

Spatial mapping of artesian zone at Iraqi southern desert using a GIS-based random forest machine learning model

Alaa M. Al-Abadi¹  · Shamsuddin Shahid²

Received: 3 June 2016 / Accepted: 4 June 2016 / Published online: 14 June 2016
© Springer International Publishing Switzerland 2016

Abstract Random forest (RF) machine learning technique and geographical information system (GIS) have been applied to delineate groundwater flowing well zones in the southern desert of Iraq. A spatial database consists of target variable, i.e., geographic locations of 93 flowing wells and predictor variables, i.e., the factors that control groundwater occurrence was prepared for this purpose. Eleven predictor variables were selected based on data availability, literature review, and field conditions which include elevation, slope, profile curvature, aspect, topographic wetness index, stream power index, distance to Abu Jir fault, distance to Euphrates River, major aquifer group, total hydraulic head, and well depth. The RF model in R package along with ArcGIS 10.2 was used to generate groundwater flowing well potential index for the study area. The obtained potential indices were classified using natural break classification scheme into five categories namely, very low, low, moderate, high, and very high. The results revealed that high or very high groundwater flowing well potential zones occupy 15 %, moderate potential zone covers 6 %, and low or very low potential zones cover 79 % of the southern desert of Iraq. The groundwater flowing well zone map was validated using relative operating characteristic (ROC) curve. The areas under the ROC curve for success and prediction rates were 0.98 and 0.97,

respectively, indicating excellent capability of RF model to delineate groundwater potential. It is expected that the method development in this study can be used for rapid but efficient evaluation of groundwater flowing well potential from limited amount of data.

Keywords Random forest · Groundwater · Southern desert of Iraq · ROC · GIS

Introduction

Iraq has abundant surface water resources compared to other countries in the Arabian Peninsula. However, mismanagement of this precise resource and interventions in the upstream of Tigris and Euphrates rivers as well as their tributaries by riparian countries bordering Iraq have made surface water gradually a scarce resource in Iraq (Al-Ansari 2013). It is anticipated that groundwater will play an important role in the area to supplement water supply to growing population and agricultural activities in near future. The southern desert of Iraq contains huge amount of groundwater resources suitable for agricultural and industrial uses and even for drinking after appropriate treatment. In the east and north parts of the desert, a set of springs and flowing wells exists which extends parallel to the Euphrates River. The spatial demarcation of this hydrogeological system can facilitate groundwater resources management and development efficiently, and thus agricultural development in the west of the Euphrates River.

Delineation of groundwater potential zones is an important prerequisite for implementation of successful groundwater development, protection, and management program (Ozdemir 2011a). Two main approaches are usually used for demarcation of groundwater potential zones namely, data-

✉ Alaa M. Al-Abadi
alaaatiaa@gmail.com
Shamsuddin Shahid
sshahid_ait@yahoo.com

¹ Department of Geology, College of Science, University of Basra, Basra, Iraq

² Faculty of Civil Engineering, Universiti Teknologi Malaysia, Johor 81310, Malaysia

driven and knowledge driven (Bonham-Carter 1994). The data-driven approaches use known locations of well/spring as dependent variable and groundwater occurrence controlling factors as independent variables. The advantages of these methods are that they need little data, and they are less affected by potential bias in human input (McKay and Harris 2015). Examples of these techniques are frequency ratio (Ozdemir 2011a; Oh et al. 2011; Manap et al. 2011; Moghaddam et al. 2013; Pourtaghi and Pourghasemi 2014; Naghibi et al. 2014; Elmahdy and Mohamed 2014; Al-Abadi 2015b), artificial neural networks (Corsini et al. 2009; Lee et al. 2012), weights of evidence (Corsini et al. 2009; Ozdemir 2011b; Lee et al. 2012; Pourtaghi and Pourghasemi 2014; Al-Abadi 2015a), maximum of entropy (Rahmati et al. 2016), evidential belief functions (Nampak et al. 2014; Mogaji et al. 2014; Pourghasemi and Beheshtirad 2015), logistic regression (Ozdemir 2011a; Pourtaghi and Pourghasemi 2014), and Shannon's entropy (Naghibi et al. 2014; Al-Abadi 2015b). In recent years, machine learning techniques such as, boosted regression tress, classification and regression tress (CART), decision tress, and random forest (Lee and Lee 2015; Naghibi et al. 2016; Rahmati et al. 2016) have also been used of spatial zoning of groundwater potential. On the other hand, knowledge driven approaches do not need geographic information of wells or springs, rather they rely on the data to determine the weights or importance of independent groundwater occurrence controlling factors. Examples of these approaches are index overly (Jha et al. 2010; Machiwal et al. 2010; Manap et al. 2011; Abdalla 2012; Pandey et al. 2013; Al-Abadi and Al-Shamma'a 2014), fuzzy logic (Shahid et al. 2002), and analytical hierarchical process (Adiat et al. 2012; Rahmati et al. 2014).

The random forests (RF) algorithm is a machine learning technique which has been applied recently as a data-driven predictive model for groundwater potential mapping (Naghibi et al. 2016; Rahmati et al. 2016). Machine learning is defined as a field of computer science that gives computers the ability to learn without being explicitly programmed. The RF is one of the most powerful, fully automated machine learning techniques (Fernandez-Delgado et al. 2014). It can handle data from various measurement scales without any statistical assumptions (Rahmati et al. 2016). At the same time, RF is computationally inexpensive than other machine learning algorithms like, neural networks or support vector machines (Rodriguez-Galiano and Chica-Olmo 2012). Another advantage of the RF is that it allows assessment of the importance of input variables in prediction (Rahmati et al. 2016). Although RF has been applied recently in many earth science disciplines including groundwater potential mapping, its capability to delineate groundwater flowing well zones is still not well explored.

The major objective of this study is to delineate the artesian zone in the southern desert of Iraq using RF and GIS. The methodology proposed in the study can be used for quick but efficient mapping of groundwater resources with limited amount of data and human interference. It is also expected that the groundwater artesian zone map development in the present study will allow efficient planning, management and development of groundwater resources in the study area.

The study area

The southern desert locates in the southern part of Iraq, south of Euphrates River, encompass an area of about 78,390 km² (Fig. 1). Four administrative governorates share the area of the southern desert namely, Najif, Samawa, Nasiriya, and Basra. The most parts of the southern desert are unpopulated due to extreme climate and severe water scarcity. Few urban centers are located along the Euphrates River, while small towns such as Al-Shbicha, Al-Salman, and Al-Busaiya are sparsely distributed through the area. The topography of the study area is mostly flat with a rising elevation from the northeast to the southwest (Al-Jiburi and Al-Basrawi 2008). The elevation ranges between 1 and 494 m with an average of 229 m (Fig. 2). The climate in major portion of the southern desert is arid with an annual average rainfall typically in range of 100–150 mm. However, the rainfall varies extremely from year to year. Most of the rainfall occurs between October and May with little or no rainfall in other months (Parsons 1955). Despite small amount of rainfall, different types of vegetation grow in the southern desert as rainfall mostly occurs in winter when evaporation is less. The climatic data at nine meteorological stations within and around the study area for the time period 1980–2015 show that the annual average of rainfall, temperature, evaporation, relative humidity, and wind speed in the area are 108 mm, 30.4°C, 352 mm, 46 %, and 2.9 m/s, respectively. The geological formations exposed in the study area, from oldest to youngest are the Aidah, Rus and Dammam with their various limestone members, and the Zahra, Euphrates, Fatha, Dibdibba and recent alluvium (Parsons 1955). In general, all the formations contain groundwater in varying amounts and quality. Since most of the springs and flowing wells occur with the Euphrates, Dammam, and Quaternary rocks, a brief description of these rocks is given here. The Euphrates Formation is the most widely spread formation of the Early Miocene sequence in Iraq with thickens up to 160 m. It comprises 8 m thick shalley chalky and well bedded recrystallized limestones with texture ranging from oolitic to chalky, which locally contain corals and shell coquinas (Jassim and

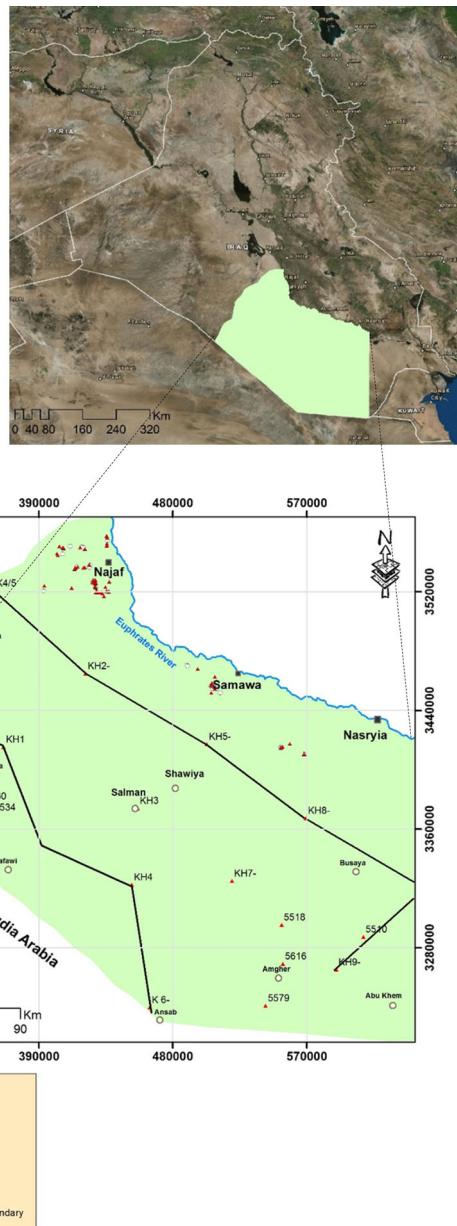


Fig. 1 Location map of the study area

Goff 2006). The Dammam Formation mainly consists of chalky limestone, dolomite, marl and Shales. The Quaternary formation in the eastern part of the study area consists of sand and alluvium deposits of recent and Pleistocene ages. The geological map of the study area is given in Fig. 3.

