



Assessment of groundwater potential in terms of the availability and quality of the resource: a case study from Iraq

Alaa M. Al-Abadi^{1,2} · Alan E. Fryar¹ · Arjan A. Rasheed³ · Biswajeet Pradhan^{4,5,6,7}

Received: 7 January 2021 / Accepted: 27 May 2021

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

A semi-confined aquifer from Kirkuk Governorate, northern Iraq was taken as a case study to map groundwater potential in terms of both the availability and quality of the resource. In terms of quantity, five machine learning (ML) algorithms were used to model the relationship between locations of 1031 wells with specific-capacity data and nine influential groundwater occurrence factors. The algorithms used were linear discriminant analysis, classification and regression trees, linear vector quantization, random forest, and K-nearest neighbor. The groundwater occurrence factors used were elevation, slope, curvature, aspect, aquifer transmissivity, specific storage, soil, geology, and groundwater depth. Analysis of the worthiness of the factors used in the analysis by the information gain ratio indicated that five out of nine factors were worthy (average merit > 0): groundwater depth, elevation, transmissivity, specific storage, and soil. The remaining factors were non-worthy (average merit = 0) and thus they were removed from the analysis. The performance of the five ML algorithms was investigated using accuracy and kappa as evaluation metrics. Applying the models in the *carte* package of R software indicated that random forest was the best model. The probability values of this model were used for mapping quantitative groundwater potential after classification into three zones: poor, moderate, and excellent. Groundwater quality for drinking was modeled using the water quality index and the weights of the chemical constituents used (pH, TDS, Ca²⁺, Mg²⁺, Na⁺, SO₄²⁻, Cl⁻, and NO₃⁻) were assigned using entropy information theory. A map of the groundwater quality index revealed five classes: < 50 (excellent), 50–100 (good), 100–150 (moderate), 150–200 (poor), and > 200 (extremely poor). Combining the groundwater quality index map with the groundwater potential map using summation operators revealed three zones of groundwater potential: poor, moderate, and excellent. Comparing this combined map with the quantitative groundwater potential map showed different patterns for the distribution of potential classes, which confirms that analysis of the groundwater potential should include groundwater quality as an important factor.

Keywords Groundwater · Random forest · Water quality index · Kirkuk · Iraq

✉ Alaa M. Al-Abadi
Alaa.Al-Abadi@uky.edu; alaa.atiaa@uobasrah.edu.iq

Alan E. Fryar
alan.fryar@uky.edu

Arjan A. Rasheed
arjantuz@yahoo.com

Biswajeet Pradhan
Biswajeet.Pradhan@uts.edu.au

¹ Department of Earth and Environmental Sciences,
University of Kentucky, Lexington, KY 40506-0053, USA

² Department of Geology, College of Science, University
of Basrah, Basra, Iraq

³ General Commission of Groundwater, Kirkuk Branch,
Ministry of Water Resources, Kirkuk, Iraq

⁴ Centre for Advanced Modelling and Geospatial Information
Systems (CAMGIS), Faculty of Engineering and IT,
University of Technology Sydney, Sydney, NSW 2007,
Australia

⁵ Department of Energy and Mineral Resources Engineering,
Sejong University, Choongmu-gwan, 209 Neungdong-ro
Gwangjin-gu, Seoul 05006, South Korea

⁶ Center of Excellence for Climate Change Research, King
Abdulaziz University, P. O. Box 80234, Jidda 21589,
Saudi Arabia

⁷ Earth Observation Center, Institute of Climate Change,
Universiti Kebangsaan Malaysia (UKM), 43600 Bangi,
Selangor, Malaysia

Abbreviations

GAOFs	Groundwater-affecting occurrence factors
CART	Classification and regression trees
GIS	Geographic information system
GWQI	Groundwater quality index
KNN	K-nearest neighbor
LDA	Linear discriminant analysis
LU/LC	Land use/land cover
LVQ	Linear vector quantization
ML	Machine learning
TWI	Topographic wetness index

Introduction

In the last few years, groundwater potential mapping has become an essential geospatial tool for aquifer management. Groundwater potential analysis gives decision-makers and hydrogeologists simple rules for managing the aquifer system with cost-effective and simple-to-construct models. To model groundwater potential, researchers worldwide have used two broad classes of techniques: data-knowledge and data-driven methods (Davoudi Moghaddam et al. 2020). In data-knowledge techniques (e.g. simple weighted overlay, fuzzy logic, and multi-criteria decision making), a specific number of groundwater-affecting occurrence factors (GAOFs) are assigned weights and are linearly combined to produce the groundwater potential map (Shahid et al. 2002; Al-Abadi and Shahid 2015; Rahmati et al. 2015; Aouragh et al. 2017; Arabameri et al. 2019). There is not yet consensus concerning the number of factors to be used, which varies depending on the nature of the problem being solved and the data availability (Kordestani et al. 2019; Termeh et al. 2019). In the data-driven approach, the maps of groundwater potential are constructed through the study of the relationship between the locations of pumping wells with specified discharge rates or specific capacity data as target variables and the GAOFs as predictors (Kim et al. 2019; Razavi-Termeh et al. 2019; Arabameri et al. 2019; Lee et al. 2020; Panahi et al. 2020). With the development of geographic information system (GIS) technology and remote sensing techniques, groundwater potential mapping becomes easier to implement without the need for very costly and time-consuming field surveys. GIS is a powerful technology for handling spatial and associated non-spatial (attribute) information (King 1991). Remote sensing is a significant source of information on Earth's surface features related to the groundwater occurrence, such as the presence of lineaments, land use/land cover (LU/LC) characteristics, and geomorphological information (Oh et al. 2011). This information can be easily integrated with other data types in GIS and then analyzed (Jha et al. 2007; Das 2019).

Groundwater potential maps are mainly used as a management tool for aquifer systems through evaluating which portion of the aquifer is more productive, i.e. as an indicator of groundwater availability. However, to manage an aquifer efficiently, information on groundwater quality is also needed. Groundwater quality assessment is one of the first tasks that should be considered in groundwater studies (Singhal and Gupta 2010; Şen 2014; Fetter 2018) because of the need to determine the suitability of this water for various uses such as drinking, agriculture, and industry. Water quality in highly-productive aquifers may be impaired by salinity (Mehta et al. 2000; Prinos et al. 2014), non-point-source pollutants such as nitrate from agriculture (Puckett et al. 2011), and geogenic pollutants such as arsenic (Mukherjee et al. 2006; Fendorf et al. 2010; Erban et al. 2013; Schaefer et al. 2017).

In this study, a methodology is introduced for modeling both quantity and quality aspects in a groundwater potential assessment. To show the advantages of the proposed approach, a semi-confined aquifer from Kirkuk Governorate in northern Iraq is taken as a case study. For modeling the groundwater potential in terms of quantity, five machine learning (ML) algorithms were used: linear discriminant analysis (LDA); classification and regression trees (CART); linear vector quantization (LVQ); random forest (RF); and K-nearest neighbor (KNN). For modeling the groundwater potential in terms of quality, the groundwater quality index (GWQI) was used. GWQI considers the composite influence of individual parameters in deriving a single number that describes overall water quality both spatially and temporally (Coletti et al. 2010). The novelty of this research lies in the fact that it considers both the quantitative and qualitative dimensions of groundwater in the study of groundwater potential, rather than relying solely on groundwater quantity as a deciding factor in aquifer management.

Characteristics of the study area

The study area is the 422-km² Lailan basin (35°07'–35°29' N; 44°30'–44°40' E) in the southern part of Kirkuk province, Iraq (Fig. 1), approximately 255 km north of Baghdad. The basin is bounded by the Kirkuk structure to the northeast, the Jambur anticline to the southwest, and two streams on the north and south (Khassa Chai and Tawuq Chai, respectively). The basin is almost flat in the middle and becomes hilly to the southwest, with an elevation varying between 254 and 410 m above mean sea level (amsl) (Fig. 2). The area has a semi-arid climate (BSh in the Köppen–Geiger classification system). The annual average rainfall of the area is 347 mm/year. Units exposed in the basin, which range from Miocene to Holocene in age, include the Fatha, Injana, Mudkadiya, and Bai Hassan formations and

