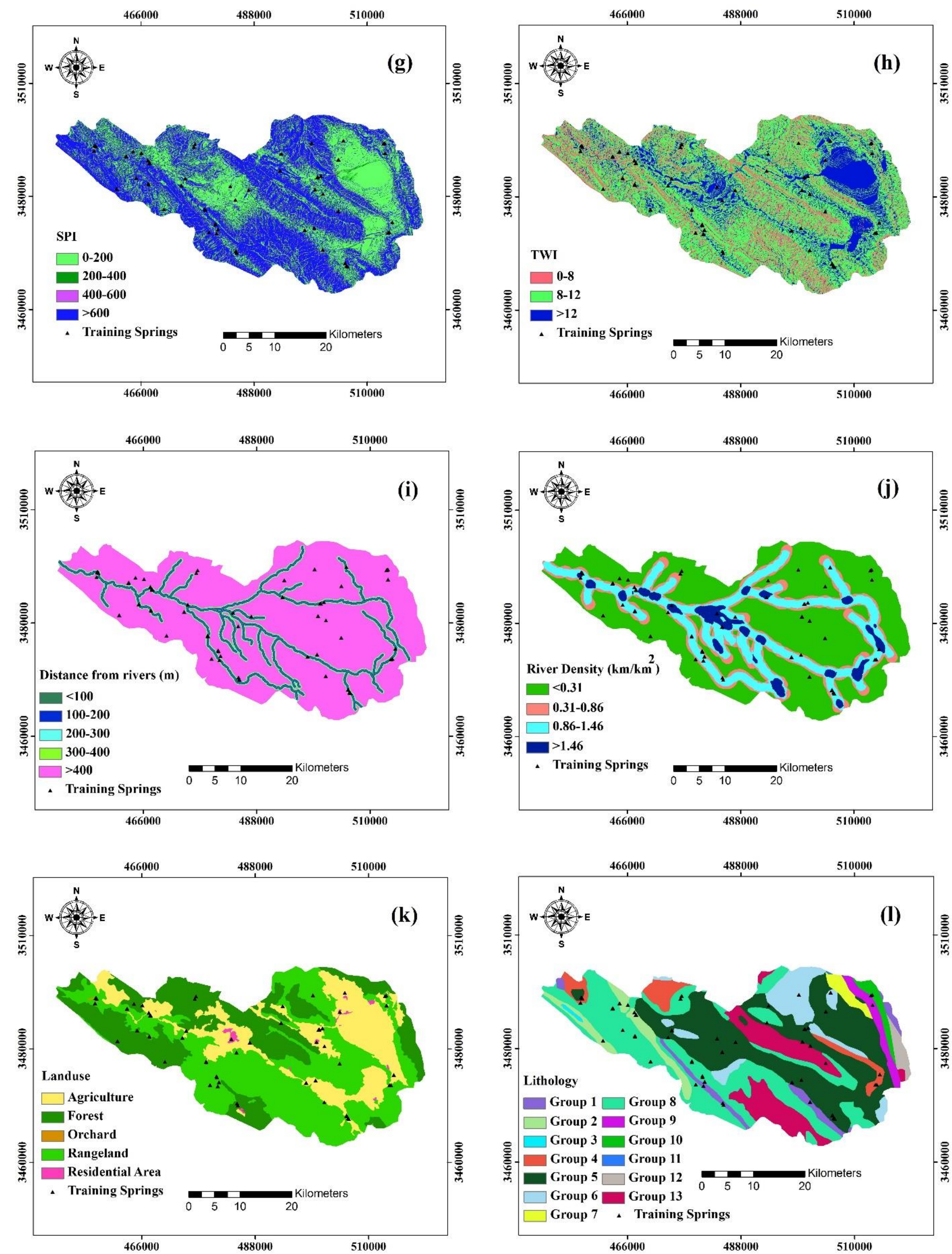