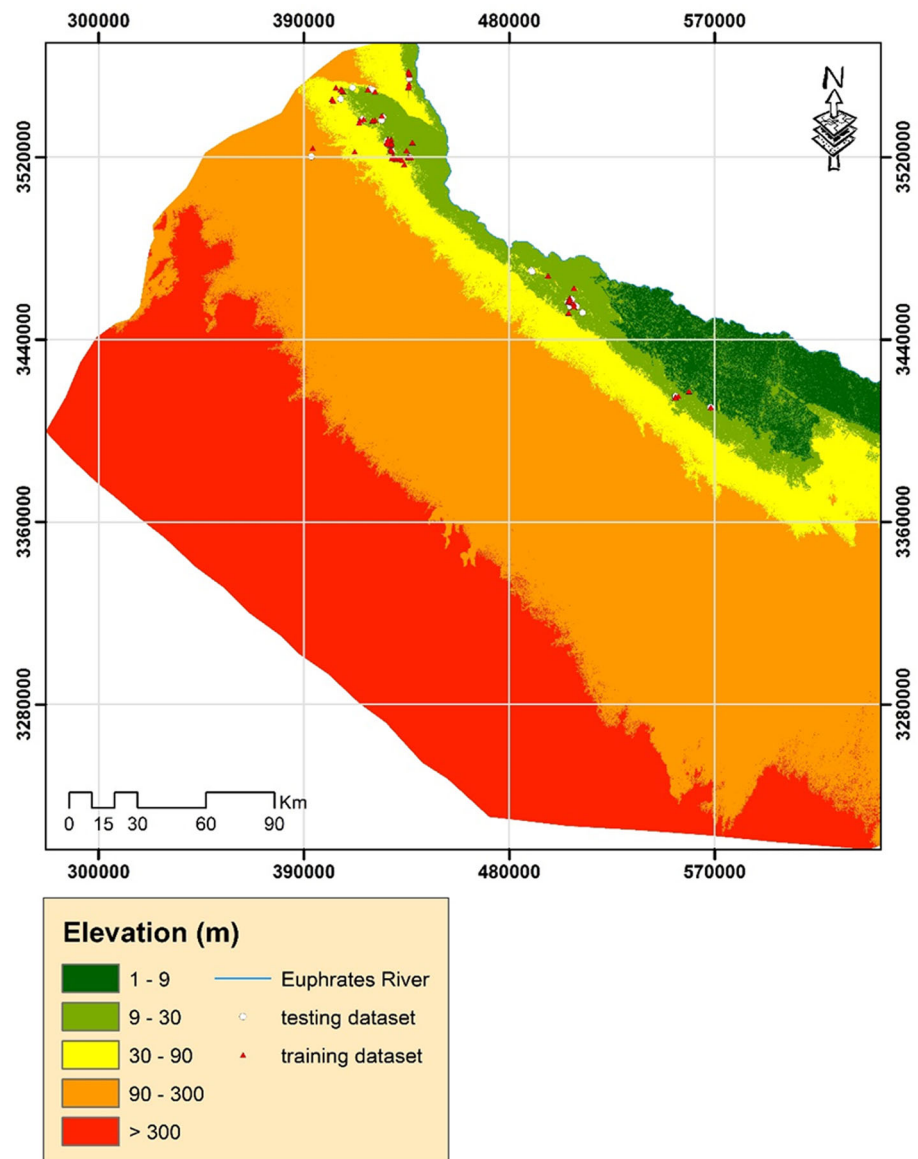
From the tectonic point of view, the southern desert is a part of the Arabian Platform, which is characterized by the presence of block tectonics and the absence of tectonic folds (Buday and Jassim 1987). The main structural features include several northeast-southwest transversal faults. The eastern boundary of the southern desert is sharply marked by the southern extension of Hit-Abu Jir fault

system. This fault system represents structural and geological boundary zone that separates the desert from the Mesopotamia Zone.

There are five major aquifer groups in the study area (Jassim and Goff 2006), which are termed as 4, 5, 6, 7, and 10 aquifer groups (Fig. 4). The aquifer group 4 represents limestone of the Palaeogene–Neogene Euphrates Formation, Kirkuk Group and Ghar formations of the Western and Southern desert of the Rutba Subzone and the Salman Zone. The aquifer group 5 represents the karstified and fractured limestones of the Palaeogene Um-Er Radhuma, and Jill and Dammam formations of the southern desert of the Salman zone. The aquifer group 6 depicts the sandstone and conglomerate of the Miocene to Pleistocene Ghar and Zahra formations and the Nukhaib Graben fill of the western and southern deserts of the Rutba subzone and the Salman zone (Jassim and Goff 2006). The aquifer group 7 represents the sandstone of the Mio-Pliocene Dibdibba Formation of the southern desert in the Salman Zone, while the aquifer group 10 represents the sands of the Quaternary Mesopotamian flood plain of central and the south of Iraq in the Mesopotamian zone. A detailed description of these aquifer units with relevant information can be found in Jassim and Goff (2006). Groundwater in the study area moves from the west and the southwest (recharge areas) to the east and the northeast (discharge areas) (Fig. 5). Groundwater level varies from ten meters in recharge areas to near surface or artesian in discharge areas (Al-Jiburi and Al-Basrawi 2008). Groundwater quality in the study area can be classified into three types namely, mixed saline with either sulfates of chlorides dominant, carbonate water and high nitrate waters (Parsons 1955).

The main aquifers underlying the study area from oldest to youngest are: Hartha, Tayart, Umm Er Radhuma, Dammam, Ghar-Euphrates, Dibdibba, and Quaternary (Fig. 5). A hydraulic connection is possible between these water bearing layers as a result of the piezometric changes throughout these aquifers. Since Dammam aquifer has high hydraulic pressure, most of the wells drilled in this aquifer is flowing artesian well. A brief description of these aquifers is given here. Dammam formation comprises of limestone, dolomite, limestone and dolomite, with marl and evaporates. Dammam aquifer is characterized by high permeability due to presence of cavities, karstified features, fractures, fissures, and joints. The Dammam aquifer is considered as the main regional groundwater aquifer in the southern desert due to its wide extension and huge amount of stored groundwater (GEOSURV 1983). The hydrological characteristic of the Dammam aquifer is presented in Table 1. The transmissivity of the aquifer ranges from 3.1 to 4752 m²/day, and thus it regards as extremely heterogeneous. The hydraulic conductivity ranges from 0.1 to 100 m/day, while static water level ranges from 0 to 170 m

Fig. 2 Elevation (m) map of the study area



below the earth surface. The total dissolved solid is in the range of 350 to 8530 mg/l. The dominant water type is sulphatic, in addition to chloride and biocarbonatic water types. The source of sulphate is attributed to the presence of evaporates within the rocks or gypsiferous soil (Al-Jiburi and Al-Basrawi 2000).

Random forest algorithm

The RF is an ensemble of learning techniques that generate many classification tresses which are aggregated to compute a classification or regression (Breiman 1984; 2001). Ensemble learning (EL) is a method that generates many classifier and aggregate their results. EL can be classified

into two well-know methodologies: boosting and bagging. In boosting, successive trees give extra weight to point incorrect prediction by earlier predictors, and finally a weighted vote is taken for prediction (Liaw and Wiener 2002). In bagging, successive tress is independently constructed using a bootstrap sample of the dataset, and do not based on generated earlier trees. The prediction is taken as a simple majority vote. RF belongs to the family of ensemble methods appeared in machine learning at the end of 1990s (Dietterich 2000). The principle of RF is to combine many binary decision tresses built using several bootstrap samples coming from the learning sample L and choosing randomly a subset of explanatory variable X at each node (Genuer et al. 2008). Roughly, two-third of the learning samples (also called bag samples) is used for