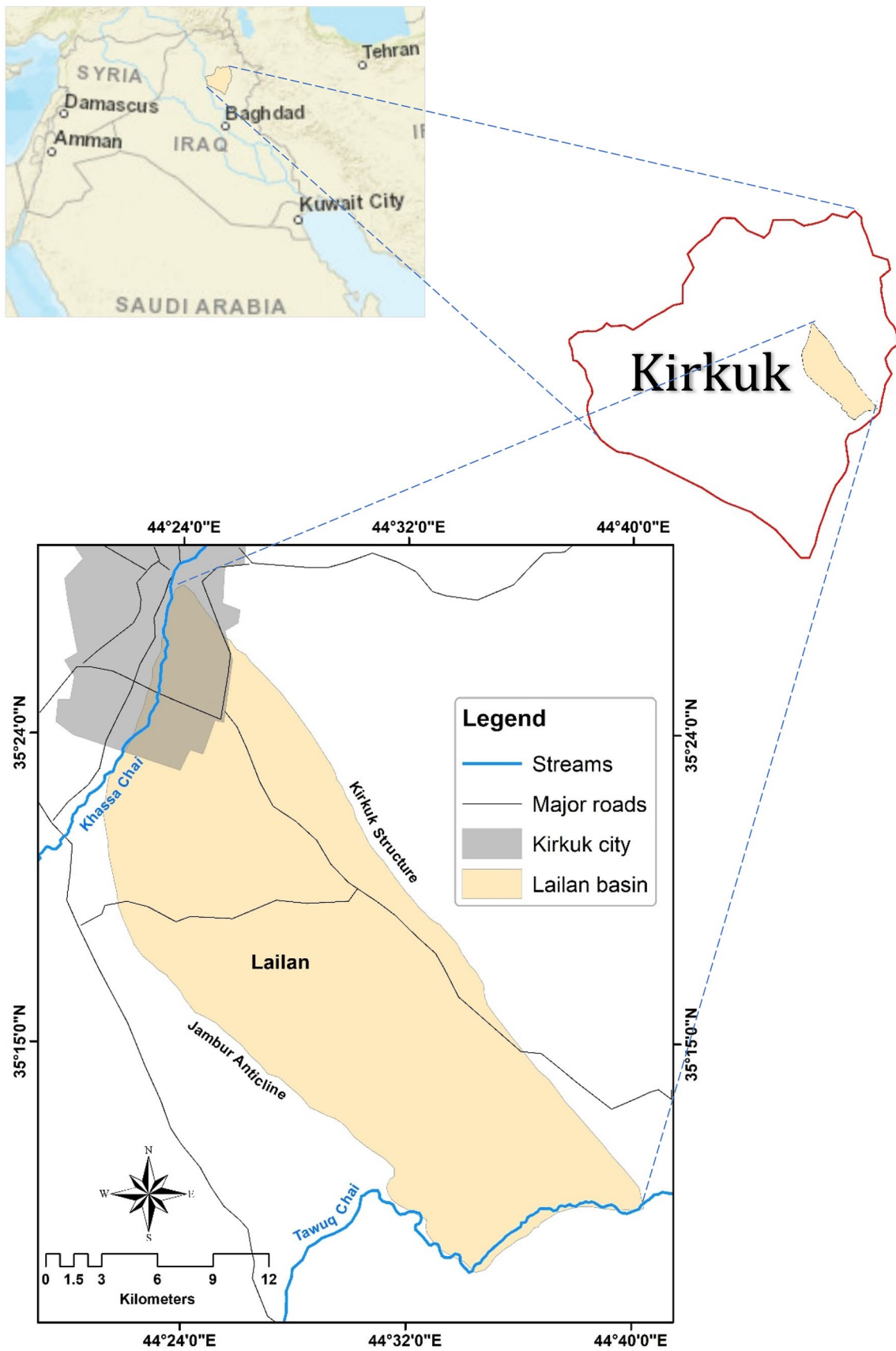


Fig. 1 Location of the Lailan basin

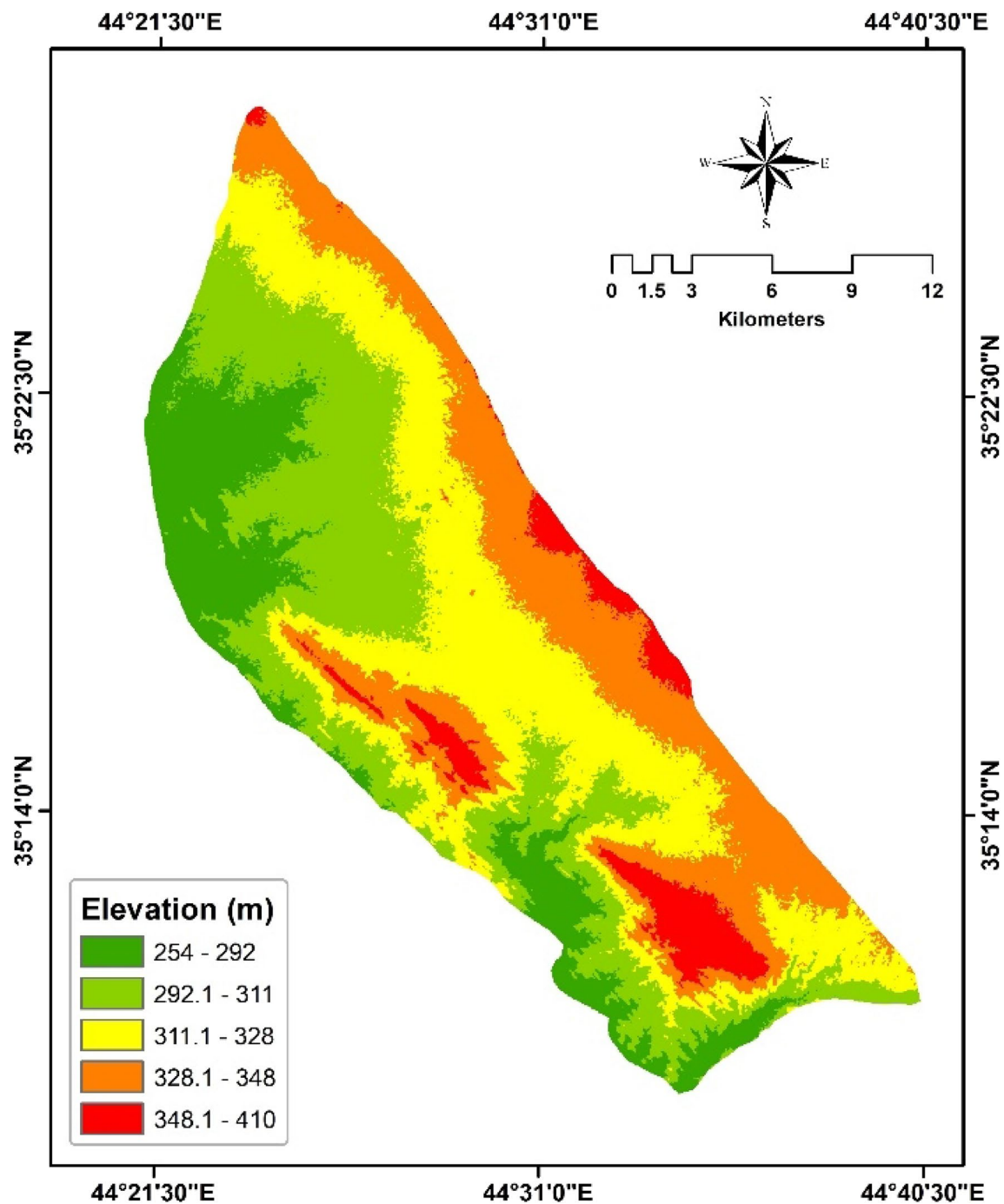


Fig. 2 Elevation of the Lailan basin (in m amsl)

Quaternary sediments (Table 1). The oldest units are usually exposed on the edge of the basin, while the Quaternary sediments cover the middle part of it (Fig. 3). The dominant LU/LC classes are unoccupied land, cropland and urban areas (Fig. 4). Five soil types are recognized in the study area, with reddish-brown soils (S27) covering approximately half of the area (Table 2). The texture of these soils is silty loam (Jasem et al. 2016).

The aquifer system in the Lailan basin occurs in the Bai Hassan, Mukdadiya, and Quaternary units. The main semi-confined aquifer occurs within the Bai Hassan Formation (Rasheed 2019). The average saturated thickness of the aquifer is 45 m. The groundwater depth in the basin varies from 38 to 63 m and generally increases from northeast to southwest (Fig. 5a). Groundwater flows from east to west and southeast to northwest, with hydraulic head

Table 1 Geological overview of the study area (after Jassim and Goff 2006 and Buday and Jassim 1980)

Formation/geological unit	Description	Age	Depositional environment
Fatha	Cyclic deposits of anhydrite, gypsum, claystone, limestone, sandstone, and marl	Middle Miocene	Rapidly subsiding sag basin
Injana	Fine-grained pre-molasse sediments deposited initially in coastal areas, and later in fluvio-lacustrine systems	Late Miocene	Anastomosing rivers
Mukdadiya	Fining upward cycles of gravelly sandstone, sandstone, and red mudstone	Pliocene	Fluvial environment in a rapidly subsiding foredeep basin
Bai Hassan	Thick layers of gravel or conglomerates interbedded with sandstone, siltstone, and claystone	Pliocene	Alluvial fans originated from the high folded zone and the Zagros suture
Quaternary deposits (Sheet runoff/Slope deposits)	A mixture of gravel, sand, silt, and clay	Pleistocene–Holocene	–

ranging from 371 to 274 m amsl (Fig. 5b). The strategic groundwater storage is $50 \times 10^6 \text{ m}^3$ and the renewable storage is $11 \times 10^6 \text{ m}^3$, taking into consideration that the annual groundwater recharge is 25 mm/year (Rasheed 2019). Groundwater assessment is critical within the basin, since groundwater supports drinking, aquaculture, and irrigation, and many people living in the basin rely on farming for their livelihood.

Materials and methods

The proposed methodology in this study has three distinct phases (Fig. 6). The *first phase* includes the following steps: (i) preparing the database of the operating groundwater wells and GAOFs; (ii) testing the importance of the GAOFs in the analysis of the groundwater potential using the information gain-ratio feature-selection approach; (iii) applying, comparing and validating ML models to select the best model; and (iv) mapping the groundwater potential using the best performing ML model. The *second phase* of the analysis involves: (i) the collection of groundwater samples and analysis for chemical parameters using standard methods; (ii) calculating the weight of each chemical parameter using the entropy theory approach; and (iii) the calculation of the GWQI and interpolation of the results to show the GWQI spatial distribution in the basin. In the *third phase*, the maps of groundwater potential and GWQI are combined to show the groundwater potential of the study area in terms of quantity and quality.

Groundwater operating wells inventory

The archive of the General Commission of Groundwater in Kirkuk and field surveys were relied on to prepare an inventory map of the 1031 groundwater operating wells. The specific capacity of these wells, which ranges from 0.60 to 5.36

L/s/m, was partitioned into high-potential ($> 2 \text{ L/s/m}$) and low-potential ($\leq 2 \text{ L/s/m}$) categories (Al-Abadi et al. 2019). For the classification problem adapted here for modeling groundwater potential, wells with high and low potential were coded as *yes* and *no*, respectively. After assigning the appropriate code for each well, the total number of wells was then randomly partitioned into two groups: 70% of the data (721 wells) were utilized for ML model training, while the remaining 30% (310 wells) were allocated for testing.

Preparing groundwater-affecting occurrence factors

There is no consensus about the number and type of GAOFs used in the analysis of groundwater potential. In general, the types of GAOFs used can be classified into two broad categories. The first is surficial factors, which primarily affect the groundwater recharge (renewable storage of the aquifer), such as topographic factors [elevation, slope, curvature, aspect, topographic wetness index (TWI), stream power index], drainage density, geology, soil, LU/LC, normalized difference vegetation index, mean annual rainfall, and distance to surface features such as streams. The second type includes factors controlling the strategic groundwater storage, such as the aquifer saturated thickness, hydraulic characteristics (transmissivity and storativity), distance to subsurface features such as faults (which may be associated with lineaments), and fault density. For this study, nine GAOFs were selected depending on the data availability: elevation, slope, curvature, aspect, TWI, soil, transmissivity, specific storage, and depth to groundwater. The importance of these factors, especially topographical factors (Al-Abadi et al. 2016; Naghibi et al. 2016; Khosravi et al. 2018), has been discussed extensively in other studies and is not repeated here. The elevation layer was created from a digital elevation model (DEM) with a $30 \times 30 \text{ m}$ grid size (Fig. 2). This was obtained from the NASA Shuttle Radar Topography Mission. The DEM was used as input to extract slope,

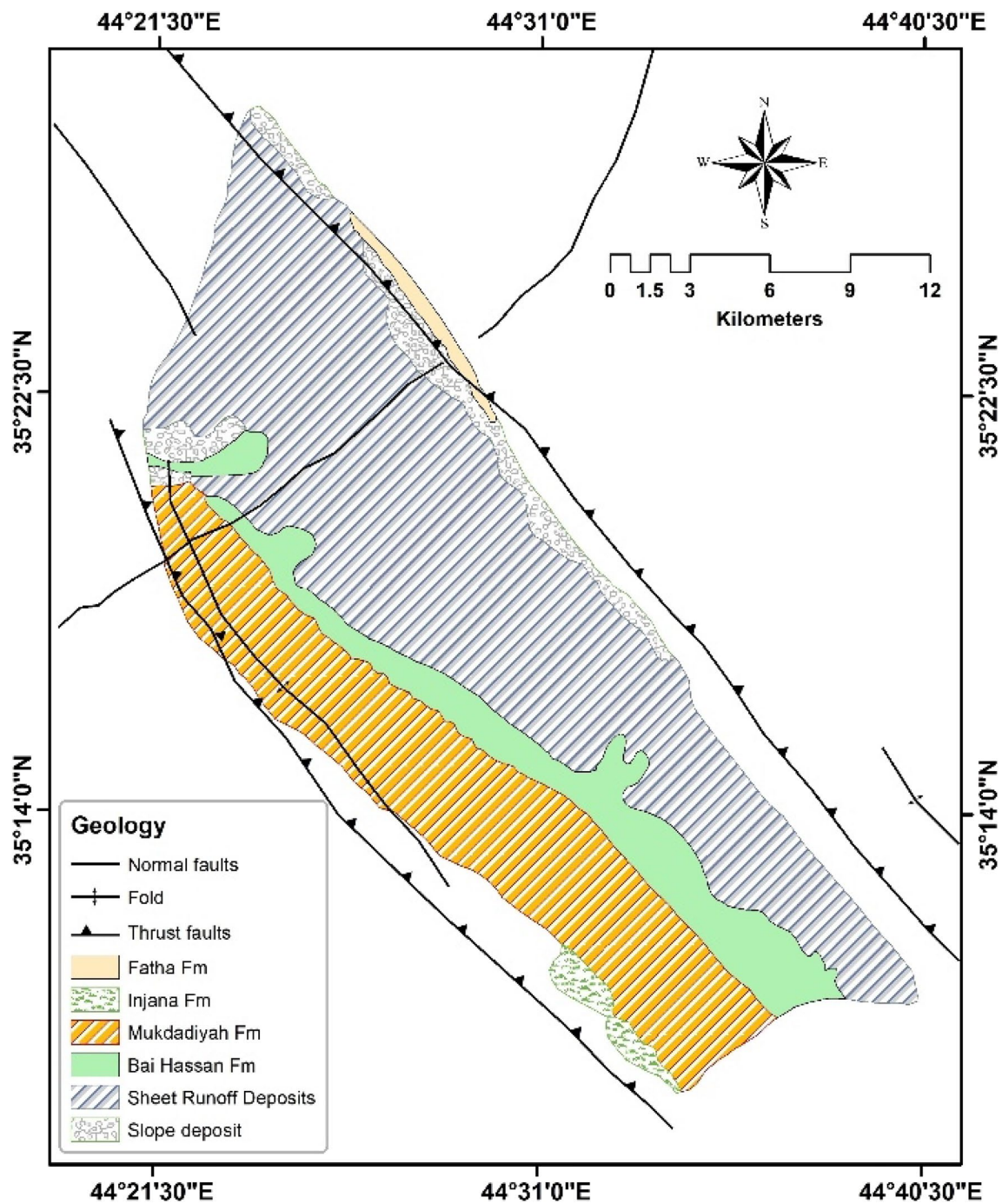


Fig. 3 Geology of the Lailan basin

curvature, aspect, and TWI (Fig. 8a–d). The geological map of Iraq (1:1,000,000 scale) was obtained from the Geological Survey of Iraq, and a digital copy was created using ArcGIS 10.5 (Fig. 3). The soil map (Fig. 7) was generated according to the work of Muhaimeed et al. (2014) and the exploratory soil map of Iraq with a scale of 1:1,000,000 (<https://esdac.jrc.ec.europa.eu/content/exploratory-soil-map-iraq-map-1>). To generate transmissivity and storage coefficient thematic

layers, pumping test data from 12 wells evenly distributed through the study area were analyzed by the Hantush and Jacob (1955) analytical solution for a leaky confined aquifer in AQTESOLV 4.5 software. The estimated transmissivity for the area was found to range from 0.003×10^{-3} to 1.1×10^{-2} with an average of $125 \text{ m}^2/\text{day}$, while the estimated storage coefficient ranged from 0.003×10^{-3} to 1.1×10^{-2} with an average of 2.6×10^{-3} . The obtained values of hydraulic

