Environmental Science and Pollution Research (2023) 30:63991–64005
https://doi.org/10.1007/s11356-023-26961-y

RESEARCH ARTICLE

The potential evaluation of groundwater by integrating rank sum ratio (RSR) and machine learning algorithms in the Qaidam Basin

Zitao Wang[1,2,3] · Jianping Wang[1,2] · Dongmei Yu[1,2,3] · Kai Chen[4]

Received: 7 November 2022 / Accepted: 8 April 2023 / Published online: 15 April 2023

Abstract

Groundwater is a vital resource in arid areas that sustains local industrial development and environmental preservation. Mapping groundwater potential zones and determining high-potential regions are essential for the responsible use of the local groundwater resource. When utilizing machine learning or deep learning algorithms to forecast groundwater potential in arid areas, difficulties such as inaccurate and overfitting predictions might occur due to a shortage of borehole samples. In this study, a database of groundwater conditioning factors with a size of 275,157 × 9 was created in the Qaidam Basin, and 85 known borehole samples were collected. The groundwater potential was evaluated using a combination of rank sum ratio (RSR), projection pursuit regression (PPR) and random forest (RF) algorithms, resulting in four models: PPR, RSR-PPR, RSR-RF, and RF. Results indicated that the groundwater potential was higher in mountainous regions surrounding the Qaidam Basin and decreased progressively towards the central and northwestern regions where most industries and facilities are located. The two primary factors, according to the PPR and RF models, were evapotranspiration (0.246, 0.225) and landform (0.176, 0.294). In terms of their ability to accurately forecast the borehole samples, the four models ranked as follows: RF > RSR-RF > RSR-PPR > PPR. The accuracy of the four models in the low-potential area was 0.73 (PPR), 0.60 (RSR-PPR), 0.87 (RSR-RF), and 0.80 (RF), respectively. However, the RF model showed overfitting due to a lack of samples, especially in high-potential regions, which limits its applicability. The RSR-RF method was applied directly to evaluate the entire factor database, avoiding the risk of overfitting caused by a limited number of training samples. The results demonstrate that the RSR-RF model is effective for classifying groundwater potential types in samples and mapping groundwater potential of the study area. This research presents a novel approach for groundwater potential predictions in areas with insufficient sample sizes, providing a reference for policymakers and researchers.

Keywords Groundwater potential · Qaidam Basin · Rank sum ratio (RSR) · Projection pursuit regression (PPR) · Random forest (RF) · Overfitting

Introduction

Surface water scarcity has become a serious issue in arid areas, which is restricting the growth of the local industry and agriculture (Band et al. 2021; Morsy and Othman 2021). Compared with surface water, groundwater is more abundant and widespread in arid regions. Statistics reveal that northwest China, which occupies 26.73% of the country's land area, contains 1/8 of the groundwater resources (Chen 1986; Wang et al. 2008). As a result, groundwater is used to support sustainable development and provide drinking water for humans and animals (Cui and Shao 2005; Anand et al. 2021). Additionally, groundwater plays a crucial role in maintaining the local ecology and environment by regulating soil water and salt transport, preventing soil degradation, erosion, and plant mortality (Zamani et al. 2022).

Responsible Editor: Marcus Schulz

✉ Jianping Wang
jpwang.team@gmail.com

[1] Key Laboratory of Comprehensive and Highly Efficient Utilization of Salt Lake Resources, Qinghai Institute of Salt Lakes, Chinese Academy of Sciences, Xining 810008, China

[2] Qinghai Provincial Key Laboratory of Geology and Environment of Salt Lakes, Xining 810008, China

[3] University of Chinese Academy of Sciences, Beijing 100049, China

[4] School of Earth and Environment, Anhui University of Science and Technology, Huainan 232001, China