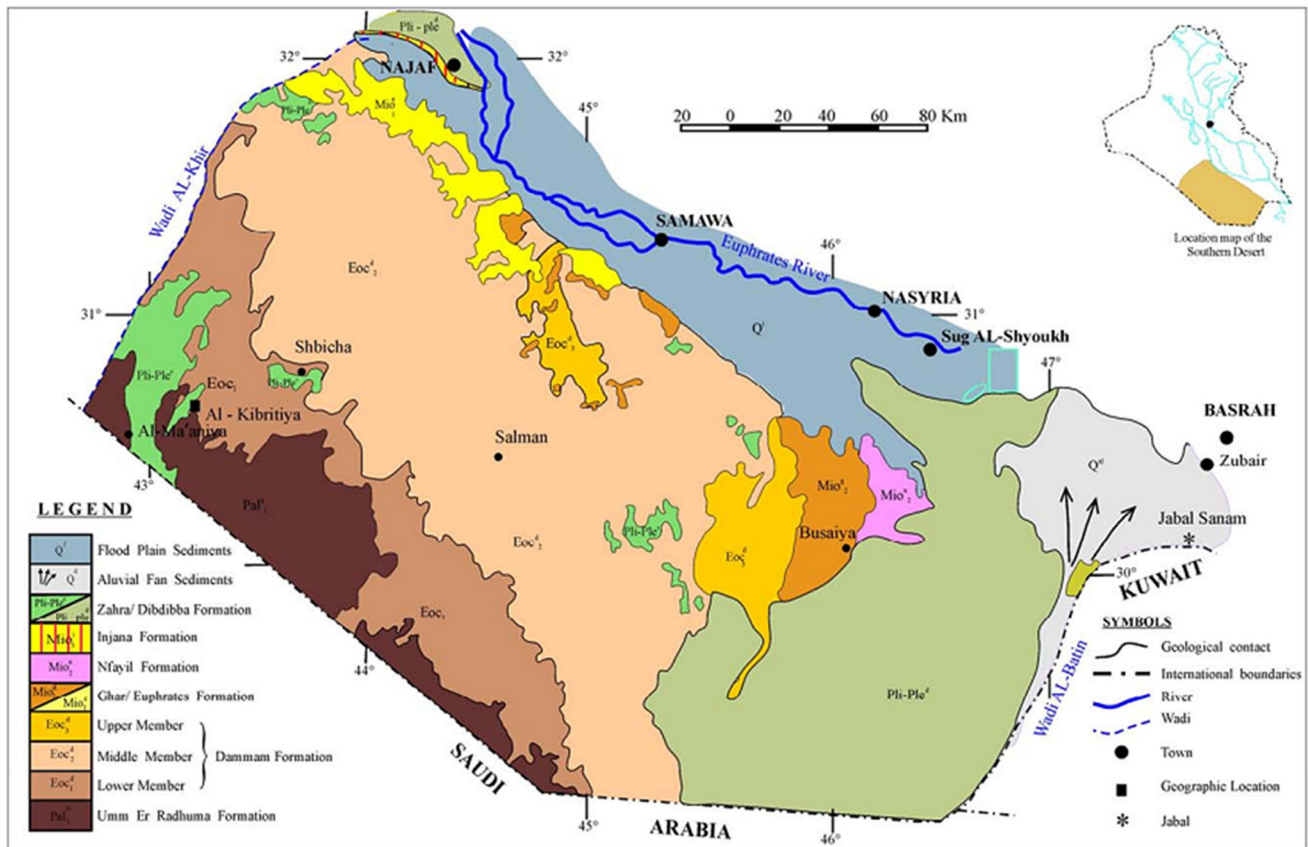


Fig. 3 Geological map of the study area (after Sissakian 2000)

prediction, while the remaining one-third [also called out-of-bag (OOB)] is used for validation. RF algorithm splits the target variable (parent node), into binary pieces, where the child nodes are ‘purer’ than the parent node (Carranza and Laborte 2015a). Each node is split using the best among a subset of predictors randomly chosen at that node. The process of data splitting in each internal node is iterated until a pre-specified stop requirement is reached (Carranza and Laborte 2015a). After that, a simple regression model is attached for every child node (leaf). The final output of RF is the majority of output from all decision trees. The RF algorithm for growing a random forest of k classification trees is as follows (Peters et al. 2007; Hastie et al. 2009):

- i. for $i = 1$ to k do:
 1. Draw a bootstrap sample (subset X_i) from the original dataset X ;
 2. Use X_i to grow an unpruned classification tree to the maximum depth with the following modification compared with standard classification tree building:
 - (a) randomly select m variables from p variables;
 - (b) choose the best split among these variables; and
 - (c) split the node into two daughter nodes.

- ii. Predict new data according to the majority vote of the ensemble of k trees.

An unbiased estimate of the error rate is obtained during the construction of a RF based on the training data as:

- i. At each bootstrap iteration, predict the data in OOB that are not used in the construction of the i th tree.
- ii. At the end of the run, aggregate the OOB predictions, calculate the error rate, and call it the OOB estimate of error rate.

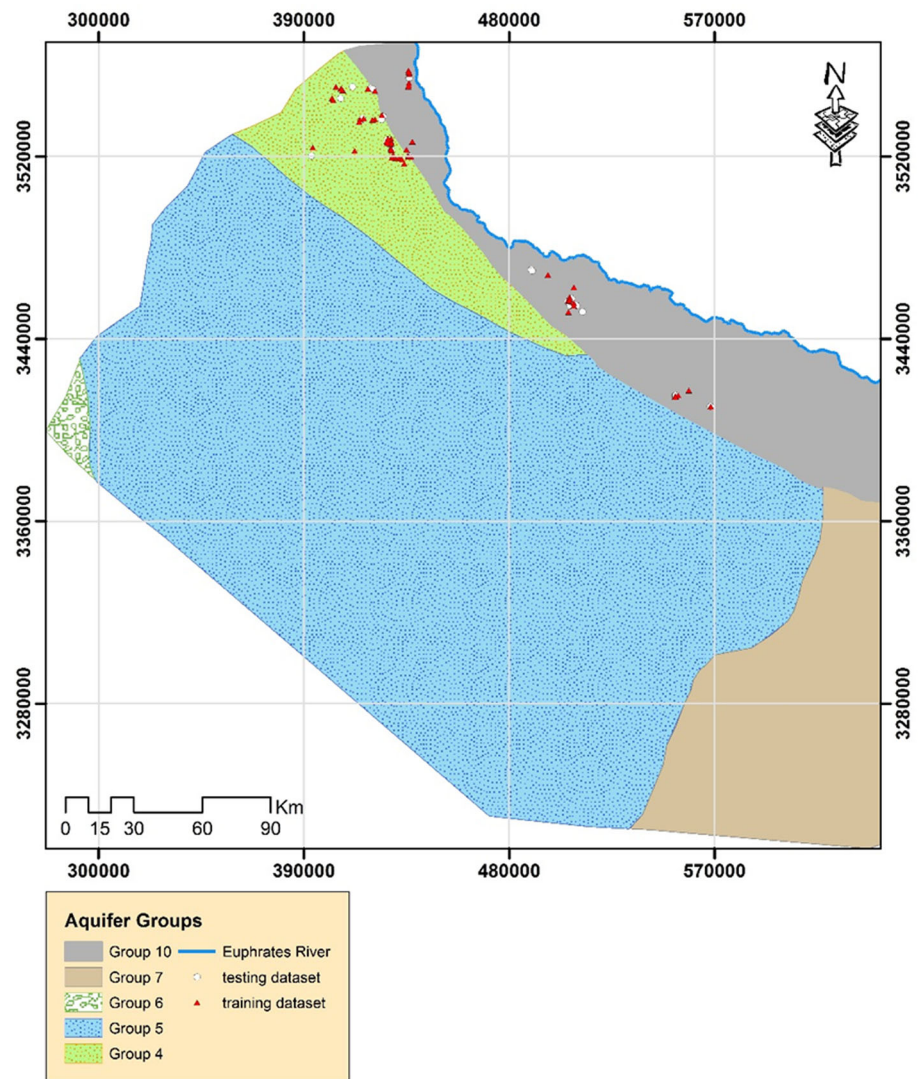
A more detailed description of the RF algorithm can be found in Breiman (2001) and Liaw and Wiener (2002).

Materials and methods

Data used

The data used in the creation of spatial zones of groundwater flowing wells involve target variable (geographic location of the flowing well) and various predictor variables represented by thematic layers of groundwater affecting occurrence factors. The flowing well inventory

Fig. 4 Spatial distribution of aquifer groups in the southern desert of Iraq [after Jassim and Goff (2006)]



was developed by a team of General Commission of Groundwater, Ministry of Water Resource of Iraq through extensive field survey in the year 2013. In total, 93 perennial flowing wells were identified. This dataset is randomly partitioned using random algorithm in MINITAB 16 software into two sets: training and testing. Of the 93 flowing well locations, 65 wells (70 %) were used as training dataset, and the remaining 28 wells (30 %) were used as validation dataset.

Eleven predictors were used in this study, which were selected based on literature review, expert opinion, data availability, and field conditions. These variables are elevation, slope, profile curvature, aspect, topographic wetness index (TWI), stream power index (SPI), distance to Abu Jir fault, distance to Euphrates River, major aquifer group, total hydraulic head, and well depth. The thematic maps of all predictors were prepared as raster layer with 30×30 m resolution in ArcGIS 10.2 software.

The total number of pixels for each thematic raster layer was 144,007,200 (12,204 columns and 11,800 rows). Topographic variables namely, elevation, slope, profile curvature, aspect, TWI, and SPI were prepared using Advanced Spaceborne Thermal Emission and Reflection Radiometer-Global Digital Elevation Model (ASTER-GDEM) data with a spatial resolution of 30 m, download from United State of Geological Survey (USGS) website (earthexplorer.usgs.gov). The importance of these variables in delineating groundwater potential zones are extensively described in literatures (Ozdemir 2011a, b; Oh et al. 2011; Pourtaghi and Pourghasemi 2014; Naghibi et al. 2014). The surface elevation map was directly derived from DEM and classified into five categories (McDonald et al. 1990): plains (0–9 m), rises (9–30 m), low hills (30–90), hills (90–300 m), and mountains (>300 m) (Fig. 2). Slope (%) map was also derived from filled DEM and classified into five classes (de Winnaar