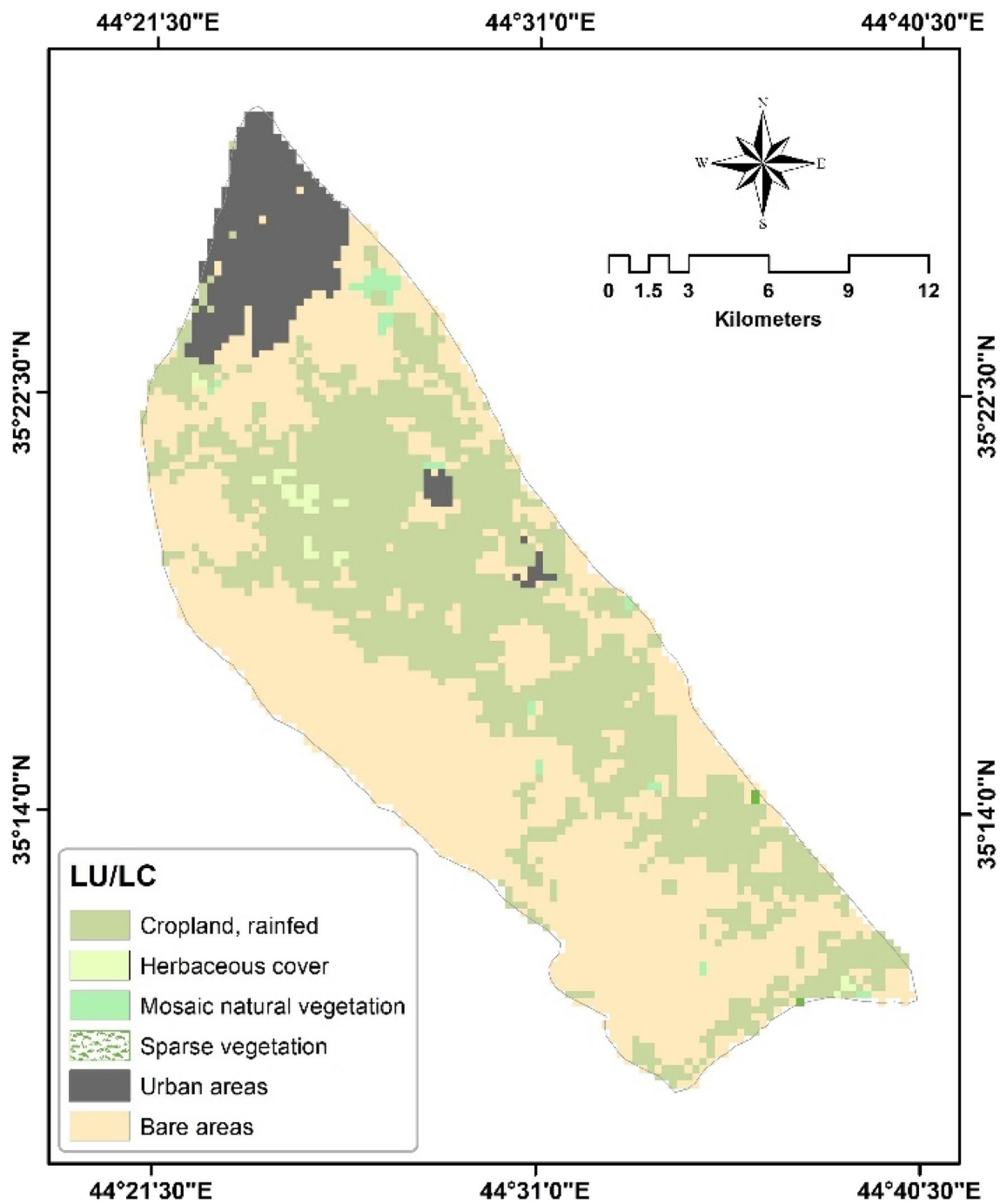


Fig. 4 LU/LC in the Lailan basin

characteristics were then interpolated in ArcGIS 10.5 throughout the studied basin using ordinary kriging (Fig. 8e, f). The transmissivity values generally increase from north to south and from east to west. On the other hand, there is no distinctive trend in the spatial distribution of specific storage coefficients. However, the central and northern parts of the study area generally have the highest values. Finally, the well log records were used for mapping depth to groundwater

(Fig. 5). All GAOFs were prepared as a raster grid with 30×30 m cells for applying the ML models. The total number of cells was 1,199,890 (970 columns and 1273 rows).

Groundwater sampling and analysis

Twenty-two groundwater samples were collected from operating wells in the basin 9–12 August 2017 and analyzed for

Table 2 Soil types in the study area

Code	Description	Occupied areas	
		km ²	%
S27	Reddish-brown soils, medium and shallow phase, over gypsum, sand, and mudstone	210	50
S28	Reddish-brown soils, deep phase	15	04
S31	Lithosolic soils in limestone	63	15
S33	Brown soils, medium and shallow phase over Mukdadiya and Bai Hasan gravel	67	16
S35	Brown soils, deep phase	67	16

chemical parameters using standard methods (Table 3). The groundwater samples were collected in 1-L polyethylene bottles. These bottles were first washed with dilute HNO₃ and then with distilled water. After that, the bottles were rinsed several times with the pumping groundwater for each well after 15 min of the pump running before taking the samples. Temperature, electrical conductivity (EC), and pH were measured on site using a calibrated multi-probe device. Samples were labeled and were kept in a field refrigerator for 2 days, then transferred to the laboratory of the General Commission of Groundwater in Kirkuk for chemical analysis

of major ions (Ca²⁺, Mg²⁺, Na⁺, K⁺, Cl⁻, SO₄²⁻, HCO₃⁻, and NO₃⁻). The charge-balance error of the groundwater samples was found to be within ± 10%.

Feature selection using information gain ratio

Feature selection techniques can be implemented in data-mining applications to reduce the dimensionality of the original data and enhance learning efficiency (Liu and Motoda 1998). By eliminating the irrelevant and redundant features in the data, feature selection accelerates data-mining algorithms, improves performance, and enhances model understandability (Zhao and Liu 2011). The information gain ratio is an improved version of gain information technique (Quinlan 2014). Information gain ratio measures the predictability of attributes (here GAOFs) by measuring the information gain for the class (Al-Abadi 2018). More information and mathematical background of this technique are found in Bui et al. (2016).

Machine learning algorithms

LDA is a generalized method of finding a linear combination of features that characterizes or distinguishes two or more groups of objects or events (Fisher 1936). The goal of the LDA is to project a feature space onto a smaller subspace

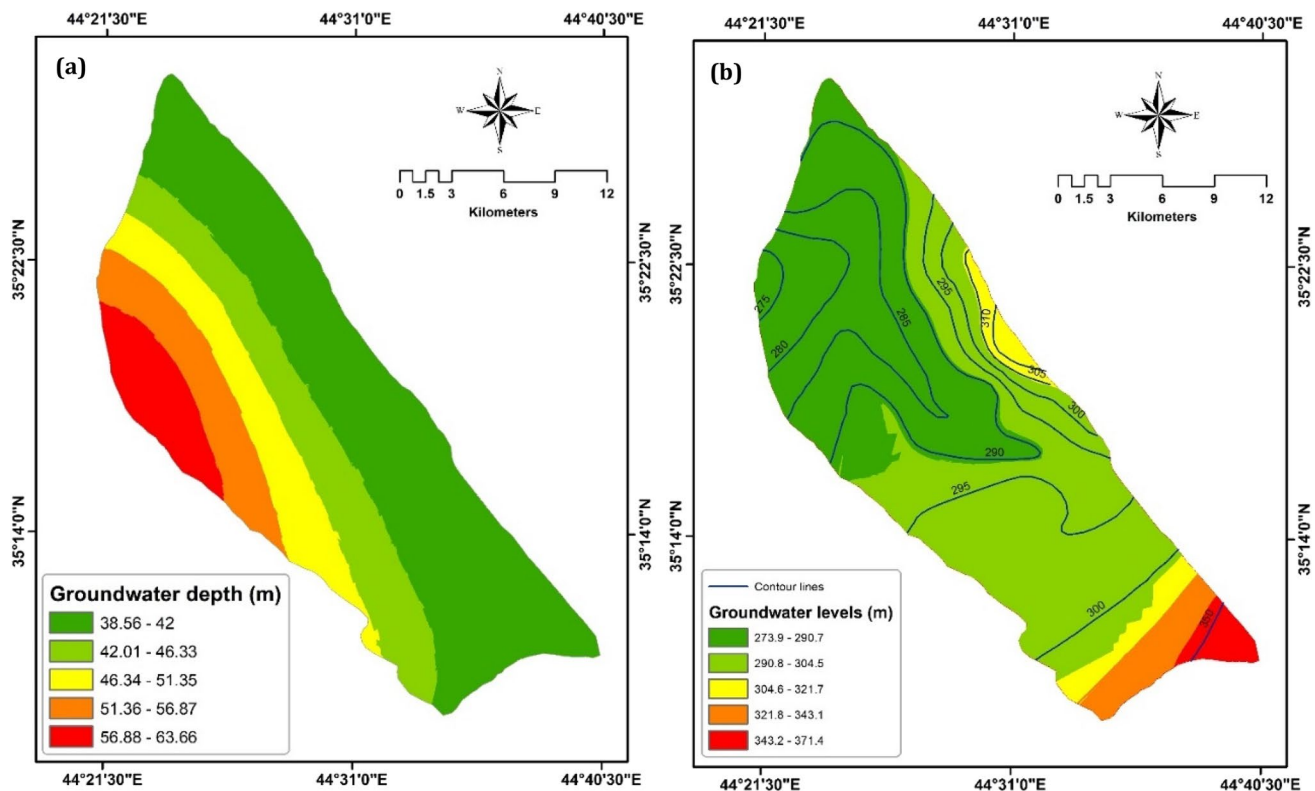


Fig. 5 **a** Depth to groundwater (m) and **b** groundwater levels (m amsl) in the study area

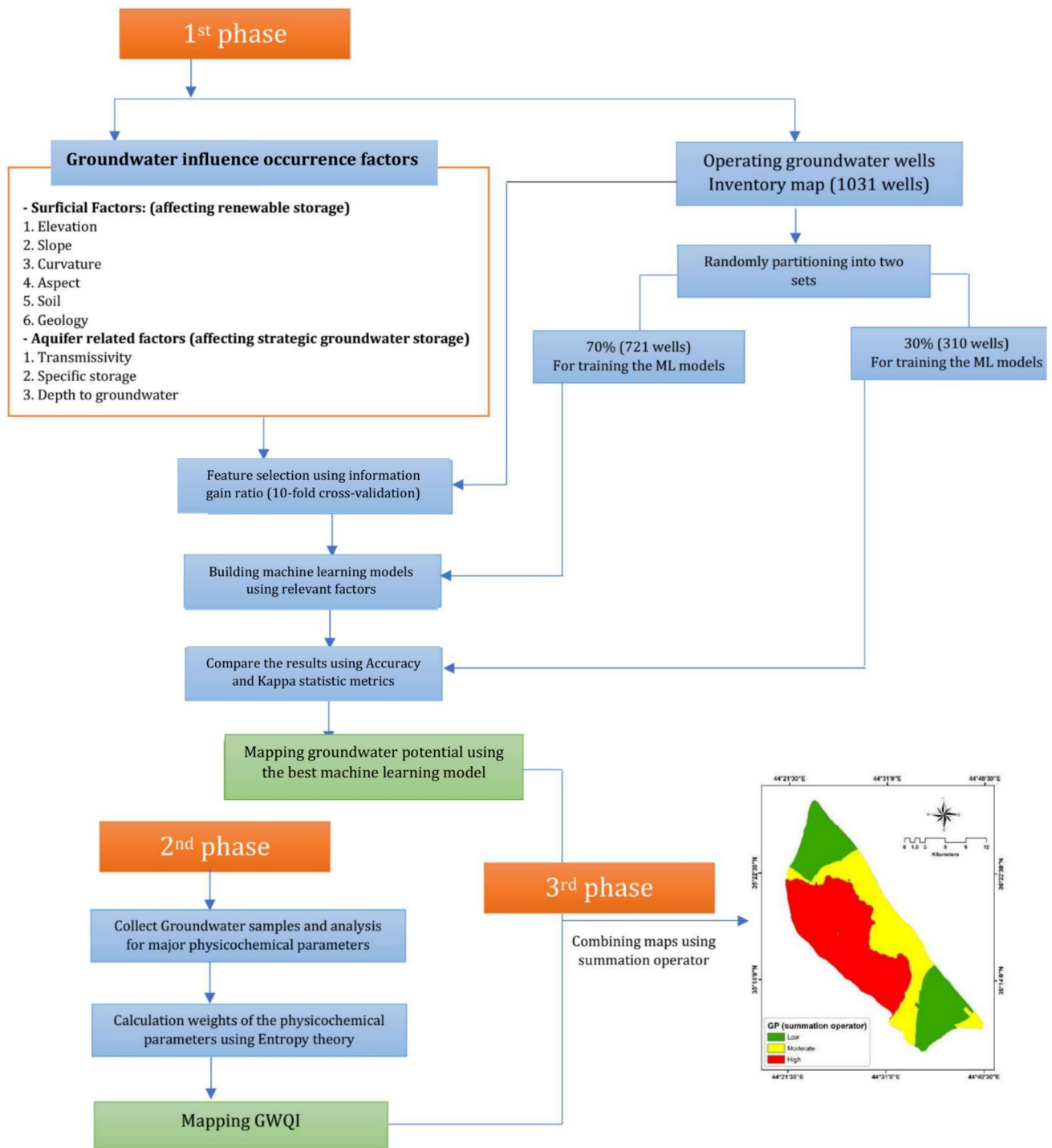


Fig. 6 Flowchart of the study procedure

(dimensional reduction) while maintaining the class-discriminatory information. The dimensional reduction helps to minimize computation cost and also contributes in reduction of overfitting. LDA has some assumptions about the data used (Brownlee 2016): (i) the input variables have a normal distribution (Gaussian distribution); (ii) the variance

determined by class grouping for each input variable is the same; and (iii) the mix of classes within the training dataset is representative of the problem.