Groundwater potential refers to the ability of soil and rock formations to store and supply water to wells, springs, and other extraction methods, and is an estimate of the amount of water that can be obtained from underground sources (Díaz-Alcaide and Martínez-Santos 2019). The most intuitive and precise method for quantifying groundwater potential is pumping test (Panahi et al. 2020; Wang et al. 2022). The distribution and circulation of groundwater, however, is a complicated system impacted by a broad variety of factors and a highly nonlinear variable of spatial heterogeneity (Wang et al. 2019). Drilling can only acquire groundwater information for specific coordinate locations, making it challenging to visualize how groundwater potential is distributed over a vast arid region. Further, since arid regions are vast and sparsely populated, drilling for groundwater resources is often costly (Ahmed et al. 2021), particularly in developing countries or regions (Zaree et al. 2019), such as the northwest China. In recent years, the mapping of groundwater potential offers an alternative approach for dealing with these challenges. Groundwater potential mapping is the process of creating a map to show the relative likelihood of finding groundwater in a specific area (Shankar and Mohan 2006; Panahi et al. 2020). The map is created by analyzing geological, hydrological, and climatic data to determine the areas with the most favorable conditions for groundwater presence. Many researchers investigating groundwater have utilized a variety of techniques, including geophysical prospecting and interpreting imagery from satellites (Sun et al. 2019; Rateb et al. 2020; Shamsudduha and Taylor 2020) and drones (Jansen 2019). In comparison to drilling, these technologies are far less costly and easier to monitor the groundwater of the whole study area. However, they tend to use physical or mathematical methods to solve the problem rather than being integrated with the local geological and environmental features (Wang et al. 2022).

The use of machine learning (ML) and deep learning (DL) techniques to forecast groundwater potential is growing in popularity (Arabameri et al. 2019, 2021; Tegegne 2022) as artificial intelligence advances. Groundwater data for training was typically collected through hydraulic discharge detected during the drilling process, or by observing multiclass or binary class values of groundwater water-richness in the field. These data were then combined with indicators of geological, environmental, hydrological, and human activity at the drilling location to form the training dataset (Granata et al. 2018). The dataset was subsequently trained with ML or DL algorithms, such as decision trees (DTs) (Lee and Lee 2015; Naghibi et al. 2015), random forest (RF) (Sachdeva and Kumar 2021), support vector machine (Panahi et al. 2020), deep neural networks (Pradhan et al. 2021), and convolutional neural network (Tegegne 2022). Among these techniques, the RF model stands out for its strong generalization ability, fast training speed, and frequent high accuracy (Wang et al. 2020). Additionally, it provides feature importances after training (Breiman 2001), making it an attractive option for combining with other evaluation models. When the models were reliable enough, they were applied to undrilled regions to evaluate the groundwater potential of the whole study area (Pham et al. 2021). Any machine learning model often exhibits inadequate sensitivity or overfitting when the sample size is insufficient; however, there are very few drilling samples obtained in arid areas. For example, many studies only employ 100 or fewer drill samples to train ML or DL models (Chen et al. 2019; Panahi et al. 2020; Arabameri et al. 2021), while the study area to be predicted may be thousands or even tens of thousands of square kilometers. From a geological perspective, drilling work is mainly focused in areas with human activity due to the harsh environment and financial constraints. This leads to the limited representation of the groundwater potential in the entire study region by the obtained samples (Wang et al. 2022). Therefore, in dry regions with few samples, ML and DL algorithms may not always be effective.

Another way to predict groundwater potential is to use evaluation models or rank algorithms (Mandal et al. 2021). After the study area was discretized into many grids or vector points, each factor values of all points were extracted, and the feature database was then formed. The relative values of the groundwater potential can be obtained by calculating weights and overlaying them with the database (Akhtar et al. 2022), or by ranking each item of the database. There are a large number of published studies that describe the application of these evaluation models in the prediction, such as analytic hierarchy process (Arulbalaji et al. 2019; Doke et al. 2021), entropy (Al-Abadi et al. 2017; Zhang et al. 2021), and technique for order preference by similarity to ideal solution (Li et al. 2019). Without the borehole data, the methods can calculate the groundwater potential by only considering the factors. The rank sum ratio (RSR) is a commonly used evaluation model. It differs from other evaluation models in that it incorporates secondary correction during the calculation process, resulting in improved reliability in its practical applications (Wang et al. 2015). RSR has been applied in a range of fields, including medicine (Wu and Shen 2019), social science (Chen et al. 2020), and economics (Pan et al. 2016). However, its use in predicting groundwater potential has not been explored yet.

