



Groundwater potential mapping using a novel data-mining ensemble model

Mojtaba Dolat Kordestani¹ · Seyed Amir Naghibi² · Hossein Hashemi² · Kourosh Ahmadi³ · Bahareh Kalantar⁴ · Biswajeet Pradhan^{5,6}

Received: 12 January 2018 / Accepted: 5 August 2018 / Published online: 1 September 2018
© The Author(s) 2018

Abstract

Freshwater scarcity is an ever-increasing problem throughout the arid and semi-arid countries, and it often results in poverty. Thus, it is necessary to enhance understanding of freshwater resources availability, particularly for groundwater, and to be able to implement functional water resources plans. This study introduces a novel statistical approach combined with a 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, 12 groundwater-conditioning factors (GCFs), including topographical and hydrogeological factors, are prepared for the modeling process. The mentioned factors are employed in the application of the EBF model. Then, the EBF values of the GCFs are implemented as input to the BRT algorithm. The results of the modeling process are plotted to produce spring (groundwater) potential maps. To verify the results, the receiver operating characteristics (ROC) test is applied to the model's output. The findings of the test indicated that the areas under the ROC curves are 75 and 82% for the EBF and EBF-BRT models, respectively. Therefore, it can be inferred that the combination of the two techniques could increase the efficacy of these methods in groundwater potential mapping.

Keywords Geographic information system (GIS) · Groundwater management · Data mining · Iran

Introduction

Groundwater could be regarded as the water in the saturated parts of the Earth that fills the pore section of geologic formations and soil beneath the water table (Freeze and Cherry 1979). Groundwater has broad advantages over surface water as a resource, 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 meeting freshwater demand 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 resource indicators, while others used qanat and well locations. According to the literature, the frequency ratio (Oh et al. 2011; Pourtaghi and Pourghasemi 2014; Naghibi et al. 2015), weights-of-evidence

✉ Seyed Amir Naghibi
Amimaghbi2010@yahoo.com; Amir.naghibi@cme.lu.se

¹ University of Hormozgan, Jiroft University Scholarship, Department of Combat Desertification, Faculty of Rangeland and Watershed Management, Jiroft University, Jiroft, 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 RIKEN Center for Advanced Intelligence Project, Goal-Oriented Technology Research Group, Disaster Resilience Science Team, Tokyo 103-0027, Japan

⁵ Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney, Sydney, NSW 2007, Australia

⁶ Department of Energy and Mineral Resources Engineering, Choongmu-gwan, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea

(Ozdemir 2011a; Corsini et al. 2009; Razandi et al. 2015; Tahmassebpour 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 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 reliable 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 Dashtpajardi (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 the 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. In another research project, 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 an acceptable performance of the EBF model.

The aforementioned 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 the 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.

Boosted regression tree as a data mining technique was selected for this purpose as it has the capability for feature selection (Naghibi et al. 2016) as well as implementing stochastic gradient boosting to diminish variance and bias (Abeare 2009). The 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 the 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 a data-mining ensemble model in groundwater potential mapping. Thus, the aims of this study are: (1) evaluating the performance of the EBF-BRT model in groundwater potential assessment, (2) ranking the importance of groundwater-conditioning factors (GCFs) and the relationship between groundwater potential and the GCFs, and (3) providing spatial information and guidance to support decision-making processes concerning groundwater management in the Lordegan aquifer in central Iran.

Materials and methods

A spring can be defined as a feature by which groundwater flows from an aquifer to the land 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, a groundwater potential map (GPM) is proposed as a tool for providing spatial information and for determining the relationship between the spring occurrence and effective factors, here called 'conditioning factors'.

For modelling of groundwater potential, two datasets were prepared, including a springs location inventory and the GCFs. Using the mentioned datasets, the EBF model was implemented, 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 a receiver operating characteristics (ROC) plot, the efficacy of the EBF and EBF-BRT methods were validated. Figure 1 shows the methodology flowchart implemented in this research.