CART is a term coined by Breiman et al. (1984) to refer to the decision tree algorithms in the predictive modeling techniques. CART can be used for solving problems in

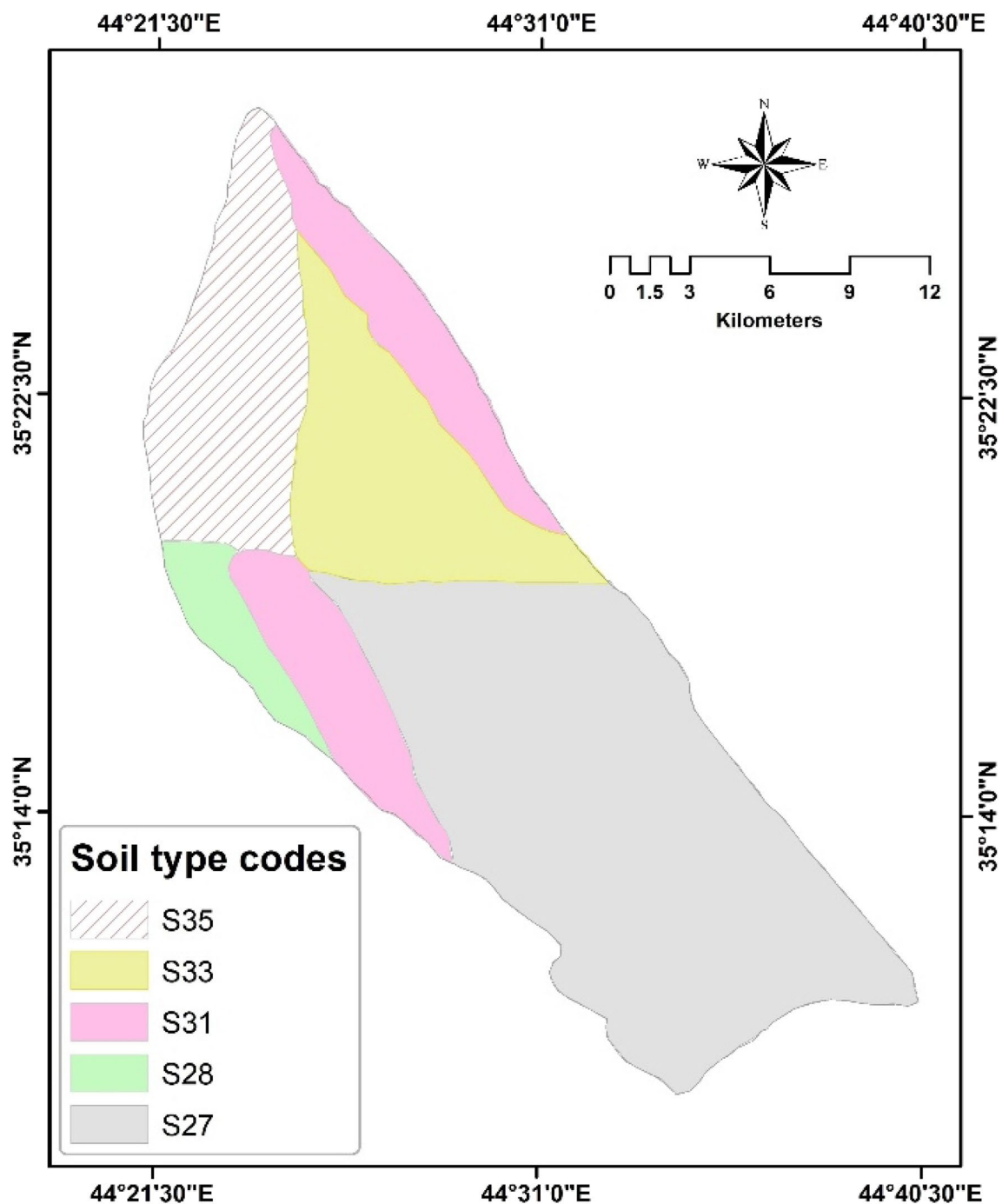


Fig. 7 Soil type in the Lailan basin

classification, regression, and feature selection. Decision trees are simple but effective methods for simulating the relationship between measurements of an object (predictors) and its response variable (target) (Rokach and Maimon 2008). CART uses a binary decision tree to represent the solution. Each node in the tree represents a variable with a single input and has an output variable that is used to make a prediction. CART has many advantages (Aertsen

et al. 2010): (i) the solution is simple to comprehend; (ii) the input data can be of any type (categorical, binary, numeric, etc.); and (iii) the monotonous transformations and different measurement scales of the independent variables do not affect the model results.

LVQ is a prototype-based method that allows choosing how many training instances to retain and learns exactly what those instances should look like (Brownlee 2016). It is

Fig. 8 Groundwater influencing occurrence factors used in the study: **a** slope (%), **b** curvature, **c** aspect, **d** TWI, **e** aquifer transmissivity (m²/day), and **f** aquifer specific storage $\times 10^{-3}$

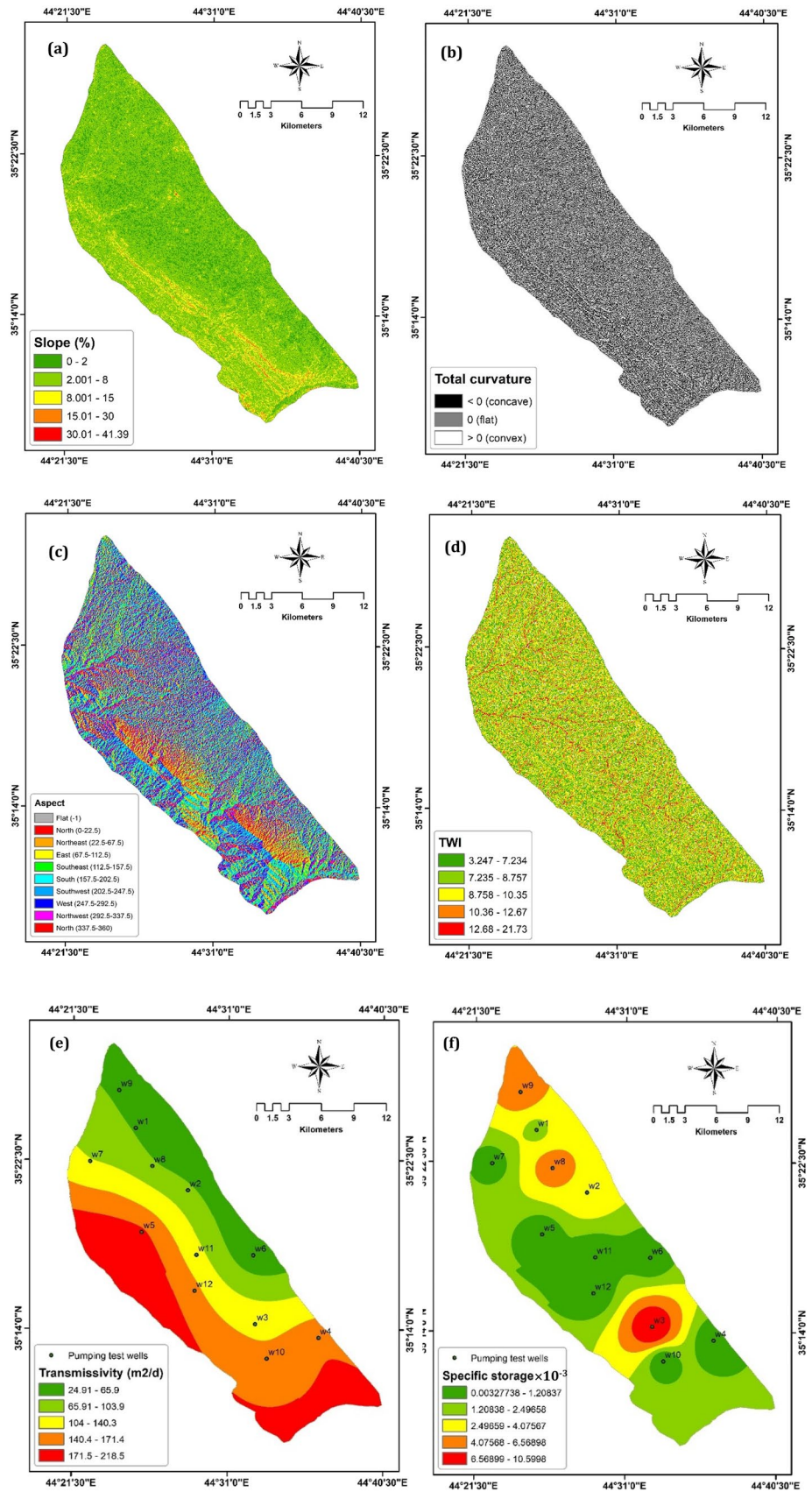


Table 3 Standard methods used for chemical analysis of groundwater samples

Parameter	Method of analysis
EC, pH, TDS, and temperature	Field multi-electrode meter
Ca ²⁺ and Mg ²⁺	Titration with 0.02 N EDTA–Na salt
Cl [−]	Titration with 0.02 N AgNO ₃
SO ₄ ^{2−}	Gravimetric method
NO ₃ [−]	UV-spectrophotometer
Na ⁺ and K ⁺	Flame photometer
HCO ₃ [−]	Titration with 0.02 N H ₂ SO ₄

a form of data compression that represents data vectors by a smaller set of codebook vectors (De Sa and Ballard 1993), which look like training instances but have adopted the values of each attribute based on the learning process. LVQ is a special type of artificial neural network where neuron, weights, and network represent codebook vector, attribute of codebook vector, and codebook vector collection in the structure of LVQ, respectively (Brownlee 2016).

RF is an ensemble-supervised ML technique (Breiman 2001), which involves choosing a set of features randomly and creating a model with a bootstrapped sample of the training data. The building blocks of RF are decision trees. RF increases diversity between classification trees by resampling the data with replacement and randomly modifying data sets over the different induction processes (Peters et al. 2008). During bootstrapping, 1/3 of observations that are not used during tree construction, referred to out-of-bag (OOB) samples, are used as a test set to evaluate misclassification error rate and estimate predictive accuracy (Pourghasemi and Rahmati 2018). The most appealing characteristic of RF is the inherent capability for calculating the variable importance (Al-Abadi 2018). Two hyperparameters should be tuned to get the best results: the total trees that need to be grown and the number of available variables for splitting at each tree node (Kalantar et al. 2019).

KNN is a simple non-parametric algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN is a lazy learning technique, in which the function is only estimated locally and all calculations were postponed until evaluation of function. The main advantages of KNN are: (i) its robustness to noisy data and (ii) simplicity and lack of parametric assumptions (Shmueli et al. 2017).

Software used for applying machine learning algorithms and error evaluation metrics

The R statistical software and related *caret* package (Kuhn 2008) were used in this study for modeling groundwater potential using ML algorithms. R offers a number

of different metrics to distinguish the best performance model. For this study, accuracy and Cohen's κ were used. Accuracy measures how many observations, both positive and negative, were correctly classified. Cohen's κ explains how much better a predictive model is compared to a random model that predicts based on class frequencies. The predictive model (classifier) is said to be slight, fair, moderate, substantial, and almost perfect if Cohen's κ is 0.01–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80, and > 0.80, respectively (Landis and Koch 1977).

Groundwater quality index

GWQI describes the composite spatial and temporal influence of individual parameters on the overall quality of groundwater using a single number (Amiri et al. 2014; Coletti et al. 2010). In general, GWQI computation involves four steps: parameter selection, development of sub-indices, assignment of weights and aggregation of sub-indices to produce an overall index (Akoteyon 2013). Weights reflect the importance of each quality parameter in the final score and can be either subjective or objective. The subjective methods are totally dependent on expert opinions, while in the objective method, entropy information theory is usually used to determine the weights. Peiyue et al. (2010) indicated that the second method is unbiased and more accurate.

The entropy indicates the extent of the instability, disorder, imbalance, and uncertainty of a system (Yufeng and Fengxiang 2009). The degree of a system's disorder can be easily captured by the information entropy, which measures the amount of useful information with the provided data. When the difference of the value among the evaluating objects for the same parameter is high, while the entropy is small, it means that this parameter provides more information, and the weight of this parameter should be set high (Zou et al. 2006). If the difference is smaller and the entropy is higher, the relative weight would be smaller. Therefore, information entropy theory provides an objective method for calculating weights.