The Qaidam Basin is an arid endorheic region that is abundant in mineral resources but deficient in water resource (Zhang 1987). In recent years, the geological, ecological, and environmental systems in the region have been severely impacted by the development of basic industries such as nonferrous metal mining, extracting oil and gas, producing chemicals from salt lakes (Xiao et al. 2018). Due to the shortage and the drastic imbalance of surface water over time and space, many industries and facilities for material processing lack an assurance of a steady and dependable water resource (Wang et al. 2008). Additionally, the undeveloped state of the majority of the

Qaidam Basin highlights the need for a spatial division of groundwater potential to guide future drilling activities. However, to the small number of drill samples currently available, it is difficult to use conventional machine learning methods to precisely anticipate the groundwater potential in the region. In this work, the RSR, a correctable evaluation technique, was used to evaluate a database of factors impacting groundwater potential for groundwater potential mapping in the Qaidam Basin. With drilling data, we trained a random forest (RF) model and a projection pursuit regression (PPR) method optimized by a genetic algorithm (GA) to obtain the feature weights. The factor weights were subsequently coupled as a reference value in the RSR to determine the groundwater potential of the Qaidam Basin. The predictions of the PPR and RF were used for comparison as well.

In the following, the "Data and data processing" section describes the study area and database built, the "Methodology" section introduces technical details of the approaches, and results and discussion are provided in the "Results and discussion" section.

Data and data processing

Description of the study area

The Qaidam Basin is located in the arid area of Northwest China (Fig. 1a). It is a part of the northern region of the Qinghai Tibet Plateau (Liu et al. 2012). The longitude and latitude of the Qaidam Basin are 90°16′E to 99°16′E and 35°00′N to 39°20′N, respectively (Fig. 1b), and the whole area is 275,127 km^2. The research area is bordered by the mountains Altun, Qilian, and Kunlun, which are situated in the northwest, northeast, and south, respectively (Han et al. 2021). It has an altitude range of 2429 to 6821 m and a slope range of 0 to 73.19°. The study area has a plateau continental climate and an arid environment, with an annual average temperature of 4.5 °C, annual average rainfall of 18 to 336 mm, and annual evaporation of 1600 to 2630 mm (Wang et al. 2022). The rivers recharge from precipitation and snowfall in the high mountains around the Qaidam basin, flowing to the center of the study area, and forming an endorheic system. There are 37 larger rivers in the Qaidam Basin (Xiao et al. 2018). Total surface water resources are 4.971 billion m^3 per year, and about 85% of those are converted to groundwater (Wang et al. 2008). The study region has a limited population, and animal husbandry is the predominant agricultural activity. The highest yearly demand for water is attributed to industrial activities (Liu et al. 2012), for example, the extraction of various mineral resources, which use an average of 1.2–1.6 billion m^3 water annually. However, the groundwater supplies in the Qaidam basin are distributed unevenly in space (Wang et al. 2008). Groundwater pumping and subsequent drilling activities in the area can be supported by having a solid understanding of the spatial distribution of groundwater potential in the Qaidam basin.