Study area and preparation of the conditioning factors

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

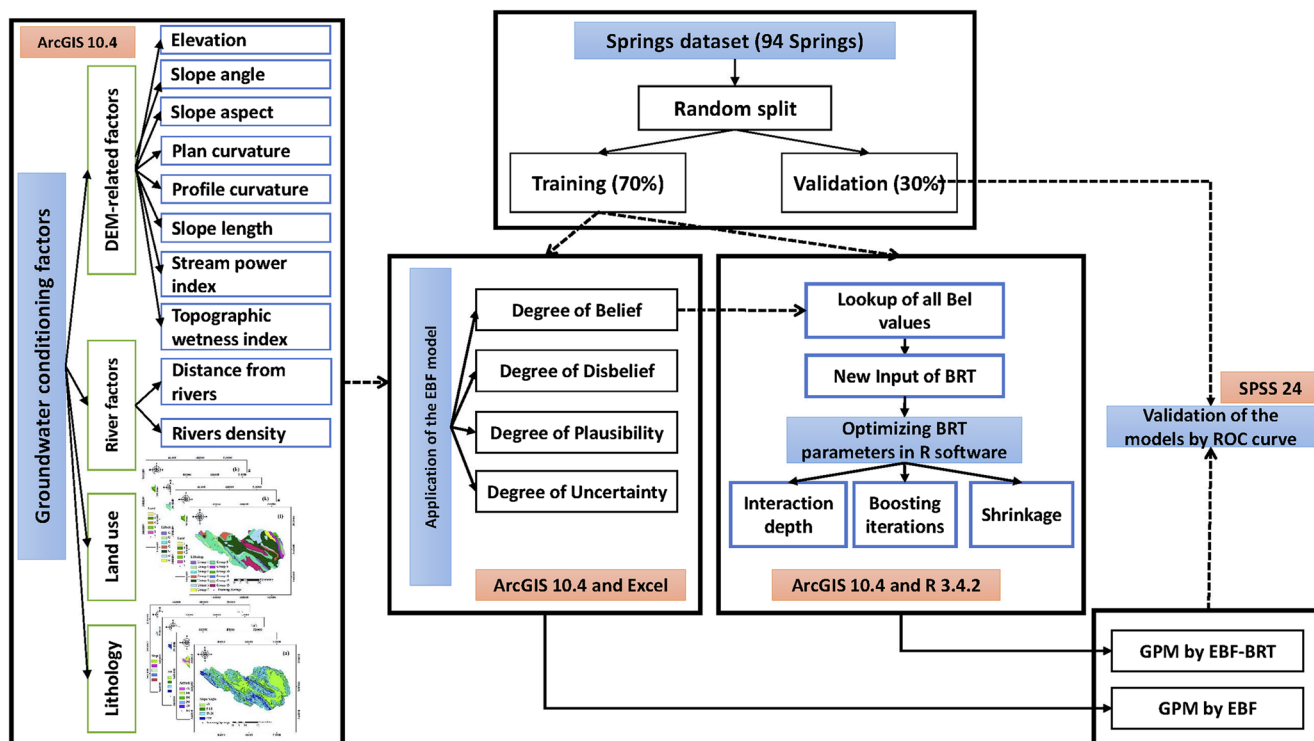


Fig. 1 Flowchart of the methodology implemented in this study

Province, Iran. Lordegan Basin covers an area of 1,486 km². The topographic elevation in Lordegan Basin ranges between 850 and 3,640 m above mean sea level (amsl) with a mean elevation 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 (GSI 1997; Table 1). The dominant 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 elevations; hence, the study was carried out at the basin scale.

Data preparation

In this study, a spring inventory dataset including 94 springs (in 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, 12 GCFs were selected for the modelling process. The 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 the EBF works with classified factors, the 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 the DEM was split into four ranges of 0–5, 5–15, 15–30, and >30° (Fig. 3a). Slope aspect was also derived from DEM data and then classified into nine classes (Fig. 3b). Elevation is another important GCF (Ozdemir 2011a, b) that was employed in this investigation (Fig. 3c). The elevation 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).

The stream power index (SPI) could be implemented to show potential flow erosion at a specific location of the basin (Moore and Burch 1986; Fig. 3g). Further, the topographic