In this study, GWQI values were calculated for assessing the suitability of groundwater for human consumption. In the first of three steps, the eight chemical physical parameters (pH, TDS, Ca²⁺, Mg²⁺, Na⁺, SO₄^{2−}, Cl[−], and NO₃[−]) that were considered as indicators of groundwater suitability for drinking water in this study were assigned weights using entropy information theory. Suppose m groundwater samples are taken to evaluate the chemical quality of an area. For each groundwater sample, n chemical constituents are analyzed ($j = 1, 2, \dots, n$). The eigenvalue matrix X can be constructed from real data as:

$$X = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \tag{1}$$

To make the data dimensionless and to facilitate the comparison between chemical constituents, a standardization process must be implemented. To standardize X into the range from 0 to 1, the following equations were utilized.

For the cost type (larger values the better):

$$y_i = \frac{x_i - x_{i(\min)}}{x_{i(\max)} - x_{i(\min)}} \tag{2}$$

For the efficiency type (smaller values the better)

$$y_i = \frac{x_{i(\max)} - x_i}{x_{i(\max)} - x_{i(\min)}} \tag{3}$$

where i is the index or attribute, x_i is the original value of i , and $x_{i(\max)}$ and $x_{i(\min)}$ are the maximum and minimum values of original data. After the standardization process, the standard-grade matrix Y can be obtained as:

$$Y = \begin{bmatrix} y_{11} & y_{21} & \dots & y_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \\ \dots & \dots & \dots & \dots \\ y_{m1} & y_{m2} & \dots & y_{mn} \end{bmatrix} \tag{4}$$

The ratio of the index value of the j index in sample i is calculated as:

$$P_{ij} = \frac{y_{ij}}{\sum_{i=1}^m y_{ij}} \tag{5}$$

The information entropy is expressed by the following formula (Amiri et al. 2014):

$$e_j = -\frac{1}{\ln m} \sum_{i=1}^m P_{ij} \ln P_{ij} \tag{6}$$

The smaller the value of e_j is, the bigger the effect of the j index. The entropy weight is computed using Eq. 7:

$$w_{ij} = \frac{1 - e_j}{\sum_{j=1}^n (1 - e_j)} \tag{7}$$

In the second step of the analysis, the quality rating scale q_j was calculated for each groundwater parameter as:

$$q_j = \frac{C_j}{S_j} \times 100, \tag{8}$$

where C_j is the concentration of each solute (mg/L) or EC ($\mu\text{S/cm}$) in each groundwater sample, and S_j is the WHO

(2017) standard for drinking water for each chemical constituent.

The third step is to calculate GWQI using the linear combination technique:

$$\text{GWQI} = \sum_{j=1}^n w_{ij} q_j. \tag{9}$$

Results and discussion

Feature selection

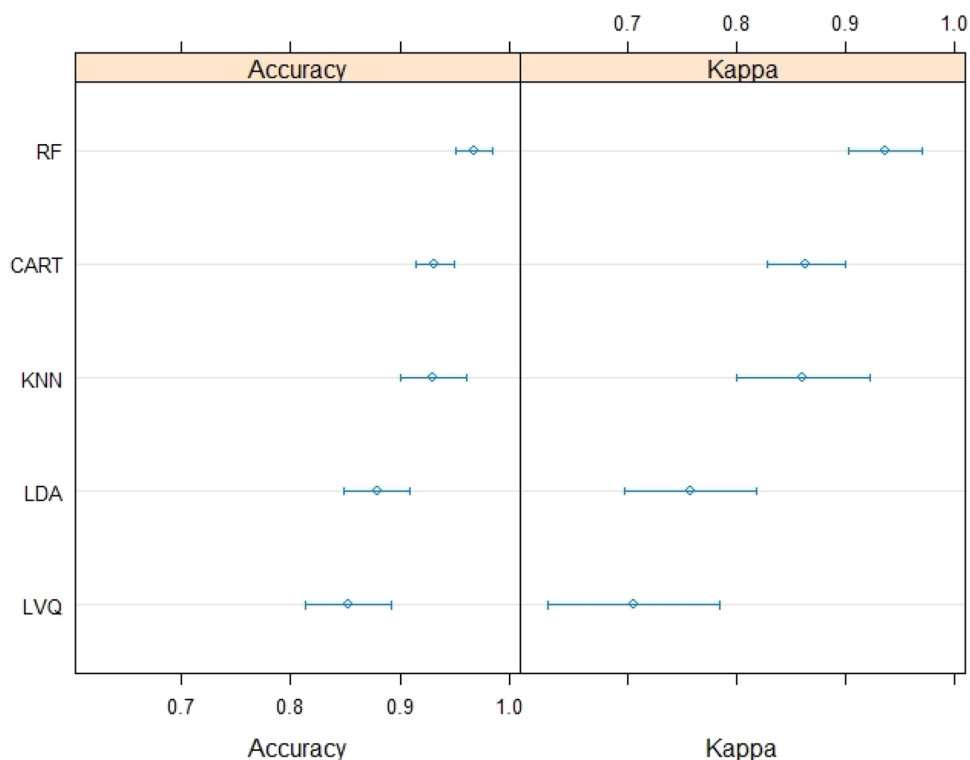
The information gain ratio was implemented in WEKA 3.8 (Witten et al. 2005) and the outputs are presented in Table 4. The highest average merit value was assigned for groundwater depth (0.57), followed by transmissivity (0.526), elevation (0.519), specific storage (0.058), and soil (0.043). The average merit values for TWI, slope, aspect, and curvature all were 0, so these factors were not used in further analysis. The lack of importance of these factors in the modeling of groundwater potential may reflect the fact that they are topographically derived factors that mainly affect the renewable storage of the aquifer. Because the aquifer is relatively deep (average depth about 45 m) and semi-confined, these factors will have little impact on determining the groundwater potential.

Training and validation of machine learning algorithms

The results of training models using the caret package are summarized in Fig. 9 and Table 5. The hyperparameters for each model were estimated automatically using the grid search tuning function in the caret package (Table 6). Depending on the outputs of the training stage, the RF model

Table 4 Feature selection with information gain ratio (tenfold cross-validation)

Attribute	Average merit	Average rank
Depth	0.573 ± 0.013	1.1 ± 0.3
Transmissivity	0.526 ± 0.027	2.5 ± 0.67
Elevation	0.519 ± 0.008	2.4 ± 0.49
Geology	0.191 ± 0.004	4 ± 0
Specific storage	0.085 ± 0.021	5 ± 0
Soil	0.043 ± 0.003	6 ± 0
TWI	0 ± 0	7 ± 0
Slope	0 ± 0	8 ± 0
Aspect	0 ± 0	9 ± 0
Curvature	0 ± 0	10 ± 0

Fig. 9 Comparison of machine learning algorithms (dot plots)**Table 5** Comparing the performance of machine learning models in training and testing stages

Model	Training		Testing	
	Accuracy	Cohen's κ	Accuracy	Cohen's κ
LDA	0.954	0.906	0.879	0.758
CART	0.948	0.895	0.931	0.863
LVQ	0.919	0.837	0.852	0.705
RF	0.996	0.993	0.968	0.936
KNN	0.957	0.907	0.930	0.861

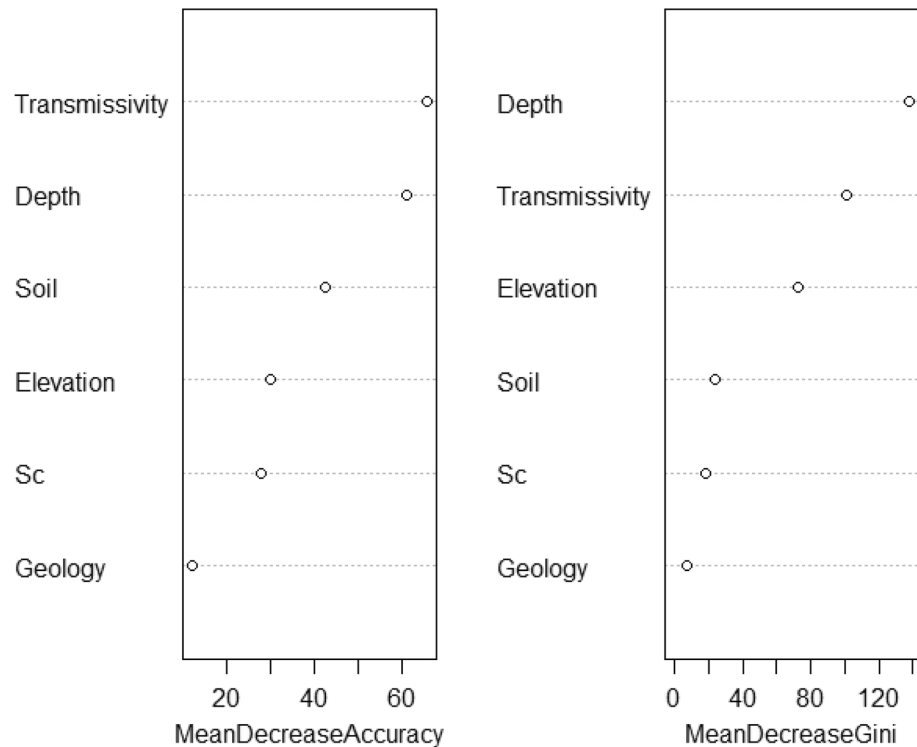
Table 6 Optimal hyperparameters of the used machine learning models

ML model	Hyperparameters	Obtained optimal values
LDA	–	–
CART	Complexity parameter (cp)	0.0308
LVQ	Codebook size	12
	Number of prototypes (k)	1
RF	mtry	7
	ntree	500
KNN	Number of neighbors (k)	7

showed the highest accuracy (0.996), followed by KNN (0.957), LDA (0.954), CART (0.948), and LVQ (0.919). In terms of Cohen's κ , all models had perfect performance (>0.8), and the best model was RF, followed by KNN, LDA, CART, and LVQ. Examination of variable importance using the RF model in terms of a decrease in the Gini index and a mean decrease in accuracy (Fig. 10) showed that the most important GAOFs were the depth to groundwater and aquifer transmissivity, followed by soil type and elevation. The least important variables were specific storage and geology.

After successful training of the models, the testing dataset that was not used in the training step was passed to each algorithm and the results were compared (Table 4). The highest classification accuracy belonged to RF (0.968), followed by CART (0.931), KNN (0.930), LDA (0.879), and LVQ (0.852). In terms of Cohen's κ , the best performing model was RF (0.936), followed by KNN (0.861) and CART (0.863), all of which were classified as perfect models. LDA and LVQ were regarded as substantial models (0.7–0.8). The results revealed that all ML models performed well, but the RF model performed much better than others in both training and testing stages. Previous researchers have shown the superiority of the RF algorithm in studies of groundwater potential (Rahmati et al. 2016; Naghibi et al. 2017; Al-Abadi et al. 2019). RF has many advantages over other ML algorithms, including the capability to handle missing values, resistance to overfitting, and ability to accept a large spectrum of data types (Al-Abadi and Alsamaani 2020).

Fig. 10 Variable importance (RF model); S_c specific storage



The number of hyperparameters is not high and they are easy to understand (Razavi-Termeh et al. 2019 et al. 2019). Therefore, the RF model was selected to map groundwater potential in the study area.

Mapping groundwater potential

The probability values of RF models for both training and testing stages were exported to ArcGIS 10.5 and visualized utilizing three categories of groundwater potential: poor, moderate, and excellent (Fig. 11a). We used a natural-break classification method to categorize probability values into groundwater potential categories because it is the most common scheme used in groundwater potential and spring potential mapping (Al-Abadi et al. 2019; Arabameri et al. 2019; Razavi-Termeh et al. 2019; Chen et al. 2020; Nguyen et al. 2020). The areas occupied by these three GP categories are presented in Table 7. The poor groundwater potential category occupies about half (53%) of the study area in the eastern, northern, and southeastern parts in Quaternary sediments. The excellent category occupies 38% of the basin in the middle and northwest parts. The moderate groundwater potential zone encompasses a small area of the basin as a strip between the poor and excellent zones. In general, the excellent groundwater potential zone (Fig. 11a) coincides with the high values of aquifer transmissivity (Fig. 8e) and occurs in areas with greater groundwater depths (Fig. 5a). This zone mainly occurs in the Mukdadiya and Bai Hassan aquifers, which consist primarily of gravel and sand.

Aquifer transmissivity and groundwater depth were the most influential GAOFs in the study area. Groundwater potential also varies inversely with elevation, which is consistent with topographically driven groundwater flow (Tóth 1962).