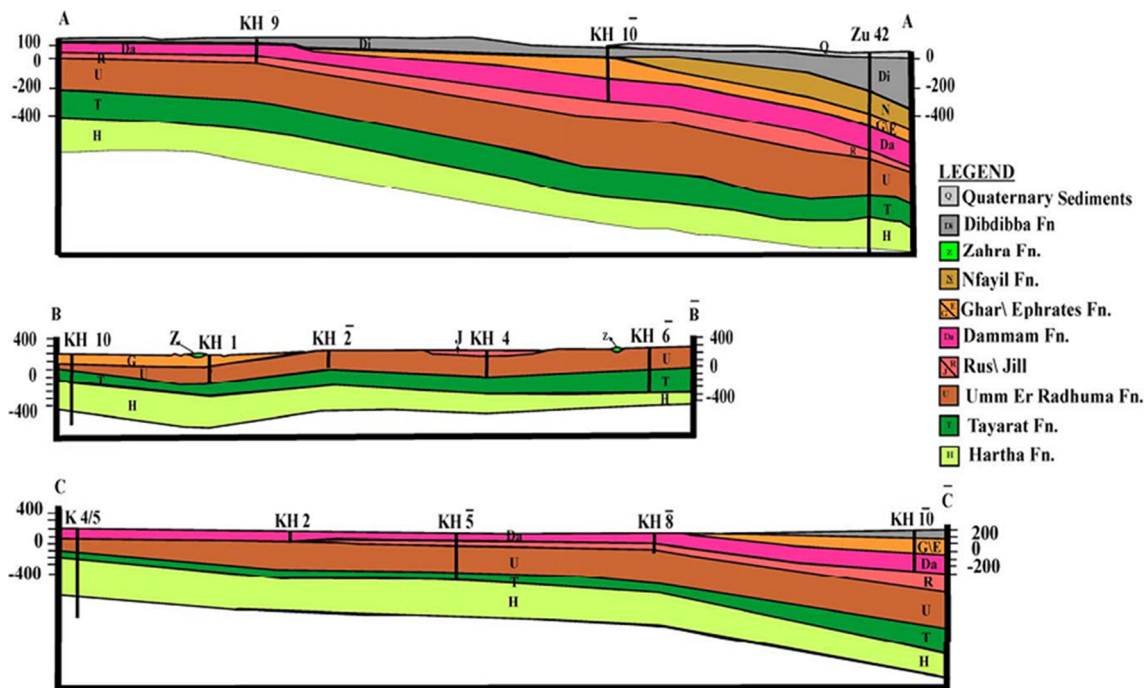


Fig. 5 Geological cross sections in the Iraqi Southern Desert [after Araim (1984)]

Table 1 Hydrogeological data of Dammam aquifer in southern desert of Iraq [after Al-Jiburi and Al-Basrawi (2008)]

Well No.	Well depth (m)	Static water level (m)	Flow rate (m ³ /day)	K (m/day)	T (m ² /day)	TDS (mg/l)	Water type
5534	22	6.0	118	20.4	309	3481	Sulphatic
5560	30	7.5	691	30.9	879	7536	Sulphatic
5510	170	136	44	2.0	5	6478	Chloridic
5518	106	57	588	17.9	878	3159	Sulphatic
5579	200	88.2	396	2.4	227	2382	Sulphatic
5616	120	57	660	7.2	446	3686	Sulphatic
K 4/10	940	104.3	1037	100	4750	3612	Sulphatic
K 4/5							
KH 1	400	60.5	584	–	7	2504	Choridic
KH 2 ⁻	140	60	86	–	23	4210	Sulphatic
KH 3	420	6.5	2877	–	1636	3120	Choridic
KH 4	419	42.7	501	–	1549	1080	Sulphatic
KH 5 ⁻	605	69	2160	–	v. high	4648	Choridic
KH 7 ⁻	585	52.6	1469.3	–	625	4228	Choridic
KH 8 ⁻	300	75.5	164	–	3	6510	Sulphatic
KH 9 ⁻	195	113.5	1123	–	171	6120	Sulphatic

et al. 2007): flat (<2 %), undulating (2–8 %), rolling (8–15 %), hilly (15–30 %), and mountainous (>30 %) (Fig. 6a). Profile curvature was classified into three classes namely, concave (<0), flat (0), and convex (>0) (Fig. 6b). Aspect was classified into ten classes (Fig. 6c): Flat (–1), North (0–22.5), Northeast (22.5–67.5), East

(67.5–112.5), Southeast (112.5–157.5), South (157.5–202.5), Southwest (202.5–247.5), West (247.5–292.5), Northwest (292.5–337.5), and North (337.5–360.0). The secondary topographic predictor variables namely, TWI and SPI were derived using following equations (Moore et al. 1991),

Fig. 6 **a** Map showing the distribution of slope (%) in the study area. **b** Map showing spatial distribution of total curvature in the study area. **c** Aspect map of the study area. **d** The topographic wetness index (TWI) map of the study area. **e** The stream power index (SPI) map of the study area. **f** Map showing the distances to Abu Jir fault (m). **g** Map showing the distances to Euphrates River. **h** Total head distribution over the study area (after Al-Jiburi and Al-Basrawi 2008). **i** Spatial distribution of groundwater depth in the study area

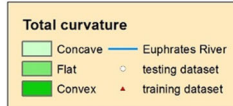
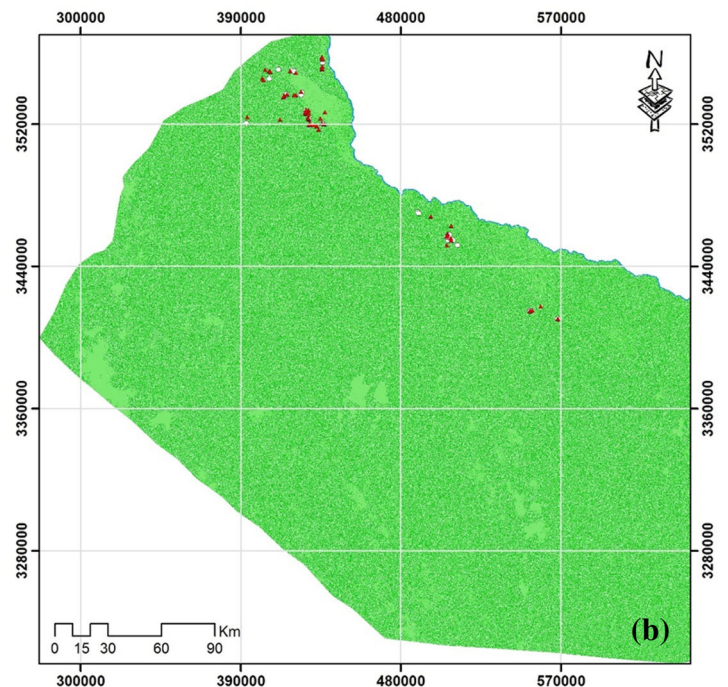
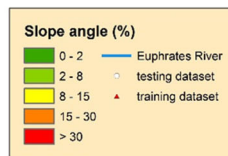
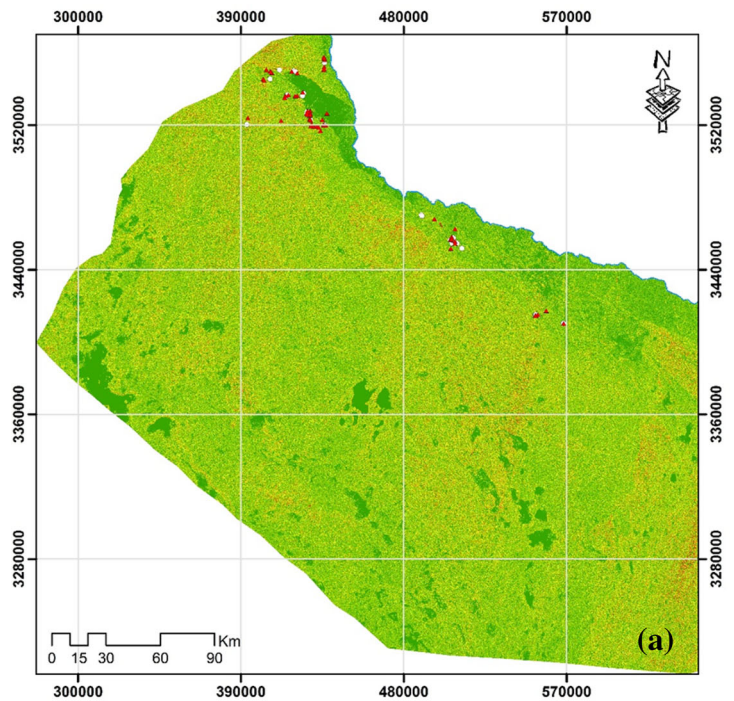