Table 1 Lithology characteristics of Lordegan Basin, Iran

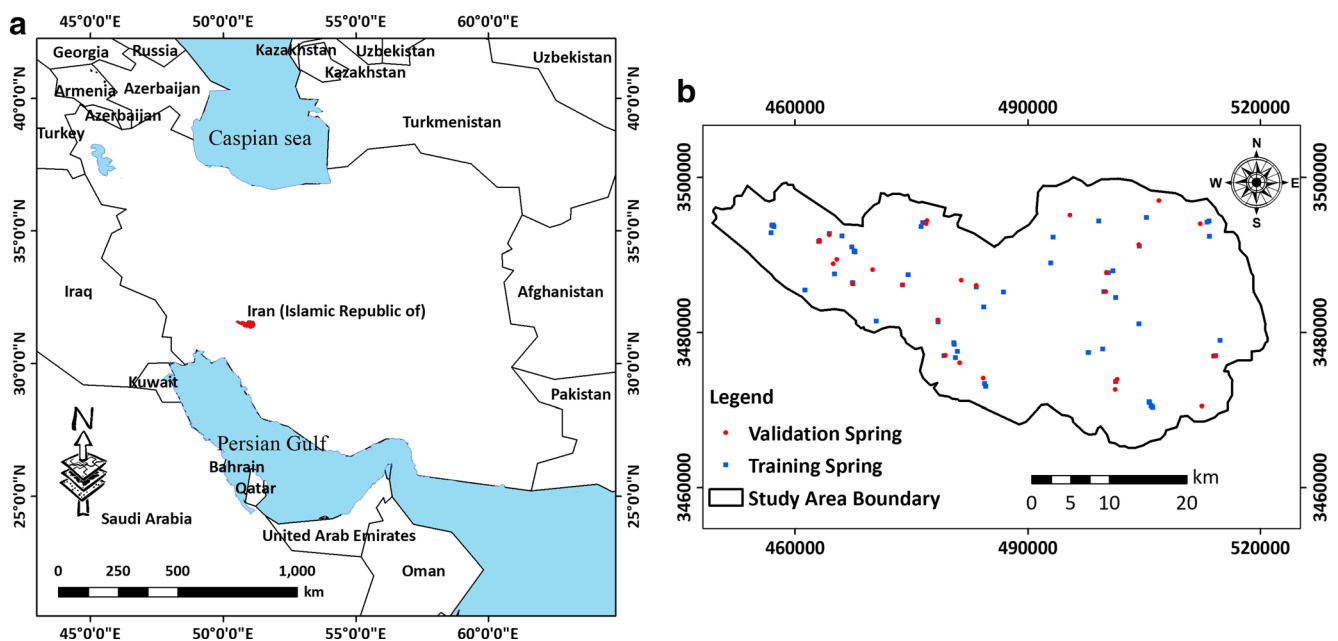
Lithology group	Lithology characteristics
1	Anhydrite, salt, grey and red marl, alternating with anhydrite, argillaceous limestone and limestone
2	Blue and purple shale and marl inter-bedded with the argillaceous limestone
3	Bluish grey marl and shale with subordinate thin-bedded argillaceous-limestone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

wetness index (TWI) was taken into account in this investigation. TWI denotes the spatial changes of soil moisture (Moore and Burch 1986; Fig. 3h).

Distance from rivers and river density are two crucial GCFs that affect the groundwater potentiality (Naghbi 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 chosen, and the distances were then classified into five groups (Fig. 3i). A rivers density map was partitioned into four categories by a natural break classification method (Fig. 3j).

A 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: 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 (Naghbi 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 faults; therefore, fault-related factors were not considered in the current research. Based on a