Hydrochemistry and groundwater quality index mapping

Results of chemical analyses of the groundwater samples from the study area are listed in Table 8. These indicate that pH values were slightly alkaline, ranging from 7.24 to 8.06 (average 7.63). EC values ranged from 312 to 3459 $\mu\text{S}/\text{cm}$ (average 1567 $\mu\text{S}/\text{cm}$) and TDS, which is linearly related to EC, ranged from 312 to 3459 mg/L (average 1333 mg/L). EC values exceeded 500 $\mu\text{S}/\text{cm}$ for all but three samples (W11, W16, and W18), which are located in the Quaternary deposits along the eastern edge of the study area (Fig. 12). For cations, divalent species were most abundant, with Ca^{2+} concentrations ranged from 41 to 353 mg/L (average 142 mg/L) and Mg^{2+} ranged from 23 to 272 mg/L (average 97 mg/L). Among monovalent cations, Na^+ ranged from 19 to 216 mg/L (average of 82 mg/L) and K^+ ranged from 1 to 6 mg/L (average of 4 mg/L). Four samples (W1, W2, W3, and W13) had Ca^{2+} concentrations > 250 mg/L and 13 samples had Mg^{2+} concentrations > 50 mg/L, whereas only one sample (W2) had Na^+ > 200 mg/L. Sulfate was the most abundant anion, with concentrations ranging from 110 to 1906 mg/L (average 655 mg/L), followed by Cl^- (39–526 mg/L, average

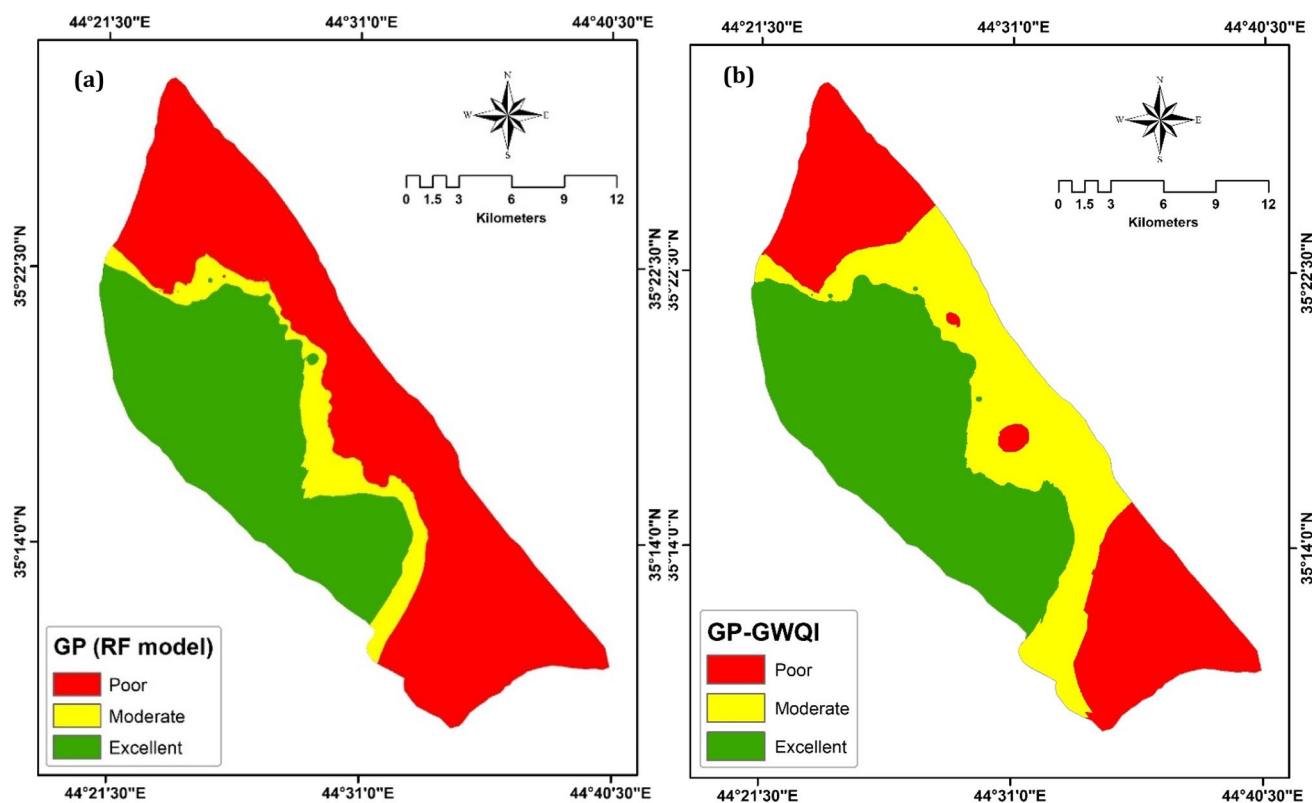


Fig. 11 Groundwater potential (GP) maps: **a** RF model, and **b** GP–GWQI

Table 7 Areas occupied by groundwater potential zones

GP zone	Groundwater potential (R model)		Groundwater potential–GWQI model	
	Area (%)	Area (km ²)	Area (%)	Area (km ²)
Poor	53	224	27	117
Moderate	9	39	30	125
Excellent	37	159	43	180

140 mg/L), HCO_3^- (35–309 mg/L, average 141 mg/L), and NO_3^- (11–69 mg/L, average 28 mg/L). Samples W3, W13, and W21 had Cl^- concentrations > 250 mg/L. For NO_3^- , all groundwater samples were within the WHO (2017) standard (50 mg/L) except W1 and W2, which were taken from wells located in Kirkuk City that may be affected by sewage.

The dominant major-ion hydrochemical type (Back 1966) was Mg–Ca– SO_4 (10 samples), followed by Ca–Mg– SO_4 (6 samples), Mg–Ca– SO_4 –Cl (3 samples), Ca– SO_4 (2 samples), and Ca–Mg– SO_4 –Cl (1 sample). The major-ion composition is consistent with dissolution of anhydrite, gypsum, and calcite from sediments in the Fatha Formation which are exposed in the upgradient (eastern and southeastern) parts of the groundwater flow system. In particular, a plot of milliequivalent concentrations of $(\text{Ca}^{2+} + \text{Mg}^{2+})$ versus

$(\text{SO}_4^{2-} + \text{HCO}_3^-)$ shows a strong linear trend with near-unit slope ($y = 0.920x + 0.405$; $r^2 = 0.98$) (Fig. 13a). A plot of Ca^{2+} versus SO_4^{2-} likewise shows a strong linear trend ($y = 0.452x + 0.940$, $r^2 = 0.92$) (Fig. 13b), while a plot of $(\text{Ca}^{2+} + \text{Mg}^{2+})$ versus HCO_3^- shows a weaker but still positive trend ($y = 5.35x + 2.70$, $r^2 = 0.51$) (Fig. 13c). The positive linear trend of Na^+ versus Cl^- ($y = 0.405 + 25.6x$, $r^2 = 0.63$) (Fig. 13d) and negative linear trend of $([\text{Ca}^{2+} + \text{Mg}^{2+}] - [\text{SO}_4^{2-} + \text{HCO}_3^-])$ versus $(\text{Na}^+ - \text{Cl}^-)$ ($y = -0.381 + 0.819x$, $r^2 = 0.68$) (Fig. 13e) suggest that halite dissolution and cation exchange on clay minerals, respectively, exert lesser influences on groundwater chemistry.

To determine GWQI values, parameter weights were calculated using the entropy approach (Eqs. 1–7). The cost-normalization formula (Eq. 2) was used for pH and the efficient type (Eq. 3) was utilized for the remaining parameters. The highest entropy weight belonged to pH (0.159), followed by Ca^{2+} (0.157), Na^+ (0.140), TDS (0.135), and Mg^{2+} (0.126). The lowest entropy weights were 0.071, 0.100, and 0.112 for Cl^- , NO_3^- , and SO_4^{2-} , respectively. The obtained GWQI values were exported to ArcGIS 10.5 as a point shapefile and interpolated using ordinary kriging to create the GWQI map, Fig. 12. The interpolated GWQI values, which range from 43 to 303, were manually classified into five classes (Jianhua et al.

Table 8 Results of chemical analysis of groundwater samples

Wells	T (°C)	pH	EC (µS/cm)	TDS (mg/L)	Cations (mg/L)				Anions (mg/L)			NO ₃ ⁻	CBE
					Ca ²⁺	Mg ²⁺	Na ⁺	K ⁺	HCO ₃ ⁻	SO ₄ ²⁻	Cl ⁻		
W1	24.0	7.24	2781	2502	323.4	158.9	131.0	03.52	199.8	1408	188.8	68.70	4.0
W2	24.0	7.30	3844	3459	352.7	272.3	215.8	03.90	109.8	1906	195.3	65.20	2.6
W3	24.1	7.46	2793	2513	304.5	126.4	196.9	04.20	219.6	1268	383.3	25.80	8.7
W4	22.4	7.52	2348	2113	206.3	177.3	106.6	02.35	298.2	1049	200.2	37.80	4.5
W5	22.3	7.69	1333	1066	128.2	67.99	48.59	01.80	195.2	537.7	85.20	24.10	8.5
W6	26.3	7.80	955	668	97.41	24.30	41.30	02.40	134.6	298.4	65.29	10.90	8.1
W7	23.8	7.57	1688	1519	224.4	97.80	87.50	02.30	183.0	804.6	106.5	15.70	0.8
W8	24.0	7.73	950	665	72.00	58.31	29.29	01.20	122.0	322.6	57.61	36.20	3.2
W9	24.8	7.61	1041	832	48.00	77.80	78.90	02.40	158.6	399.5	63.89	26.20	1.7
W10	23.4	7.66	717	430	40.19	36.56	38.04	04.31	55.07	145.9	85.97	38.30	3.2
W11	19.8	7.96	478	330	40.19	24.38	23.06	01.84	47.39	112.5	54.96	11.10	4.0
W12	22.5	7.54	902	631	62.40	57.60	48.91	06.50	66.67	312.2	74.49	22.10	2.3
W13	24.1	7.35	3013	2711	297.9	212.4	161.3	05.22	279.1	1190	526.3	22.90	5.6
W14	23.3	7.75	725	580	96.49	23.20	42.71	03.30	34.70	281.2	95.79	25.20	2.6
W15	24.2	7.63	2827	2544	155.0	132.7	190.6	03.98	309.2	846.8	220.0	16.60	3.3
W16	22.5	7.87	462	319	41.49	25.65	30.79	03.66	45.52	115.9	53.38	42.80	9.2
W17	22.9	7.65	819	655	78.32	42.26	30.97	03.59	41.35	358.6	59.65	23.90	5.4
W18	21.5	8.06	404	312	48.00	25.80	18.52	04.08	49.50	110.0	52.36	24.80	8.5
W19	22.0	7.79	548	329	40.71	23.37	26.70	05.43	44.96	147.3	39.46	33.70	3.3
W20	24.1	7.74	605	424	48.60	29.69	32.20	05.80	46.60	201.7	53.21	22.50	0.4
W21	23.6	7.40	2496	2246	192.3	204.2	119.6	04.90	236.8	1215	255.6	15.00	6.8
W22	23.3	7.50	2756	2480	236.1	231.2	109.0	04.08	230.1	1380	153.1	14.50	1.6
Min	19.80	7.24	404	312	40.19	23.20	18.52	1.20	34.70	110.0	39.46	10.9	
Max	26.30	8.06	3844	3459	352.7	272.3	215.8	6.50	309.2	1906	526.3	68.7	
Mean	23.31	7.63	1568	1333	142.5	96.82	82.19	3.67	141.3	655.0	139.6	28.4	
Guideline		6.5–8.5 ¹	–	600 ³	250 ³	50 ³	200 ³	–	–	250 ²	250 ²	50 ¹	

Superscripts: 1 = USEPA (2020) Secondary Maximum Contaminant Level (esthetic), 2 = WHO (2017) guideline, 3 = inferred from WHO (2017)

2011): < 50 (excellent), 50–100 (good), 100–150 (moderate), 150–200 (poor), and > 200 (extremely poor). These classes encompass areas of 11 km² (3%) for excellent, 106 km² (25%) for good, 155 km² (37%) for moderate, 96 km² (23%) for poor, and 54 km² (13%) for extremely poor. The poor–extremely poor classes extend over an area of 150 km² (35%) of the basin and are distributed in the northern (Kirkuk City) and the southern parts of the study area together with some parts of the central region. The excellent–good classes, on the other hand, occupy an area of 118 km² (28%) and are mainly distributed in the middle of the basin. The moderate zone encompasses an area of 155 km² (37%) and is concentrated in the middle of the basin as well. From these results, the Lailan basin is promising in terms of quality. Comparing the GWQI map (Fig. 12) with the groundwater potential map (Fig. 11a) indicates that the groundwater quantity and quality of the basin are not coincident: the most productive parts of the aquifer are in the west whereas the most suitable groundwater quality is in the eastern part of the study area.