Fig. 6 continued

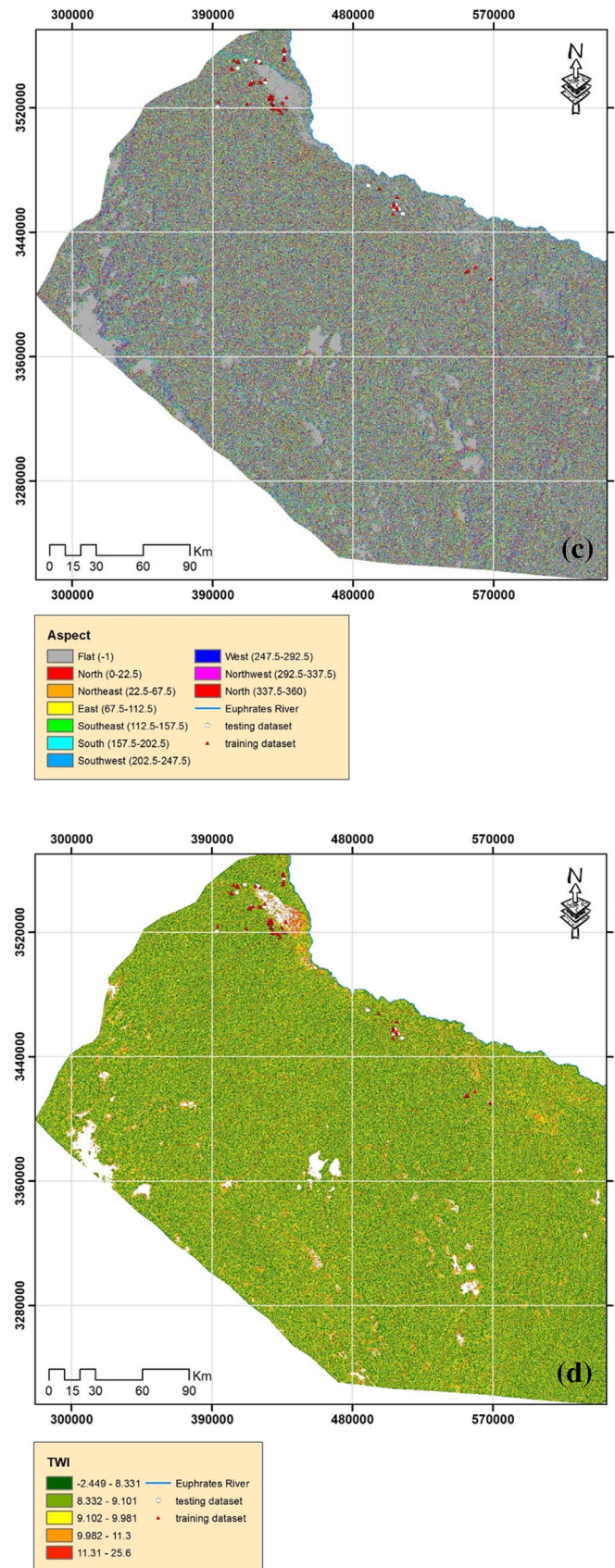


Fig. 6 continued

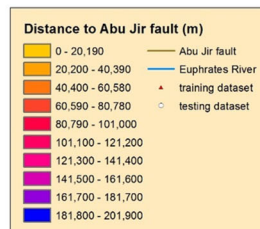
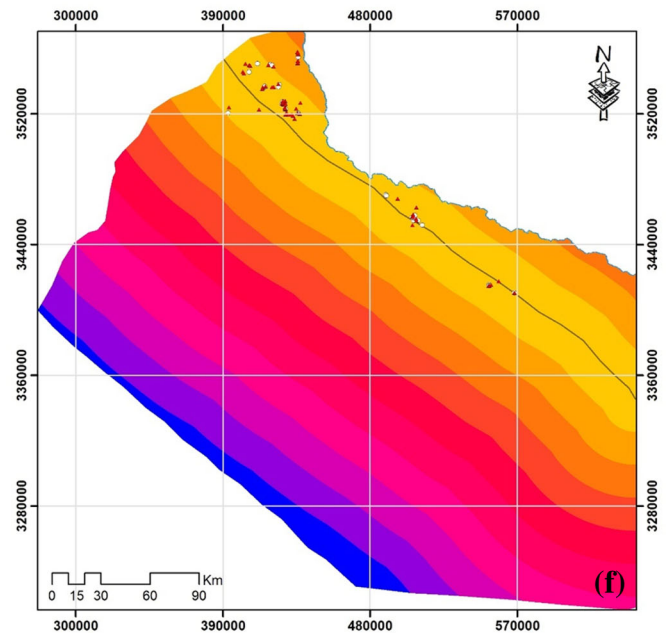
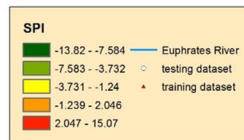
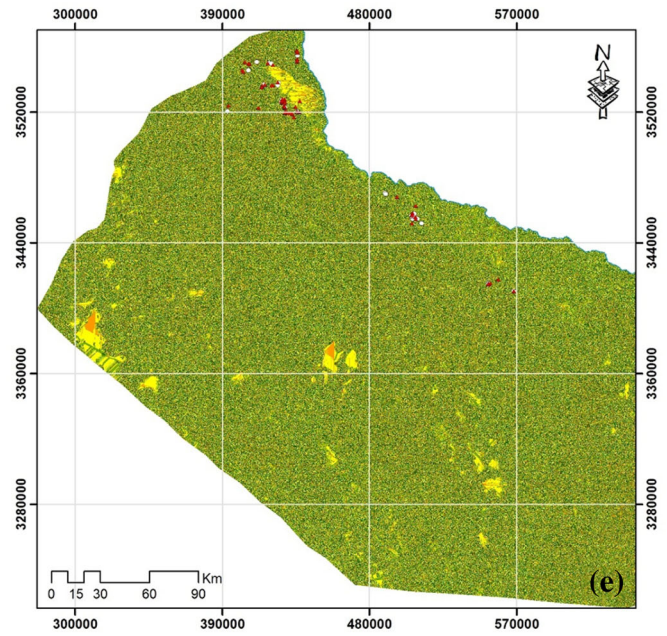
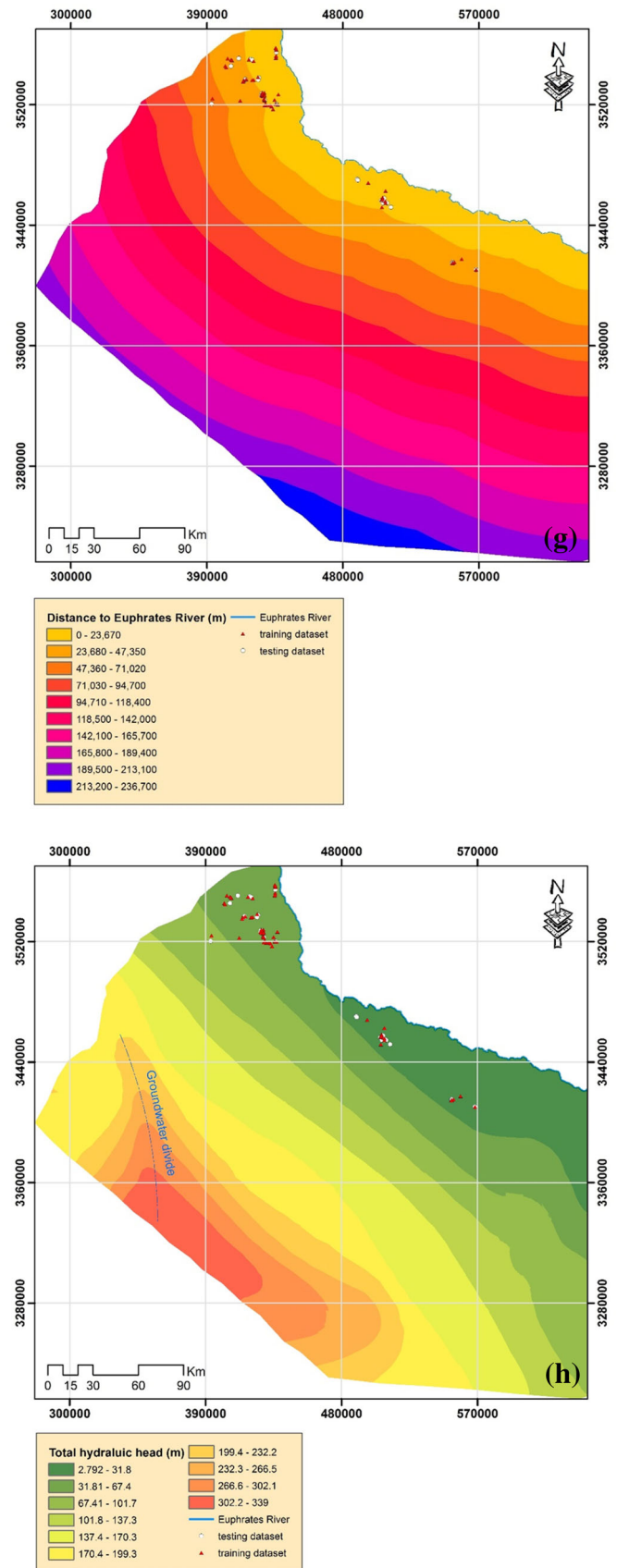


Fig. 6 continued



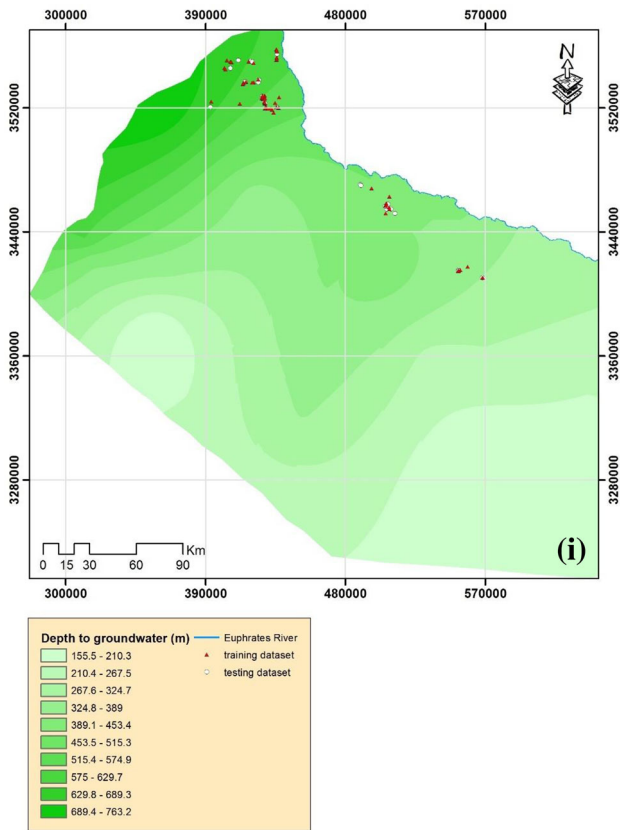


Fig. 6 continued

$$TWI = \ln\left(\frac{a}{\tan \beta}\right) \tag{1}$$

$$SPI = A_s \tan \beta \tag{2}$$

where, a is the local unslope area draining through a certain point per unit contour length and $\tan \beta$ is the local slope in degrees, and A_s is the specific catchment area. The *Raster Calculator* in ArcGIS software was used to derive TWI and SPI layers and finally classified into five categories to prepare the thematic maps of those variables as shown in Fig. 6d, e, respectively.