**Fig. 2** Locations of the study area in Iran, and the training and validation springs

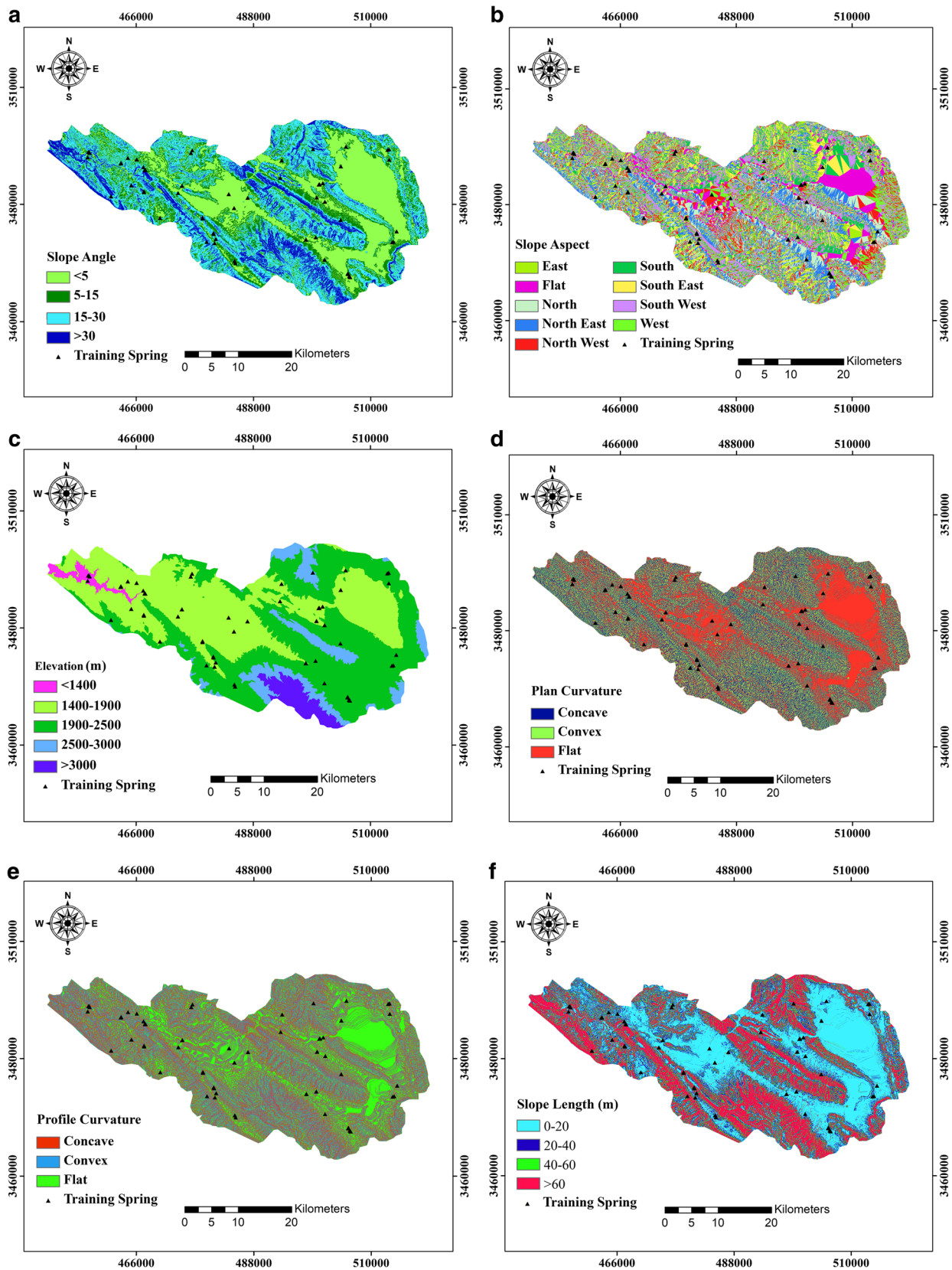


Fig. 3 The groundwater-conditioning factors (GCFs) considered in this study: **a** slope angle, **b** slope aspect, **c** elevation, **d** plan curvature, **e** profile curvature, **f** slope length, **g** stream power index (SPI), **h** topographic

wetness index (TWI), **i** distance from rivers, **j** rivers density, **k** land use, and **l** lithology