Mapping groundwater potential for both quantity and quality aspects

Adjustments are necessary to combine the groundwater potential and GWQI maps into one map. The groundwater potential values are probability values in the range of 0–1, while GWQI values are in the range 30–270 (numerical values), and groundwater quality decreases as GWQI values increase. Therefore, the GWQI raster map was normalized to a 0–1 range using a min–max scaling function (Eq. 2) in the Raster Calculator in ArcGIS 10.5. The resulting raster values were then subtracted from 1 so that the high values are preferred. After that, the two maps (Fig. 11a and modified Fig. 12) were combined using the summation operator to maximize the values of each of the two indicators and the resulting raster was classified into three groundwater potential-GWQI zones (Fig. 11b): poor, moderate, and excellent. The poor groundwater potential zone covers an area of 117 km² (about 28% of the study area), the moderate zone occupies an area of 125 km² (30%), and the excellent

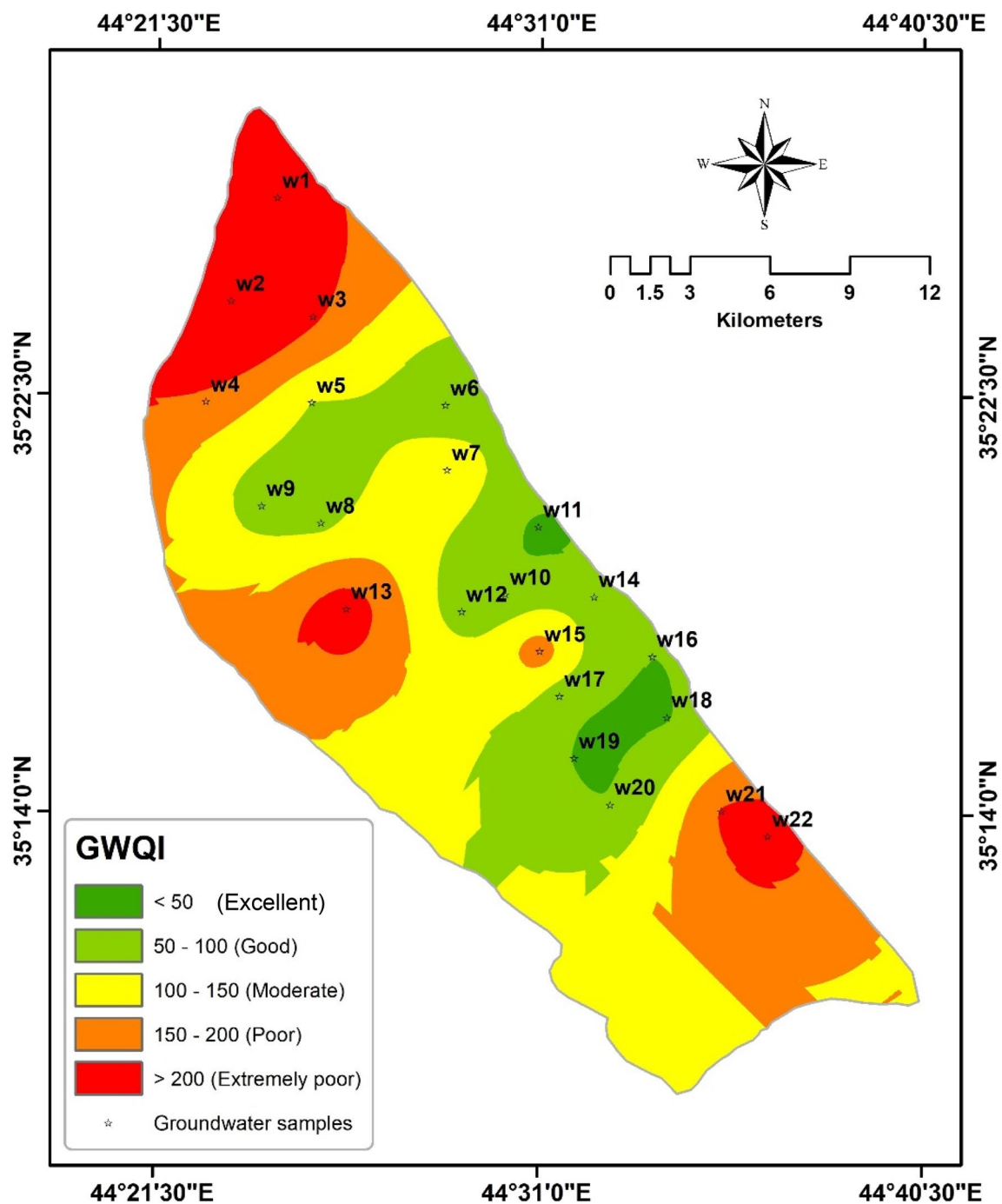


Fig. 12 GWQI of the Lailan basin

zone encompasses an area of 180 km² (43%) (Table 7). Relative to the initial groundwater potential map, which only considers groundwater quantity, the area of the poor zone decreased by 26%, while the moderate zone increased by 21% and the excellent zone increased by 6%. The moderate and excellent zones cover 73% (305 km²) of the study area, which indicates that the aquifer system in the Lailan basin is promising in both its availability and its quality aspects.

The most promising groundwater potential zone occurs in the middle of the basin, whereas groundwater potential is lower in the north (Kirkuk City) and the south.

Comparison with other studies in the region

Results of this study are broadly consistent with published hydrogeologic studies in the region. Water isotopes in

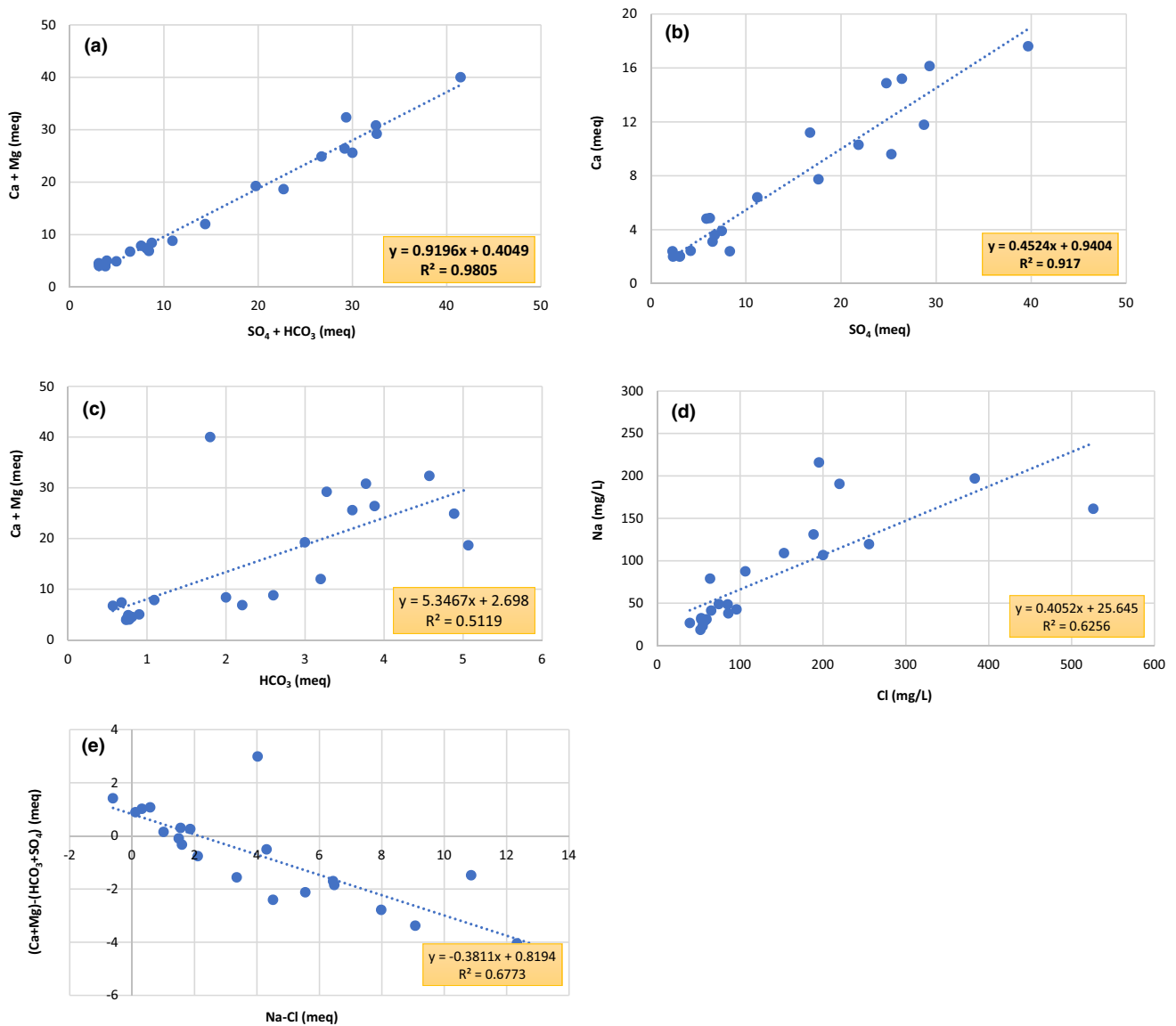


Fig. 13 Bivariate plots of **a** ($Ca^{2+} + Mg^{2+}$) vs. ($SO_4^{2-} + HCO_3^-$), **b** Ca^{2+} vs. SO_4^{2-} , **c** ($Ca^{2+} + Mg^{2+}$) vs. HCO_3^- . Bivariate plots of **d** Na^+ vs. Cl^- , **e** ($[Ca^{2+} + Mg^{2+}] - [SO_4^{2-} + HCO_3^-]$) vs. ($Na^+ - Cl^-$)

shallow groundwater samples fall close to the local meteoric water line (Sahib et al. 2016), which indicates that evaporation during recharge is limited, notwithstanding the semi-arid climate. This may reflect the timing of recharge: rainfall in the Lailan basin tends to occur during the cool season (November to April) and there is a moisture surplus from December through March (Al-Kubaisi and Rasheed 2018). In addition, evaporation may be limited by preferential infiltration through relatively permeable Quaternary slope deposits and faults along the northeastern margin of the basin (Fig. 1). Sahib et al. (2016) found that dissolution of evaporites within the Fatha Formation contributed to salinization and that cation exchange contributed to elevated Na^+ in shallow groundwater. Salinization of shallow

groundwater could also result from upward movement of oilfield brines via faults along anticlinal axes. These faults cut across the Fatha Formation, which is highly fractured and locally karstified (Sahib et al. 2016). Al-Tameemi et al. (2020) assessed groundwater quality in Kirkuk Governorate, including the Lailan basin, using data from 60 wells sampled from 2017 to 2019. Those authors found that 75% of wells exceeded 400 mg/L SO_4^{2-} (the Iraqi drinking-water standard), and they attributed instances of NO_3^- contamination to oxidation of ammonium in sewage from seepage pits. Using the Canadian Water Quality Index as an aggregate indicator based on 15 parameters, Al-Tameemi et al. (2020) concluded that groundwater was marginal to poor for drinking and fair to marginal for irrigation.

Conclusions

Even though groundwater quality is important, it is seldom included in groundwater potential mapping. We developed a technique to resolve this problem and demonstrated the technique for a semi-confined aquifer in northern Iraq. We found that groundwater depth, aquifer transmissivity, elevation, specific storage, and soil type play a major role in controlling the physical groundwater potential in the study area. Using the RF model, which exhibited the best performance of the five ML models considered, we categorized groundwater potential into poor, moderate, and excellent zones. The excellent zone, which occurs in the Mukdadiya and Bai Hassan formations and encompasses 38% of the study area, is closely related to low elevation, high transmissivity, and greater groundwater depths. Calculation of the weights of the chemical constituents for the GWQI using entropy theory assigned the highest weight for pH, followed by Ca^{2+} , Na^+ , TDS, Mg^{2+} , SO_4^{2-} , NO_3^- and Cl^- . The calculated GWQI indicates that the basin is promising for drinking-water quality except for some parts in the north (Kirkuk) and in the south. A comparison of the maps of groundwater potential and GWQI shows the most productive parts of the aquifer are in the west and the most suitable groundwater quality is in the east. Aggregating the results of groundwater potential and GWQI produced a new map for aquifer potential, which could be used to help manage groundwater use by coordinating pumping across the basin in order to avoid depleting the strategic storage and degrading the quality of the groundwater resource. We recommend this approach to take into consideration both quantity and quality aspects in future studies of groundwater potential.

Author contributions AMA-A: Conceptualization, software, writing—original draft preparation, supervision. AF: Supervision, writing—reviewing and editing. AAR: Conceptualization and writing. BP: Reviewing and editing.

Funding The author(s) received no specific funding for this work.