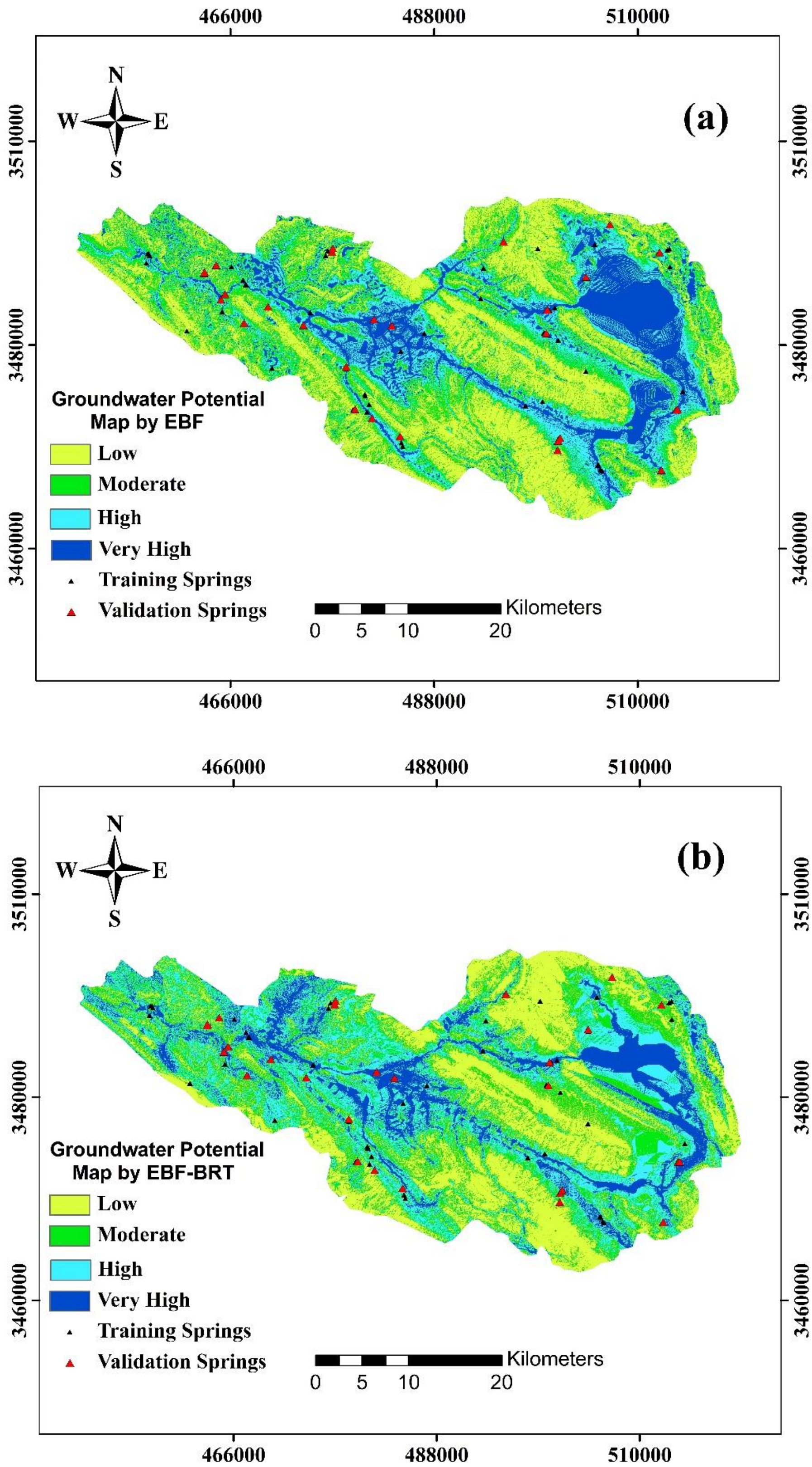