The proximity variables namely, distance to Abu Jir fault and distance to Euphrates River were prepared by using *Euclidean Distance* method in ArcGIS environment. Both variables were classified into ten classes using equal interval classification scheme (Fig. 6f, g). It was found that the correlation between well locations and distance from these linear features decreases as the distance increase, and thus a strong negative correlation exists.

Three predictor variables related to hydrogeological characteristics namely, total hydraulic head, well depth, and major aquifer groups were used in this study. Hard copy of total hydraulic head map (Al-Jiburi and Al-Basrawi 2008) was scanned, georeferenced, and then digitized

using ArcGIS software (Fig. 6h). Generally, groundwater moves from recharge areas (northeast) to discharge areas (southwest). Groundwater either discharges in form of springs or flows underground into Mesopotamian plan sediments (GEOSURV 1983). Data presented in Table 1 were used to create the map of groundwater depth. The values of groundwater depth were interpolated using inverse distance weighting technique (IDW) method to prepare the map. IDW is a deterministic interpolation technique traditionally used to interpolate groundwater depth (Reed et al. 2000). In IDW, deterministic interpolation techniques create surface from sample points using mathematical functions, based on either the extent of similarity or the degree of smoothing (radial basic function RBF) (Adhikary and Dash 2014). In mathematical terms, IDW is written as:

$$z(x_o) = \frac{\sum_{i=1}^n \frac{x_i}{h_{ij}^\beta}}{\sum_{i=1}^n \frac{1}{h_{ij}^\beta}} \tag{3}$$

where, $z(x_o)$ is the interpolated value, n is total number of sample data, x_i is the i th data value, h_{ij} is the separation distance between interpolated value and sample data value, and β is the weighting power factor. The optimal weighting power depends on the spatial structure of the data and is influenced by the coefficient of variation, skewness and kurtosis of the data (Mueller et al. 2001). The map of groundwater depth is shown in Fig. 6i. The figure shows that the depth to groundwater increases from south to north. To create the aquifer group layer, a hard copy of this predictive variable (Fig. 4) was scanned, georeferenced, digitized and finally convert from vector to raster using conversion tool in ArcGIS.

Mapping groundwater flowing well potential zone

The *randomForest* package in R software was used for the development of RF model. The target variable, the locations of well was represented by 1 for flowing wells and 0 for non-flowing wells. For the training of RF mode, equal number of flowing and non-flowing wells was selected to get the optimal result (Carranza and Laborte 2015b). Non-flowing locations should be distal to any flow location because locations proximal to existing potential zone are likely to have similar multivariate spatial data signatures as the potential zone and thus preclude achievement of desired results (Carranza and Laborte 2015a). Therefore, the point pattern analysis was used to find the distance from any flowing location and corresponding probability that there is one flowing location situated next to it. The point pattern analysis showed that the distance for any non-flowing well location in which there is 100 % probability

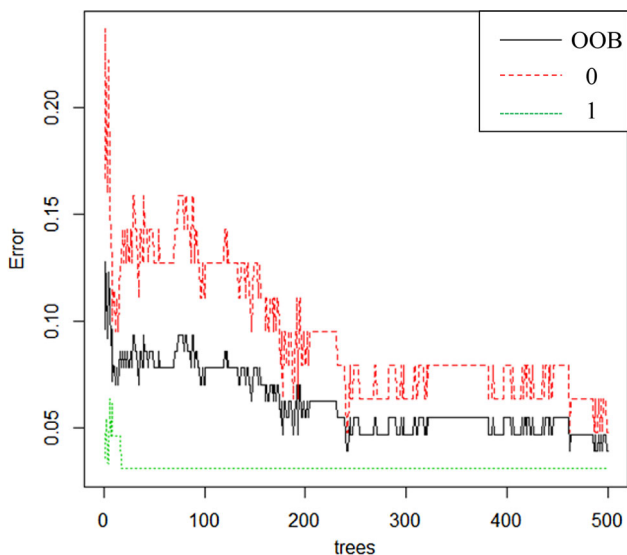


Fig. 7 Plot of out-of-bag (OOB) error of random forest model

of a neighboring flowing well location is approximately 10 km. Hence, the 65 non-flowing well locations were selected from areas beyond 10 km of every flowing well location. At flowing and non-flowing well locations, the values of predictor variables were extracted using ArcGIS. These data were stored as *.csv text file and exported to R statistical software to run RF model. Obtained results of RF model were the values between 0 and 1, where 1 refers to high probability of getting flowing well, while 0 refers to

non-flowing well. Finally, the probability of getting 1 were stored as text file and exported to the ArcGIS as point shape file. The point data were finally interpolated to prepare the map of groundwater flowing well potential.

Validation of the results

The accuracy of the RF model developed in this study to delineate groundwater flowing well zone are investigated using relative operating characteristics (ROC) curve. The ROC curve is commonly used for examining the quality of deterministic and probabilistic detection and forecast system (Swets 1988). It is a common method used to assess the accuracy of a diagnostic test (Egan 1975). The ROC plots the sensitivity (false positive rate) on X axis against 100-specificity (true positive rate) on Y axis. In ROC analysis, the area under the ROC curves (AUC) used to measure the prediction accuracy qualitatively (Maier and Dandy 2000). The predictive capability of a model is excellent if $AUC = 1-9$; very good 0.8-0.9; good 0.8-0.7; average 0.7-0.6; and poor 0.6-0.5 (Yesilnacar 2005). Usually, the AUC are used to evaluate the performance during both model training (success rate) and testing (prediction rate). The success rate explains how well the resulting groundwater potential map classified the area of flowing well locations during training (Al-Abadi 2015a), while prediction rate provides a measure of the accuracy of predictive model with unseen data (testing dataset).

Fig. 8 Mean decrease accuracy and mean decrease GINI of effective predictors

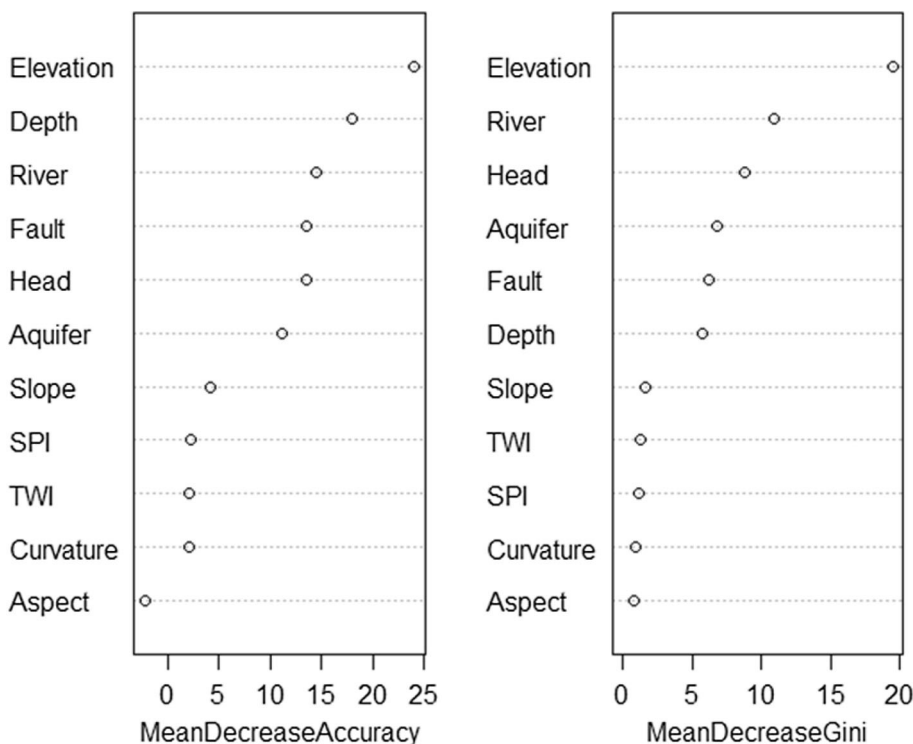
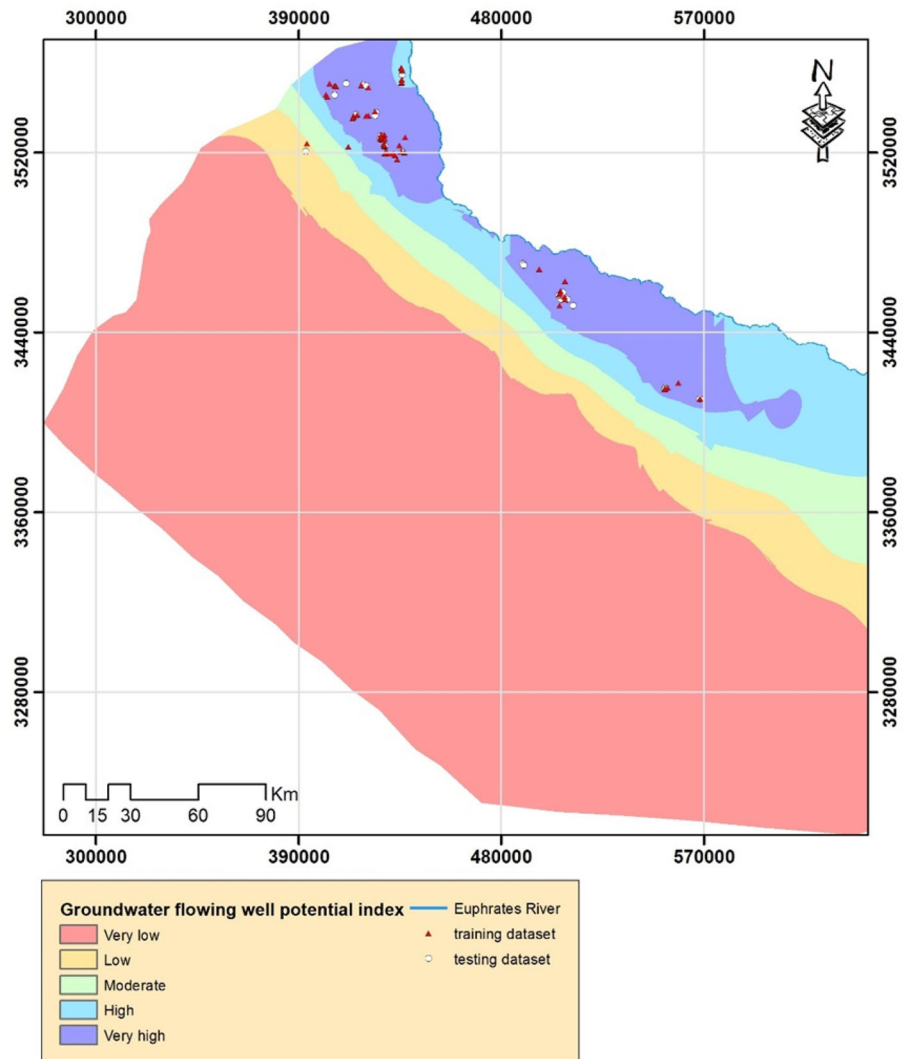