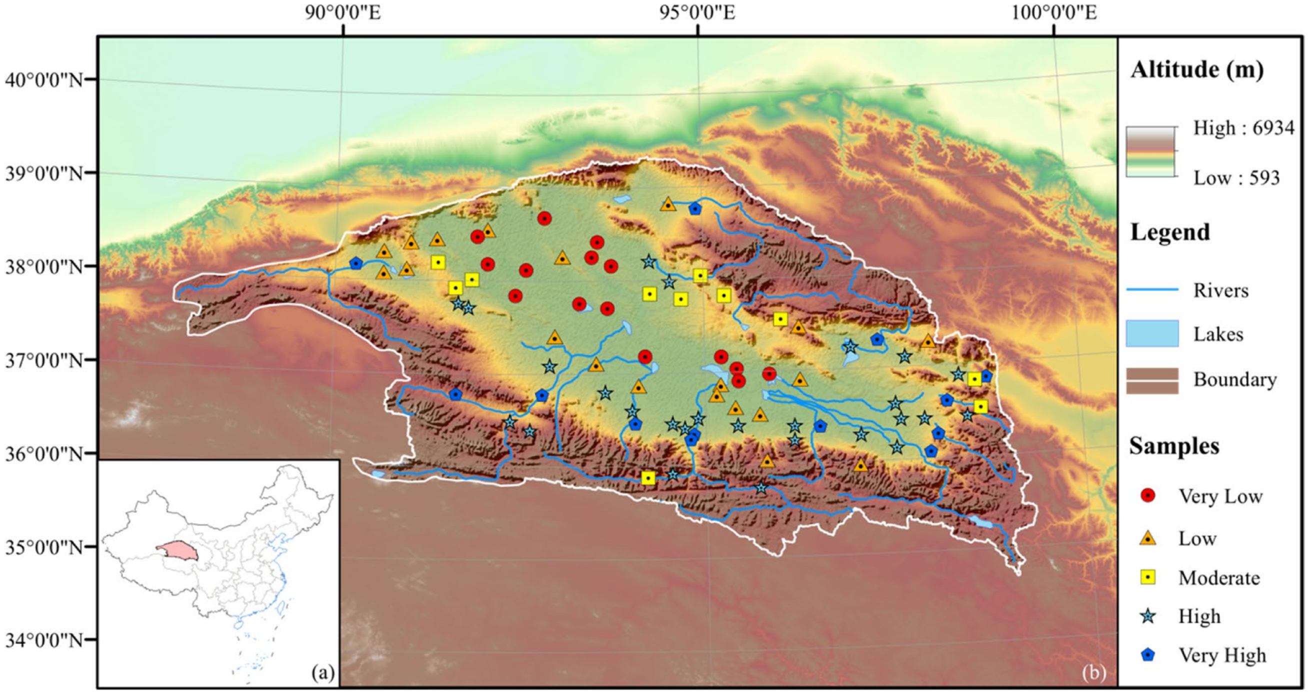

Fig. 1 The study area characteristics and the location of samples

Map of the groundwater borehole inventory

The most precise method of determining groundwater potential is drilling. However, as previously stated, the amount of borehole samples that can be obtained is severely constrained since large-scale drilling in the harsh natural environment is both exceedingly difficult and costly. In this study, a total of 85 sets of groundwater borehole data were collected (Fig. 1) from GeoCloud (http://geoscience.cn), Tibetan Plateau Data Center (TPDC, https://data.tpdc.ac.cn/) and past investigations by our team in the Qaidam basin. The borehole data consisted of their coordinates, subsurface depths, aquifer type and lithology, and hydraulic discharge. According to hydraulic discharge, the boreholes were categorized into five groups: 1, 1–5, 5–10, 10–30, and > 30 t/h, which correspond to very low, low, moderate, high, and very high groundwater potential, respectively. Figure 1 demonstrated that the majority of boreholes are located in the sparsely populated Piedmont plain and the center of the basin, where brine industries are mostly concentrated. In contrast, there is limited borehole data available in the high mountain regions surrounding the study area. The borehole samples collected in the study area showed a gradual change in groundwater potential from very low or low to high or very high, as the samples were taken from north to south and from west to east. However, the distribution of these samples is complex and it is challenging to discern their boundaries by visual inspection alone.