(g)
SPI
0-200
200-400
400-600
>600
Training Springs
Kilometers
0 5 10 20
(h)
TWI
0-8
8-12
>12
Training Springs
(i)
Distance from rivers (m)
<100
100-200
200-300
300-400
>400
Training Springs
(j)
River Density (km/km^2)
<0.31
0.31-0.86
0.86-1.46
>1.46
Training Springs
(k)
Landuse
Agriculture
Forest
Orchard
Rangeland
Residential Area
Training Springs
(l)
Lithology
Group 1
Group 2
Group 3
Group 4
Group 5
Group 6
Group 7
Group 8
Group 9
Group 10
Group 11
Group 12
Group 13
Training Springs
466000
488000
510000
3460000
3480000
3510000
N
S
E
W

Figure 3. The GCFs considered in this study (a) slope angle, (b) slope aspect, (c) altitude, (d) plan curvature, (f) profile curvature, (g) slope length, (h) stream power index, (i) topographic wetness index, (j) distance from rivers, (k) rivers density, (k) land use, and (l) lithology.



518 **Figure 4.** Groundwater potential map produced by (a) EBF and (b) EBF-BRT models.

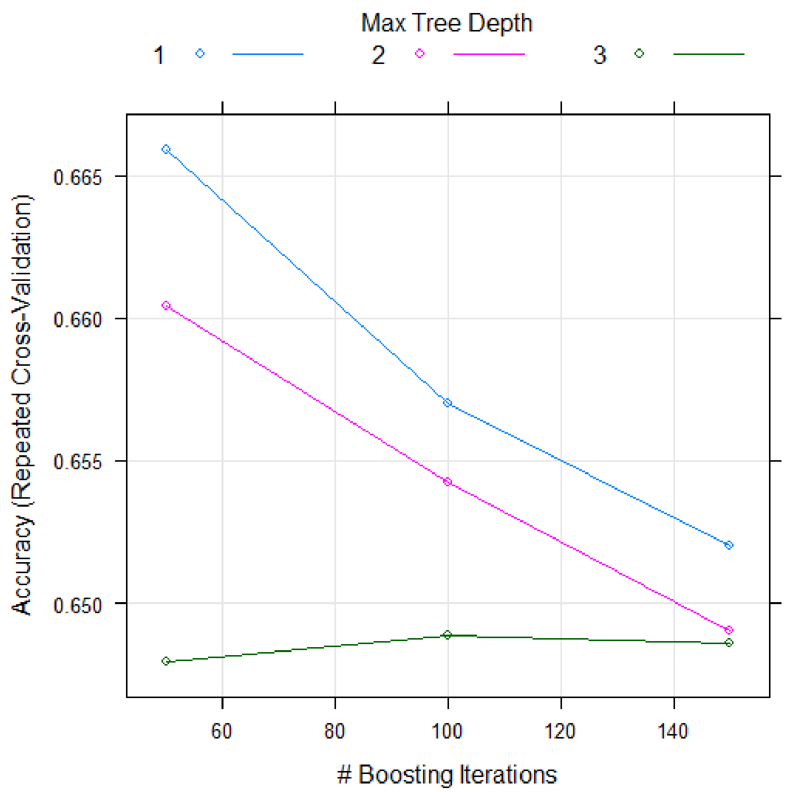


Figure 5. Results of the EBF-BRT application.