Table 2 Accuracy of RF models using all predictor and only the five most important predictors of target variable

Parameters	Developed RF models	
	All variables (model I)	The most important variables (model II)
OOB error	0.0391	0.0547
Correctly classified instances	120 (93.75 %)	122 (95.31 %)
Incorrectly classified instances	8 (6.25 %)	6 (4.68 %)
Kappa statistic	0.875	0.906
Mean absolute error	0.109	0.086
Root mean square error	0.231	0.215

Fig. 9 Groundwater flowing well potential index map of the study area



Results and discussion

The parameters required to run RF algorithm in R packages are the number of trees (*ntree*) and the number of predictors (*mtry*) randomly sampled at each split node. These parameters are determined using OOB error. The OOB error rate is a helpful estimator of the generalization error

depending on the number of trees (Rahmati et al. 2016). The OOB error rate depicted in Fig. 7 shows that the OOB error rate decreases as the number of trees increase. When the OOB equals to 0.01, *mtry* and *ntree* are equal to 3 and 500, respectively. The *mtry* value is further checked using the equation proposed by Breiman (2001), which postulates that the *mtry* should be less than $\log_2(M + 1)$ in order to

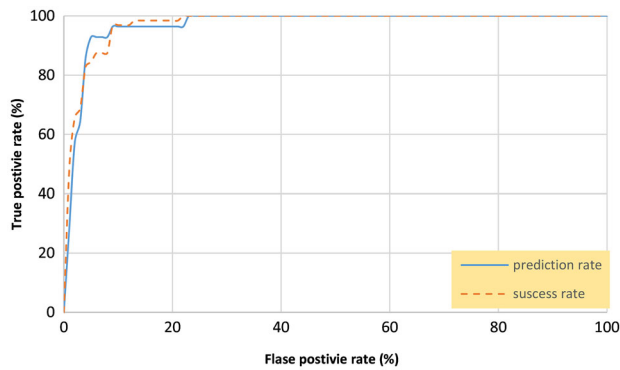


Fig. 10 ROC plot of random forest model

minimize the generalization error and correlation among decision trees, where M is the number of predictors used. As the number of predictors in this study is 11, the $mtry$ should be less than $\text{int}(\log_2(11 + 1))$ or 3. On the other hand, Rodrigues-Galiano et al. (2014) proposed that $n\text{tree}$ value of 1000 results relatively low prediction errors and most stable predictions (ZhenJie et al. 2015). From OOB plot in (Fig. 7), it is obvious that error rate stabilizes at 500 ($n\text{tree}$), and therefore, $n\text{tree}$ value of 500 was adopted in the study.

One of the most attractive features of the RF algorithm is its capability to rank the importance of predictors according to predictor's marginal effect on the target variable while keeping all the other predictors constant (Carranza and Laborte 2015a). In order to assess the predictor's importance in RF model developed in the study, two parameters were used namely, mean decrease accuracy (*MeanDecreaseAccuracy*) and mean decrease in Gini coefficient (*MeanDecreaseGini*) (Fig. 8). The mean decrease accuracy is a measure to explain how the model fit decreases as a variable dropped from the analysis. The greater the drop, the more significant the variable. On the other hand, the mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting RF model. The Gini index is often used to describe the overall explanatory power of the predictors. Therefore, mean decrease accuracy is more important for variable selection, while Gini index is important in defining the explanatory association among the variables selected. The mean decrease accuracy plot (Fig. 8) identified the elevation, well depth, distance to Euphrates River, Distance to Abu-Jir fault, aquifer groups, and groundwater heads as most importance predictors. The same results were obtained using Gini index with different ranks of importance. Both measures indicate that the slope, curvature, aspect, TWI, and SPI have lower effect on the groundwater flowing well potential in the study area. Therefore, these predictors were removed and the RF model was run again. The overall

accuracy for both RF runs are presented in Table 2. It is obvious from Table 2 that removal of the less important predictors caused an increase in model accuracy from 93.75 % (model I) to 95.31 % (model II). According to Landis and Koch (1977) the coefficient is the best index of fit between the predictors and the target variable. Kappa values >0.8 = strong fit, $0.4-0.8$ = moderate fit, and <0.4 = poor fit. Both RF model have Kappa coefficients greater than 0.8 and thus regard as excellent accuracy models but the model II more accurate than model I (0.87 versus 0.91 kappa coefficient, for models I and II, respectively). Therefore, results of the RF model II were used for further analysis. The RF model gave a probability value between 0 and 1 at 128 observation points. In ArcGIS, the probability of getting 1 was interpolated using IDW algorithm to generate the flowing well potential map, as shown in (Fig. 9). The probability values were classified into five categories using brake classification scheme namely, very low (0.00–0.11), low (0.11–0.34), moderate (0.34–0.61), high (0.61–0.83), and very high (0.83–1.0). It was found that the high and very high potential zones cover an area of about 11,868 km² or 15 % of the total area, which are mainly concentrated in northeastern parts of the study area. The moderate zones encompass total area of about 4342 km² or 6 % of total study area. The majority of the study area, 62,180 km² or 79 % of total area was identified as very low and low potential for groundwater flowing well. The low potential zones are found to concentrate mainly in the southwestern parts of the study area. The high and very high potentiality zones are basically associated with low elevation values, closeness to the Abu Jir fault, closeness to the Euphrates River, aquifer groups 10 and 4, and low hydraulic heads.

The plot of ROC curves for RF model is shown in (Fig. 10). The AUC for success and prediction rates were 0.98 and 0.97, respectively, which correspond to 98 and 97 % accuracy, respectively. This indicates the excellent capability of RF model in delineating groundwater flowing well potential zone in the study area.

Conclusions

The efficacy of RF machine learning technique in demarcating flowing well potential zone at southern desert of Iraq has been investigated in the present study. The spatial associations between target variable (locations of flowing wells) and set of predictors (groundwater occurrence controlling factors) were used to model groundwater potential using RF model. Eleven predictor variables namely, elevation, slope angle, curvature, aspect, TWI, SPI, distance to Abu Jir fault, distance to Euphrates River, major aquifer group, total hydraulic head, and well depth were used for

demarcation of groundwater flowing well potential zones. The study revealed that elevation, well depth, distance to Euphrates River, distance to Abu-Jir fault, aquifer groups, and groundwater heads are the most importance predictors, while the slope, curvature, aspect, TWI, and SPI have less influence in delineating groundwater potential in the study area. The groundwater flowing well potential index map shows that the high to very high, moderate, and low to very low potential zones occupy 15, 6, and 79 % of the total area of southern desert of Iraq, respectively. The validation of RF model using ROC curve revealed that the AUC's of success and prediction rates were 0.98 and 0.97, respectively, which indicate the excellent capability of RF model in delineating groundwater flowing well potential in GIS. It is expected that the map developed in this study will provide valuable information for the development of groundwater resources to solve the long lasting water scarcity in the region. In future, the performance of RF model can be compared with other conventional methods to show its efficacy. Furthermore, other state of art data driven models can also be used in future for mapping groundwater potential zone.