Database of groundwater conditioning factors (DGCF)

The accuracy and applicability of groundwater potential prediction are impacted by the choice of groundwater conditioning factors (Chen et al. 2019; Panahi et al. 2020). To characterize groundwater potential, we must gather various groundwater data to use as input variables for the model. Common types of data used for training may include hydrological, geological, topographic, and climatic data. The specific data required will depend on local and regional characteristics and may reflect recharge, runoff, and discharge conditions in the area, for example, the APLIS model for karst areas (Zaree et al. 2019). In a recent study, we evaluated 17 factors that may potentially impact the groundwater potential of the arid endorheic basins based on a geographical detector model (Wang et al. 2022). The top 8 driving factors (landform, evaporation, soil, geology, river density, precipitation, distance to faults, and slope) that contributed the most to the groundwater potential were identified. In this work, these indicators were used to build the DGCF (Figs. 2 and 3). In addition, considering that desertification is one of the most significant features of the Qaidam basin that has received increasing attention in recent years (Jin et al. 2016; Huang and Jiang 2017; Han et al. 2021), we included fractional vegetation cover (FVC) to the DGCF in order to measure how it affects the groundwater potential of the region. The objective of this study is to understand

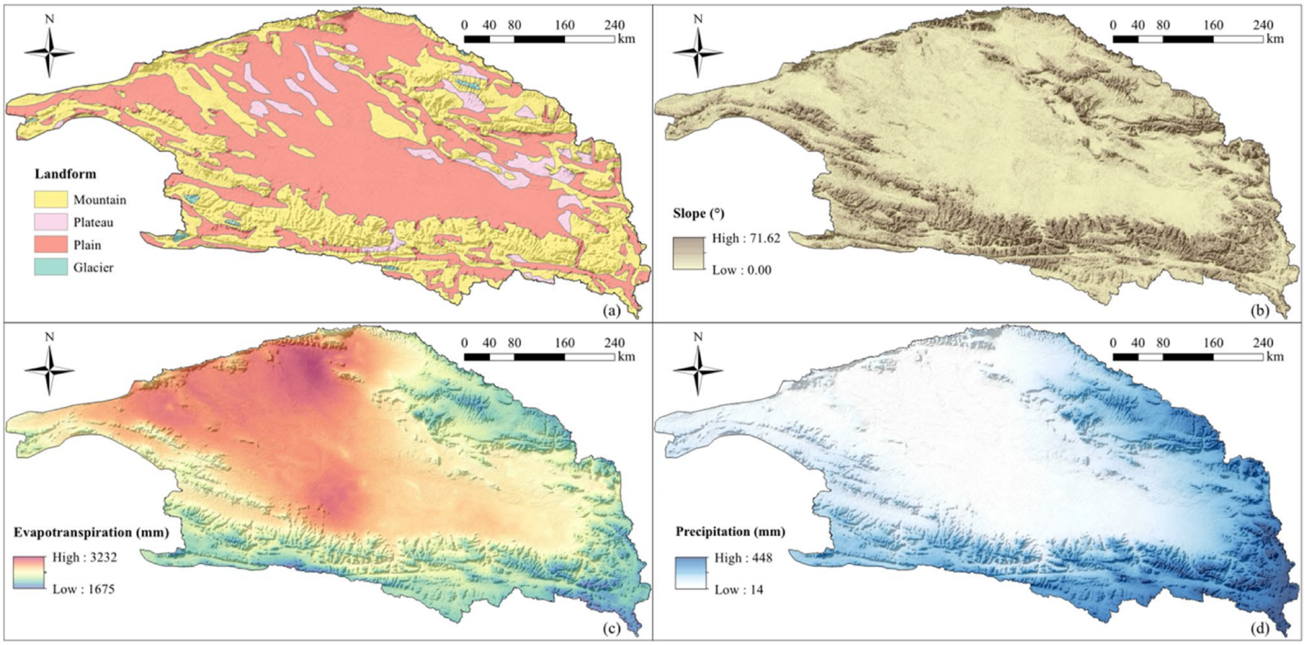

Fig. 2 Groundwater conditioning factors: (**a**) landform, (**b**) slope (°), (**c**) evapotranspiration (mm), (**d**) precipitation (mm)