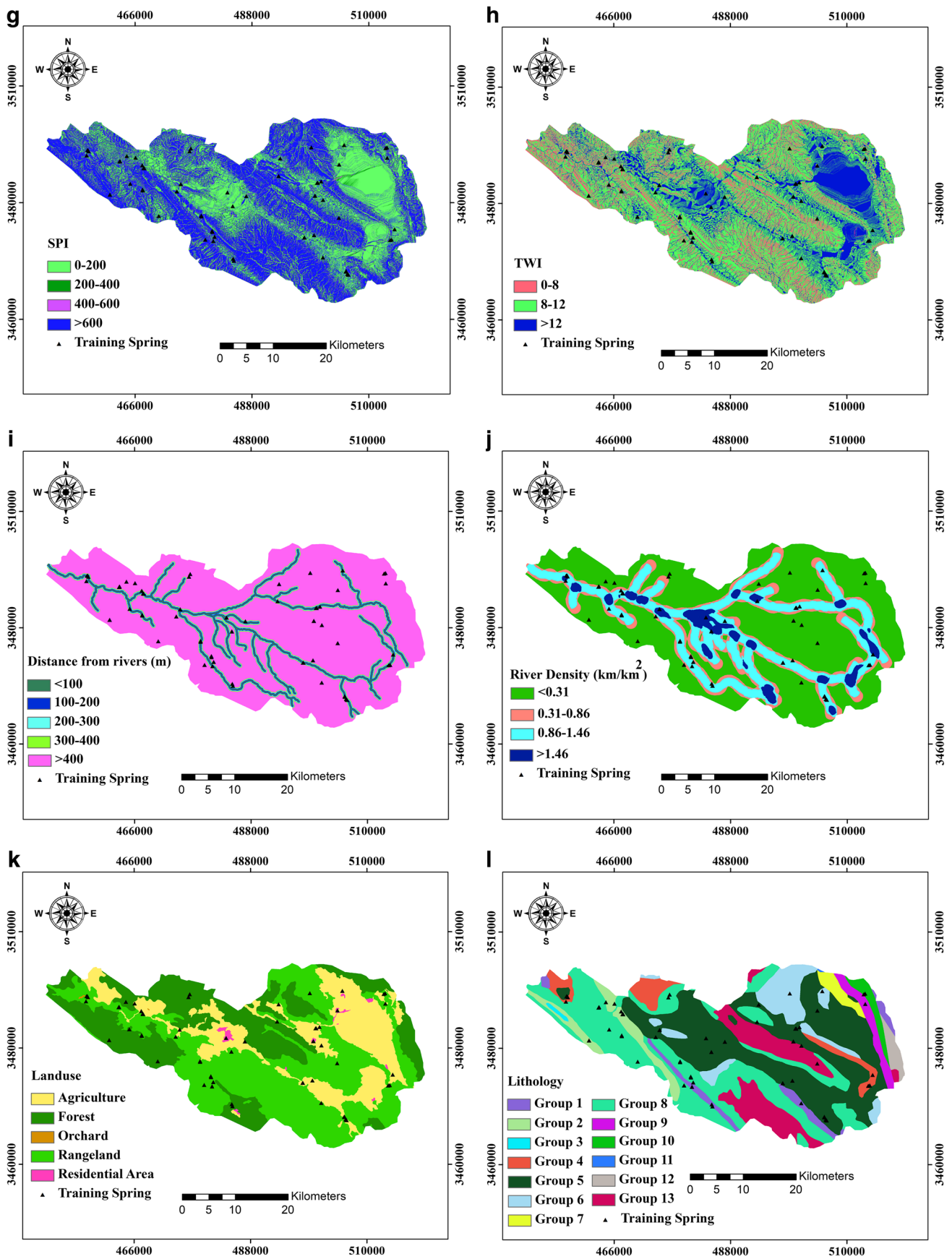


Fig. 3 (continued)

1:100,000-scale geological map, the geological units were partitioned into thirteen units including groups 1–13 (Table 1; Fig. 31).

Modelling process

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

Evidential belief function (EBF) model

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 the generalized Bayesian theorem, 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):

$$m : 2^\Theta = \{ \phi, T_P, \overline{T_P}, \Theta \} \quad \text{with } \Theta = \{ T_P, \overline{T_P} \} \quad (2)$$

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

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

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

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

where $N(S \cap A_{ij})$ denotes the 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).

The novel 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 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. Recalculate weights $w_i = w_i \exp[\alpha_m I(y_i \neq C_m)]$

Finally, the majority vote can be obtained by:

$$\text{sign} = \left[\sum_{m=1}^M \alpha_m C_m(X) \right]$$

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

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

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

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

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

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

To apply a novel data-mining ensemble model, first, the EBF model was applied and belief values were assigned to different classes of the GCFs. Then, new maps of each factor were produced by the 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 the spring and 0 was assigned to nonspring locations. It is noted that the

nonspring locations were randomly defined using ArcGIS 10.4. Using the new training dataset and new GCFs layers with Bel values, the BRT model was conducted using R open-source software via the gbm package (Ridgeway 2006). The BRT model was run using a 10-fold cross-validation, deemed to be a sufficient number of runs for optimization of the assigned parameters. It needs to be clarified that the GPMs produced by the EBF and BFB-BRT methods are classified into four classes—low, moderate, high, and very high—by the natural break classification method (Naghibi et al. 2018).

Results and discussion

GPM production by 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 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 a 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. In contrast, 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 elevation, the results indicated that an inverse relationship exists between GCF and spring incidence. At lower elevations, water has concentrated near the rivers and, therefore, the wetness index is higher in these areas which 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 Bel value was observed in the (−0.001)–(0.001) category of the 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, the 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. In contrast, an inverse relationship was observed between the groundwater potentiality and three GCFs including elevation, slope length, and distance from rivers. Naghibi and Pourghasemi (2015) obtained the same relationship between elevation, TWI, and distance from rivers and spring occurrence. However, in some other factors such as LS, the findings of this study differ from the findings of Naghibi and Pourghasemi (2015). 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.

The GPM produced by the EBF model in the current study is presented in Fig. 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.

GPM production by the novel data-mining ensemble model

The findings of the application of BRT algorithm are presented in Fig. 5. The final BRT model was applied with the 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 Fig. 6. The results indicated that the distance from rivers, lithology, river density, and plan curvature have the highest contribution to groundwater potential estimated by the 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 the EBF-BRT method is presented in Fig. 4b and Table 3. The GPM produced by the 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.