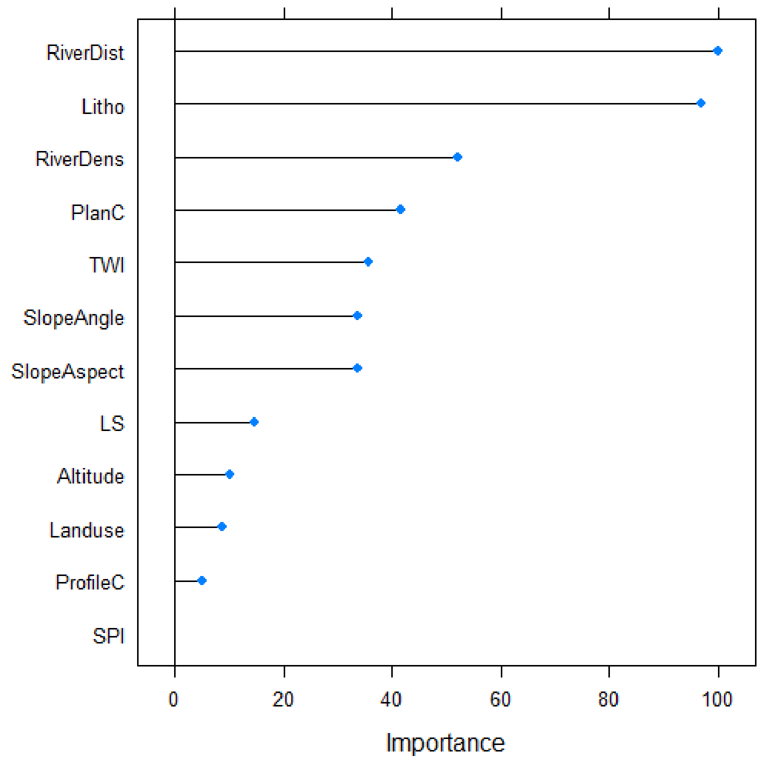


Figure 6. Importance of the groundwater conditioning factors (GCFs) in the BRT model (RiverDist: distance from rivers; Litho: lithology; RiverDens: rivers density; PlanC: plan curvature; TWI: TWI; SlopeAngle: slope angle; SlopeAspect: slope aspect; LS: LS; Altitude: altitude; Landuse: land usel; ProfileC: profile curvature; SPI: SPI).

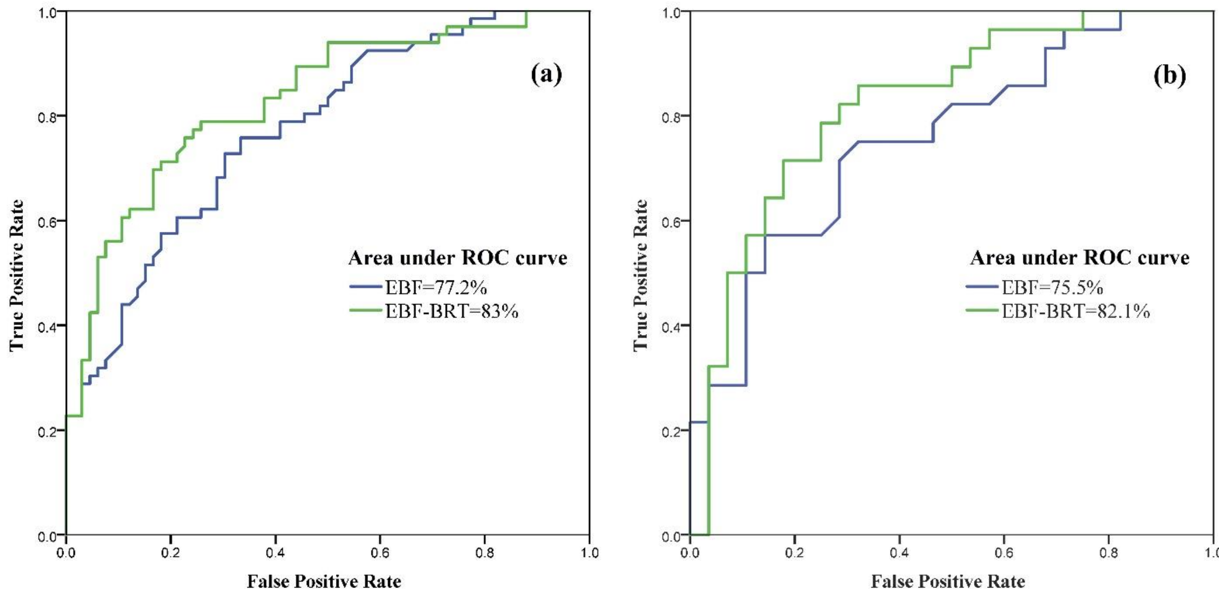


Figure 7. Receiver operating characteristics (ROC) curve calculated for the EBF and EBF-BRT models for training (a) and validation datasets (b), respectively.