References

- Abdalla F (2012) Mapping of groundwater prospective zones using remote sensing and GIS techniques: a case study from the Central Eastern Desert, Egypt. *J. Afr Earth Sci* 70:8–17
- Adiat KAN, Nawawi MNM, Abdullah K (2012) Assessing the accuracy of GIS-based elementary multi criteria decision analysis as a spatial prediction tool—a case of predicting potential zones of sustainable groundwater resources. *J Hydrol* 440:75–89. doi:10.1016/j.jhydrol.2012.03.028
- Adhikary PP, Dash CJ (2014) Comparison of deterministic and stochastic methods to predict spatial variation of groundwater depth. *Appl Water Sci*. doi:10.1007/s13201-014-0249-8
- Al-Abadi AM (2015a) Groundwater potential mapping at northeastern Wasit and Missan governorates, Iraq using a data-driven weights of evidence technique in framework of GIS. *Environ Earth Sci*. doi:10.1007/s12665-015-4097-0
- Al-Abadi AM (2015b) Modeling of groundwater productivity in northeastern Wasit Governorate. Iraq by using frequency ratio and Shannon's entropy models. *Appl Water Sci*. doi:10.1007/s13201-015-0283-1
- Al-Abadi AM, Al-Shamma'a A (2014) Groundwater potential mapping of the major aquifer in northeastern Missan governorate, south of Iraq by using analytical hierarchy process and GIS. *J. Environ Earth Sci* 10:125–149
- Al-Ansari N (2013) Management of water resources in Iraq: perspectives and prognoses. *J Eng* 5(8):667–668
- Al-Jiburi HKS, Al-Basrawi NH (2000) Hydrogeological and hydrochemical study of Al-Najaf Quadrangle, sheet NH-38-2, scale 1: 250 000. GEOSURV, Int. Rep. No. 2705
- Al-Jiburi HK, Al-Basrawi NH (2008) Hydrology. In: *Geology of Iraqi southern desert*. Iraqi Bulletin of Geology and Mining. Special issue, pp 77–91
- Araim HI (1984) Regional hydrogeology of Iraq. GEOSURV, Internal Report, No. 1450
- Bonham-Carter GF (1994) *Geographic information systems for geoscientists: modeling with GIS*. Pergamon inc., New York, p 416
- Breiman L (1984) *Classification and regression trees*. Chapman & Hall/CRC, London
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Buday T, Jassim SZ (1987) *The regional geology of Iraq, vol 2., Tectonism, magmatism, and metamorphism* Publication of GEOSURV, Baghdad, p 352
- Carranza EJM, Laborte AG (2015a) Data-driven predictive modeling of mineral prospectivity using random forests: a case study in Catanduanes Island (Philippines). *Nat Resour Res*. doi:10.1007/s11053-015-9268-x
- Carranza EJM, Laborte AG (2015b) Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). *Comput Geosci* 74:60–70
- Corsini A, Cervi F, Ronchetti F (2009) Weight of evidence and artificial neural networks for potential groundwater mapping: an application to the Mt. Modino area (northern Apennines, Italy). *Geomorphology* 111:79–87. doi:10.1016/j.geomorph.2008.03.015
- de Winnaar G, Jewitt GPW, Horan M (2007) A GIS-based approach for identifying potential runoff harvesting sites in the Thukela River basin, South Africa. *Phys Chem Earth* 32:1058–1067
- Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach Learn* 40(2):139–157
- Egan JP (1975) *Signal detection theory and ROC analysis*. Academic Press, New York
- 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
- Fernandez-Delgado M, Carnada E, Barro S (2014) Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res* 15:3133–3181
- Genier R, Poggi J-M, Tuleau C (2008) Random Forests: Some methodological insights. [arXiv:0811.3619](https://arxiv.org/abs/0811.3619)
- GEOSURV (1983) Hydrogeology, hydrochemistry and water resources in the southern desert (blocks 1, 2, 3). GEOSURV, Int. Rep. Nos. 1250–1256
- Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, p 745
- Jassim SZ, Goff JC (2006) *Geology of Iraq*. Dolin, Prague and Moravian Museum, Brno, p 431
- Jha MK, Chowdary VM, Chowdhury A (2010) Groundwater assessment in Salboni Block, West Bengal (India) using remote sensing, geographical information system and multi-criteria decision analysis techniques. *Hydrogeo J* 18:1713–1728. doi:10.1007/s10040-010-0631-z
- Landis J, Koch G (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–174
- Lee S, Lee C-W (2015) Application of decision-tree model to groundwater productivity-potential mapping. *Sustainability* 7:13416–13432. doi:10.3390/su71013416
- Lee S, Kim YS, Oh HJ (2012) Application of a weight-of-evidence method and GIS to regional groundwater productivity potential mapping. *J Environ Manage* 96:91–105. doi:10.1016/j.jenvman.2011.09.016
- Liaw A, Wiener M (2002) Classification and regression by random forest. *R News* 2(3):18–22
- Machiwal D, Madan KJ, Bimal CM (2010) Assessment of groundwater potential in a semi-arid region of India using remote sensing, GIS and MCDM techniques. *Water Resour Manage* 25:1359–1386

- Maier HG, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environ Model Softw* 15:101–124
- Manap MA, Sulaiman WN, Ramli MF, Pradhan B, Surip N (2011) A knowledge-driven GIS modeling technique for groundwater potential mapping at the Upper Langat Basin, Malaysia. *Arab J Geosci* 6:1621–1637. doi:10.1007/s12517-011-0469-2
- McDonald RC, Isbell RF, Speight JG, Walker J, Hopkins MS (1990) Australian land and soil survey field handbook, 2nd edn. Inkata Press Pty Ltd, Melbourne
- McKay G, Harris JR (2015) Comparison of the data-driven random forests model and a knowledge-driven method for mineral prospectively mapping: a case study for gold deposits around the Huritz Group and Nueltin Suite, Nunavut, Canada. *Nat Resour Res*. doi:10.1007/s11053-015-9274-z
- Mogaji KA, Lim HS, Abdullah K (2014) Regional prediction of groundwater potential mapping in a multifaceted geology terrain using GIS-based Dempster–Shafer model. *Arab J Geosci*. doi:10.1007/s12517-014-1391-1
- Moghaddam DD, Rezaei M, Pourghasemi HR, Pourtaghie ZS, Pradhan B (2013) Groundwater spring potential mapping using bivariate statistical model and GIS in the Taleghan Watershed, Iraq. *Arab J Geosci*. doi:10.1007/s12517-013-1161-5
- Moore ID, Grayson RB, Ladson AR (1991) Digital terrain modeling – a review of hydrological, geomorphological, and biological applications. *Hydrol Process* 5:3–30
- Mueller TG, Pierce FJ, Schabenberger O, Warncke DD (2001) Map quality for site-specific fertility management. *Soil Sci Soc Am J* 65(5):1547–1558
- Naghibi SA, Pourghasemi HR, Pourtaghi ZS, Rezaei A (2014) Groundwater qanat potential mapping using frequency ratio and Shannon’s entropy models in the Moghan watershed, Iraq. *Earth Sci Inf*. doi:10.1007/s12145-014-0145-7
- 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
- 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
- Oh HJ, Kim YS, Choi JK, Park E, Lee S (2011) GIS mapping of regional probabilistic groundwater potential in the area of Pohang City, Korea. *J Hydrol* 399:158–172
- Ozdemir A (2011a) 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
- Ozdemir A (2011b) 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
- Pandey VP, Shrestha S, Kazama F (2013) A GIS-based methodology to delineate potential areas for groundwater development: a case study from Kathmandu Valley, Nepal. *Appl Water Sci* 3:453–465. doi:10.1007/s13201-013-0094-1
- Parsons RM (1955) Groundwater resource of Iraq, vol 4. Development Board, Ministry of Development, Government of Iraq, Kirkuk liwa, p 142
- Peters J, Baets BD, Verhoest NEC, Samson R, Degroove S, Becker PD, Huybrechts WH (2007) Random forests as a tool for ecohydrological distribution modelling. *Ecol Model* 207:304–318
- 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(6):662–685. doi:10.1080/10106049.2014.966161
- Pourtaghi ZS, Pourghasemi HR (2014) GIS-based groundwater spring potential assessment and mapping in the Birjand Township, southern Khorasan Province, Iran. *Hydrogeol J* 22:643–662. doi:10.1007/s10040-013-1089-6
- Rahmati O, Samani AN, Mahdavi M, Rourghasemi HR, Zeinivand H (2014) Groundwater potential mapping at Kurdistan region of Iran using analytic hierarchy process and GIS. *Arab J Geosci* 8(9):7059–7071. doi:10.1007/s12517-014-1668-4
- 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
- Reed P, Minsker B, Valocchi AJ (2000) Cost-effective long-term groundwater monitoring design using a genetic algorithm and global mass interpolation. *Water Resour Res* 36(12):3731–3741
- Rodriguez-Galiano V, Chica-Olmo M (2012) Land cover change analysis of a Mediterranean area in Spain using different sources of data: multi-seasonal Landsat images, land surface temperature, digital terrain models and texture. *Appl Geogr* 35:208–218
- Shahid S, Nath SK, Kamal AS (2002) GIS integration of remote sensing and topographic data using fuzzy logic for ground water assessment in Midnapur District, India. *Geocarto Int* 17:69–74. doi:10.1080/10106040208
- Sissakian VK (2000) Geological Map of Iraq, Scale 1:1000000. GEOSURV, Baghdad
- Swets JA (1988) Measuring the accuracy of diagnostic systems. *Science* 240:1285–1293
- Yesilnacar EK (2005) The application of computational intelligence to landslide susceptibility mapping in Turkey. PhD thesis, Department of Geomatics the University of Melbourne, p 423
- ZhenJie A, RenGuang Z, YiHui X (2015) A comparative study of fuzzy weights of evidence and random forests for mapping mineral prospectivity for skarn-type Fe deposits in the southwestern Fujian metallogenic belt, China. *Sci China Earth Sci* 59(3):556–572. doi:10.1007/s11430-015-5178-3