Validation and verification of the GPMs

This section includes two steps: (1) validation of the maps using the validation dataset and ROC curve and (2) 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 the EBF and EBF-BRT

Table 2 Spatial relationship between GCFs and springs using the 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
	Elevation (m)	<1400	1.63	4	0.61	0.24
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

Table 2 (continued)

Factor	Class	% of pixels in domain	No. of springs	Bel	Dis	Unc
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

Bel belief, *Dis* disbelief, *Unc* uncertainty

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 the ROC curve. Accordingly, the area under the curve of ROC for the validation dataset was defined as 75.5 and 82.1% for EBF and EBF-BRT models, respectively. Further, the area under the ROC curve for the 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 (Naghbi et al. 2016).

To verify the resulting 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 the 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 that, 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

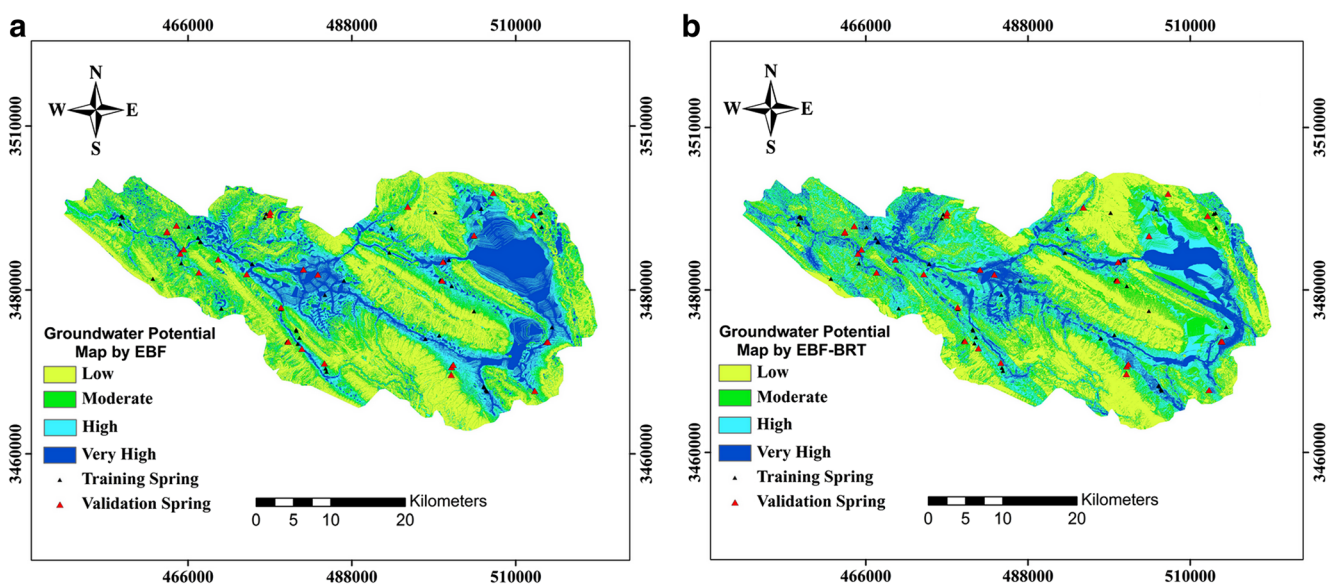


Fig. 4 Groundwater potential map produced by the **a** EBF and **b** EBF-BRT models

Table 3 Range and area of different classes of the groundwater potential map (GPM) produced by the EBF and EBF-BRT models

Class	EBF		EBF-BRT	
	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

potential zones, as well as the computed area under the ROC curve, confirm the satisfying performance of the models in this study.

Performance comparison

The findings of this study indicated superior performance of the EBF-BRT to EBF model 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 the 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 landslides. Their