Data availability The data that support the findings of this study are available from the corresponding author, upon reasonable request. Code availability The code is written in R software and we have no right to distribute it.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Aertsen W, Kint V, Van Orshoven J, Özkan K, Muys B (2010) Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecol Model* 221:1119–1130. <https://doi.org/10.1016/j.ecolmodel.2010.01.007>
- Akoteyon IS (2013) Evaluation of groundwater quality using water quality indices in parts of Lagos-Nigeria. *J Environ Geogr* 6:29–36. <https://doi.org/10.2478/v10326-012-0004-2>
- Al-Abadi AM (2018) Mapping flood susceptibility in an arid region of southern Iraq using ensemble machine learning classifiers: a comparative study. *Arab J Geosci* 11:218. <https://doi.org/10.1007/s12517-018-3584-5>
- Al-Abadi AM, Al-Temmeme AA, Al-Ghanimy MA (2016) A GIS-based combining of frequency ratio and index of entropy approaches for mapping groundwater availability zones at Badra–Al Al-Gharbi–Teeb areas, Iraq. *Sustain Water Resour Manag* 2:265–283. <https://doi.org/10.1007/s40899-016-0056-5>
- Al-Abadi AM, Handhal AM, Al-Ginamy MA (2019) Evaluating the Dibdibba aquifer productivity at the Karbala-Najaf Plateau (central Iraq) using GIS-based tree machine learning algorithms. *Nat Resour Res* 29:1989–2009. <https://doi.org/10.1007/s11053-019-09561-x>
- Al-Abadi AM, Shahid S (2015) A comparison between index of entropy and catastrophe theory methods for mapping groundwater potential in an arid region. *Environ Monit Assess* 187:576. <https://doi.org/10.1007/s10661-015-4801-2>
- Al-Abadi AM, Alsamaani JJ (2020) Spatial analysis of groundwater flowing artesian condition using machine learning techniques. *Groundw Sustain Dev* 11:100418. <https://doi.org/10.1016/j.gsd.2020.100418>
- Al-Kubaisi QY, Rasheed AA (2018) Climatic water balance and hydrogeological characteristics of Lailan Basin, southeast Kirkuk—north of Iraq. *Iraqi J Sci* 59:105–118. <https://doi.org/10.24996/ij.s.2018.59.1A.13>
- Al-Tameemi IM, Hasan MB, Al-Mussawy HA, Al-Madhachi AT (2020) Groundwater quality assessment using water quality index technique: a case study of Kirkuk Governorate, Iraq. *IOP Conf Ser Mater Sci Eng* 881:012185. <https://doi.org/10.1088/1757-899X/881/1/012185>
- Amiri V, Rezaei M, Sohrabi N (2014) Groundwater quality assessment using entropy weighted water quality index (EWQI) in Lenjanat, Iran. *Environ Earth Sci* 72:3479–3490. <https://doi.org/10.1007/s12665-014-3255-0>
- Aouragh MH, Essahlaoui A, El Ouali A, El Hmaid A, Kamel S (2017) Groundwater potential of Middle Atlas Plateaus, Morocco, using fuzzy logic approach, GIS and remote sensing. *Geomat Nat Hazards Risk* 8:194–206. <https://doi.org/10.1080/19475705.2016.1181676>
- Arabameri A, Rezaei K, Cerda A, Lombardo L, Rodrigo-Comino J (2019) GIS-based groundwater potential mapping in Shahroud Plain, Iran. A comparison among statistical (bivariate and multivariate), data mining and MCDM approaches. *Sci Total Environ* 658:160–177. <https://doi.org/10.1016/j.scitotenv.2018.12.115>
- Back W (1966) Hydrochemical facies and ground-water flow patterns in the Atlantic Coastal Plain. US Geological Survey Professional Paper 498
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, Boca Raton
- Brownlee J (2016) Master machine learning algorithms. Discover how they work and implement them from scratch. <http://machinelearningmastery.com>. Accessed 12 May 2021

- Bui DT, Tuan TA, Klempe H, Pradhan B, Revhaug I (2016) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides* 13:361–378. <https://doi.org/10.1007/s10346-015-0557-6>
- Chen W, Zhao X, Tsangaratos P, Himan S, Ilia I, Xue W, Wang X, Ahmah BB (2020) Evaluating the usage of tree-based ensemble methods in groundwater spring potential mapping. *J Hydrol* 583:124602. <https://doi.org/10.1016/j.jhydrol.2020.124602>
- Coletti C, Testezlaf R, Ribeiro TAP, Renata T, de Souza G, Pereira D (2010) Water quality index using multivariate factorial analysis. *Rev Bras Eng Agríc Ambient* 14:517–522. <https://doi.org/10.1590/S1415-43662010000500009>
- Das S (2019) Comparison among influencing factor, frequency ratio, and analytical hierarchy process techniques for groundwater potential zonation in Vaitarna Basin, Maharashtra, India. *Groundw Sustain Dev* 8:617–629. <https://doi.org/10.1016/j.gsd.2019.03.003>
- Davoudi Moghaddam D, Rahmati O, Haghizadeh A, Kalantari Z (2020) A modeling comparison of groundwater potential mapping in a mountain bedrock aquifer: QUEST, GARP, and RF Models. *Water* 12:679. <https://doi.org/10.3390/w12030679>
- De Sa VR, Ballard DH (1993) A note on learning vector quantization. In: *Advances in neural information processing systems*, pp 220–227
- Erban LE, Gorelick SM, Zebker HA, Fendorf S (2013) Release of arsenic to deep groundwater in the Mekong Delta, Vietnam, linked to pumping-induced land subsidence. *Proc Natl Acad Sci USA* 110:13751–13756. <https://doi.org/10.1073/pnas.1300503110>
- Fendorf S, Michael HA, van Geen A (2010) Spatial and temporal variations of groundwater arsenic in South and Southeast Asia. *Science* 328:1123–1127. <https://doi.org/10.1126/science.1172974>
- Fetter CW (2018) *Applied hydrogeology*. Waveland Press, Long Grove
- Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugen* 7:179–188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>
- Hantush MS, Jacob CE (1955) Non-steady radial flow in an infinite leaky aquifer. *Trans Am Geophys Union* 36:95–100. <https://doi.org/10.1029/TR036i001p00095>
- Jasem FM, Azeez DR, Hindi NJ (2016) Classification of some soils from Province of Kirkuk and the statement extent of their sensitivity to desertification. *Anbar J Agric Sci* 14:122–132
- Jha MK, Chowdhury A, Chowdhary VM, Peiffer S (2007) Groundwater management and development by integrated remote sensing and geographic information systems: prospects and constraints. *Water Resour Manag* 21:427–467. <https://doi.org/10.1007/s11269-006-9024-4>
- Jianhua W, Peiyue L, Hui Q (2011) Groundwater quality in Jingyuan County, a semi-humid area in northwest China. *J Chem* 8:787–793. <https://doi.org/10.1155/2011/163695>
- Kalantar B, Al-Najjar HAH, Pradhan B et al (2019) Optimized conditioning factors using machine learning techniques for groundwater potential mapping. *Water* 11:1909. <https://doi.org/10.3390/w11091909>
- Khosravi K, Panahi M, Tien Bui D (2018) Spatial prediction of groundwater spring potential mapping based on an adaptive neuro-fuzzy inference system and metaheuristic optimization. *Hydrol Earth Syst Sci* 22:4771–4792. <https://doi.org/10.5194/hess-22-4771-2018>
- Kim J-C, Jung H-S, Lee S (2019) Spatial mapping of the groundwater potential of the Geum River Basin using ensemble models based on remote sensing images. *Remote Sens* 11:2285. <https://doi.org/10.3390/rs11192285>
- King GQ (1991) Geography and GIS technology. *J Geogr* 90:66–72
- Kordestani MD, Naghibi SA, Hashemi H, Ahmadi K, Kalantar B, Pradhan B (2019) Groundwater potential mapping using a novel data-mining ensemble model. *Hydrogeol J* 27:211–224. <https://doi.org/10.1007/s10040-018-1848-5>
- Kuhn M (2008) Building predictive models in R using the caret package. *J Stat Softw* 28:1–26
- Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174
- Lee S, Hyun Y, Lee S, Lee M-J (2020) Groundwater potential mapping using remote sensing and GIS-based machine learning techniques. *Remote Sens* 12:1200. <https://doi.org/10.3390/rs12071200>
- Liu H, Motoda H (1998) Feature extraction, construction and selection: a data mining perspective. Springer, New York. <https://doi.org/10.1007/978-1-4615-5725-8>
- Mehta S, Fryar AE, Brady RM, Morin RH (2000) Modeling regional salinization of the Ogallala aquifer, Southern High Plains, TX, USA. *J Hydrol* 238:44–64
- Muhaimeed AS, Saloom AJ, Saliem KA, Alani KA, Muklef WM (2014) Classification and distribution of Iraqi soils. *Int J Agric Innov Res* 2:997–1002
- Mukherjee A, Sengupta MK, Hossain MA, Ahamed S, Das B, Nayak B, Lodh D, Rahman MM, Chakraborti D (2006) Arsenic contamination in groundwater: a global perspective with emphasis on the Asian scenario. *J Health Popul Nutr* 24(2):142–163
- 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, Ahmadi K, Daneshi A (2017) Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resour Manag* 31:2761–2775. <https://doi.org/10.1007/s11269-017-1660-3>
- Nguyen PT, Ha DH, Avand M, Jaafari A, Nguyen HD, Al-Ansari N, Phong TV, Sharma R, Kumar R, Van Le H, Si Ho L, Prakash I, Thai Pham B (2020) Soft computing ensemble models based on logistic regression for groundwater potential mapping. *Appl Sci* 10:2469. <https://doi.org/10.3390/app10072469>
- Oh H-J, Kim Y-S, Choi J-K, 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>
- Panahi M, Sadhasivam N, Pourghasemi HR, Rezaie F, Saro L (2020) Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR). *J Hydrol* 588:125033. <https://doi.org/10.1016/j.jhydrol.2020.125033>
- Peiyue L, Jianhua W, Hui Q (2010) Groundwater quality assessment based on entropy weighted osculating value method. *Int J Environ Sci* 1:621–630
- Peters J, Verhoest NEC, Samson R, Boeckx P, De Baets B (2008) Wetland vegetation distribution modelling for the identification of constraining environmental variables. *Landscape Ecol* 23:1049–1065. <https://doi.org/10.1007/s10980-008-9261-4>
- Pourghasemi HR, Rahmati O (2018) Prediction of the landslide susceptibility: which algorithm, which precision? *CATENA* 162:177–192. <https://doi.org/10.1016/j.catena.2017.11.022>
- Prinos ST, Wacker MA, Cunningham KJ, Fitterman D V (2014) Origins and delineation of saltwater intrusion in the Biscayne aquifer and changes in the distribution of saltwater in Miami-Dade County, Florida. US Geological Survey Scientific Investigations Report 2014-5025
- Puckett LJ, Tesoriero AJ, Dubrovsky NM (2011) Nitrogen contamination of surficial aquifers—a growing legacy. *Environ Sci Technol* 45:839–844. <https://doi.org/10.1021/es1038358>

- Quinlan JR (2014) C4.5: programs for machine learning. Elsevier Publishing House, Amsterdam. <https://doi.org/10.1016/C2009-0-27846-9>
- Rahmati O, Samani AN, Mahdavi M, Pourghasemi HR, Zeinivand H (2015) Groundwater potential mapping at Kurdistan region of Iran using analytic hierarchy process and GIS. *Arab J Geosci* 8:7059–7071. <https://doi.org/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. <https://doi.org/10.1016/j.catena.2015.10.010>
- Rasheed AA (2019) Evaluation of groundwater in Lailan Basin, south-east Kirkuk. Unpublished Doctoral Thesis, University of Baghdad, Iraq
- Razavi-Termeh SV, Sadeghi-Niaraki A, Choi S-M (2019) Groundwater potential mapping using an integrated ensemble of three bivariate statistical models with random forest and logistic model tree models. *Water* 11:1596. <https://doi.org/10.3390/w11081596>
- Rokach L, Maimon OZ (2008) Data mining with decision trees: theory and applications. World Scientific, Singapore
- Sahib LY, Marandi A, Schüth C (2016) Strontium isotopes as an indicator for groundwater salinity sources in the Kirkuk region, Iraq. *Sci Total Environ* 562:935–945. <https://doi.org/10.1016/j.scitotenv.2016.03.185>
- Schaefer MV, Guo X, Gan Y, Banner SG, Griffin AM, Gorski CA, Wang Y, Fendorf S (2017) Redox controls on arsenic enrichment and release from aquifer sediments in central Yangtze River Basin. *Geochim Cosmochim Acta* 204:104–119. <https://doi.org/10.1016/j.gca.2017.01.035>
- Şen Z (2014) Practical and applied hydrogeology. Elsevier Science, San Diego
- Shahid S, Nath SK, Maksud Kamal ASM (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. <https://doi.org/10.1080/10106040208542246>
- Shmueli G, Bruce PC, Yahav I, Patel NR, Lichtendahl KC Jr (2017) Data mining for business analytics: concepts, techniques, and applications in R. Wiley, New York
- Singhal BBS, Gupta RP (2010) Applied hydrogeology of fractured rocks. Springer, Dordrecht
- Termeh SVR, Khosravi K, Sartaj M et al (2019) Optimization of an adaptive neuro-fuzzy inference system for groundwater potential mapping. *Hydrogeol J* 27:2511–2534. <https://doi.org/10.1007/s10040-019-02017-9>
- Tóth J (1962) A theory of groundwater motion in small drainage basins in central Alberta, Canada. *J Geophys Res* 67:4375–4387
- USEPA (2020) Secondary drinking water standards: guidelines for nuisance chemicals. U.S. Environmental Protection Agency. <https://www.epa.gov/dwstandardsregulations/secondary-drinking-water-standards-guidancenuisance-chemicals>. Accessed 6 June 2020
- WHO (2017) Guidelines for drinking-water quality: fourth edition incorporating the first addendum. World Health Organization, Geneva
- Witten IH, Frank E, Hall MA (2005) Practical machine learning tools and techniques. Elsevier Science, Amsterdam
- Yufeng S, Fengxiang J (2009) Landslide stability analysis based on generalized information entropy. In: 2009 International conference on environmental science and information application technology. IEEE, New York, pp 83–85
- Zhao ZA, Liu H (2011) Spectral feature selection for data mining. CRC Press, Boca Raton
- Zou Z-H, Yi Y, Sun J-N (2006) Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment. *J Environ Sci* 18:1020–1023

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com