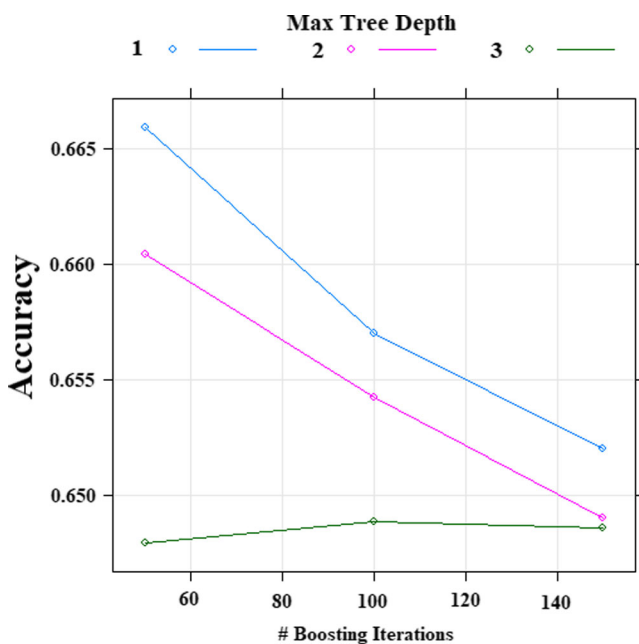


Fig. 5 Results of the EBF-BRT application

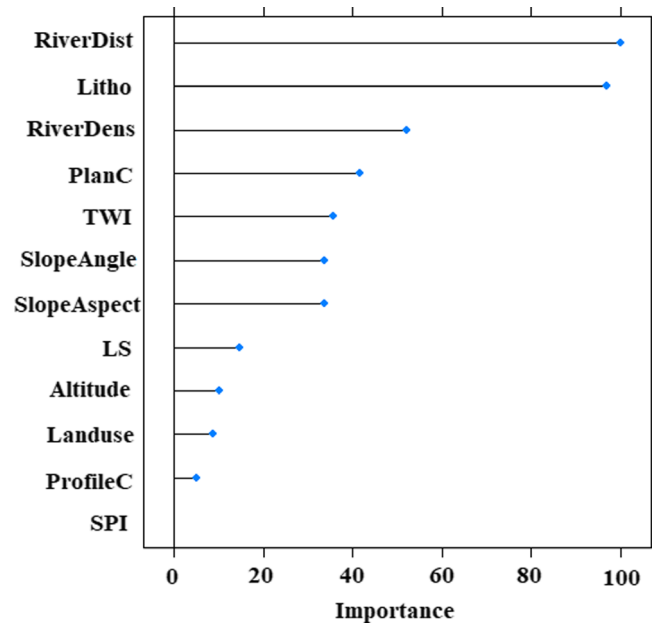


Fig. 6 Importance of the grGCFs in the BRT model (RiverDist distance from rivers; Litho lithology; RiverDens rivers density; PlanC plan curvature; TWI topographic wetness index; SlopeAngle slope angle; SlopeAspect slope aspect; LS slope length; Altitude elevation; Landuse land use; ProfileC profile curvature; SPI stream power index)

findings showed the higher efficacy of the hybrid method relative to the EBF model. In another research project, 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 the EBF model. In a related work, Naghibi et al. (2017a) used an 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 non-linear relationships (Naghibi et al. 2016). Boosted regression tree 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.

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

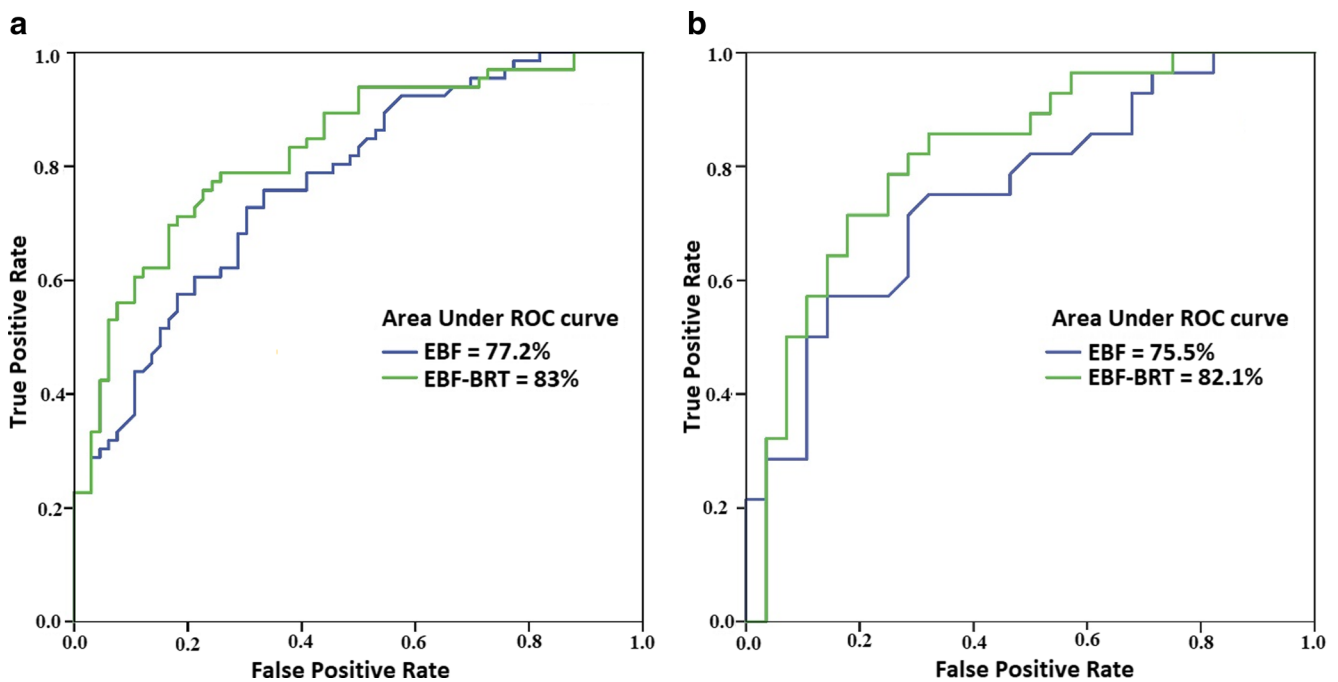


Fig. 7 Receiver operating characteristics (ROC) curve calculated for the EBF and EBF-BRT models for **a** training and **b** validation datasets

ensemble EBF-BRT model was introduced, and its performance was assessed in groundwater potential mapping. The 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 the EBF-BRT model yielded better performance than the 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 the 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 non-linear 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 the 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 groundwater extraction can be more productive. Better land use planning could lead to lower pressure on aquifers. However, it is the first step and there need to be more remediation steps, such as artificial recharge through water harvesting, and flood spreading.

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

Potential zones	EBF		BRT	
	No. of springs	Springs (%)	No. of springs	Springs (%)
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

Acknowledgements The authors would like to appreciate the editor Dr. Martin Appold for handling the paper and two anonymous reviewers for their constructive comments on the previous version of the paper.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

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 online fishery. Msc Thesis, LSU, Baton Rouge, LA, USA
- Carranza JEM, Hale M (2003) Evidential belief functions for data-driven geologically constrained mapping of gold potential, Baguio district, Philippines. *Ore Geol Rev* 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. <https://doi.org/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
- Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *J Anim Ecol* 77:802–813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>
- Freeze RA, Cherry JA (1979) *Groundwater*, vol XVI. Prentice-Hall, Englewood Cliffs, NJ, 604 pp
- Geology Survey of Iran (GSI) (1997) *Geological Survey and Mineral Exploration of Iran*. http://www.gsiir/Main/Lang_en/indexhtml. Accessed July 20, 2018
- 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. *Environ Monit Assess* 190(3):149
- Hong H, Naghibi SA, Dashtpajardi MM, Pourghasemi HR, 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. *Arab J Geosci* 10(7):167
- Jaafari A, Gholami DM (2017) Wildfire hazard mapping using an ensemble method of frequency ratio with Shannon's entropy. *Iran J Forest Poplar Res* 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 Nat Hazards Risk* 9(1):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. <https://doi.org/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 Process* 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. <https://doi.org/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 Geosci* 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. <https://doi.org/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 Inform* 8:1–16. <https://doi.org/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. <https://doi.org/10.1007/s10661-015-5049-6>
- Naghibi SA, Moradi Dashtpajardi 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. <https://doi.org/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 SA, Pourghasemi HR, Abbaspour K (2018) A comparison between ten advanced and soft computing models for groundwater qanat potential assessment in Iran using R and GIS. *Theor Appl Climatol* 131(3–4):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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/10.1016/j.jhydrol.2011.05.015>
- Pham BT, Jaafari A, Prakash I, Bui DT (2018) A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling. *Bull Eng Geol Environ*. <https://doi.org/10.1007/s10064-018-1281-y>
- 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. <https://doi.org/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. <https://doi.org/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. *Hydrogeol J* 22(3):643–662. <https://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. *Sci Total Environ* 568(15):1110–1123. <https://doi.org/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. <https://doi.org/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 Inform* 8:867–883. <https://doi.org/10.1007/s12145-015-0220-8>
- Ridgeway G (2006) *gbm: generalized boosted regression models*. R package version 1(3), 55 pp
- Sangchini EK, Emami SN, Tahmasebipour N, Pourghasemi HR, Naghibi SA, Arami SA, 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. *Arab J Geosci* 9(3):201
- Shafer G (1976) *A mathematical theory of evidence*. Princeton Univ Press, Princeton, 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. <https://doi.org/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. <https://doi.org/10.1016/j.jhydrol.2014.03.008>
- Tahmasebipour 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. <https://doi.org/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. <https://doi.org/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 and logistic regression models in West Sumatera Province, Indonesia. *Catena* 118:124–135. <https://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*. <https://doi.org/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. <https://doi.org/10.1007/s12665-016-5424-9>