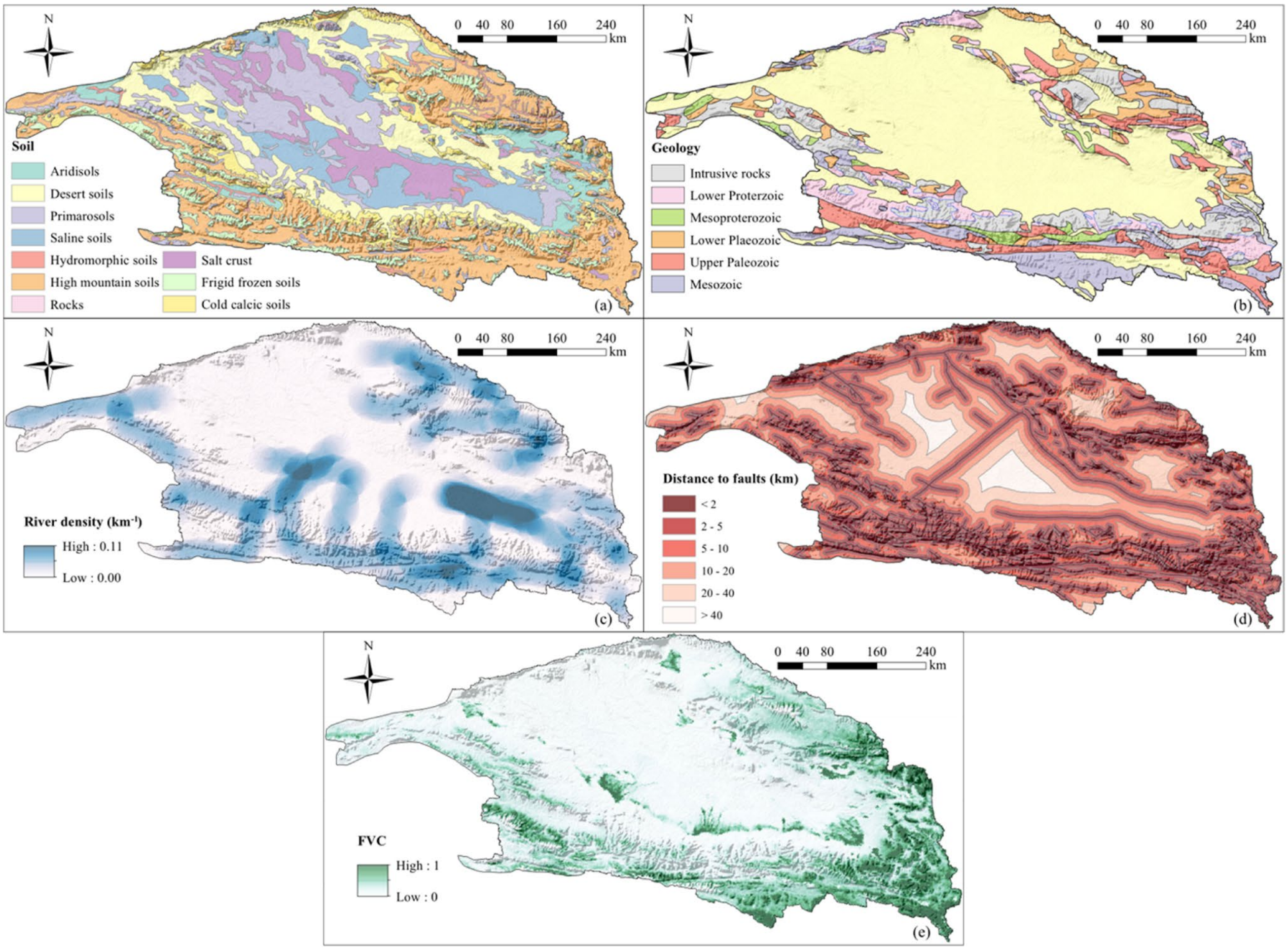

Fig. 3 Groundwater conditioning factors: (**a**) soil, (**b**) geology, (**c**) river density (km^{-1}), (**d**) distance to faults (km), (**e**) FVC

the spatial variation of groundwater potential in the Qaidam Basin. Therefore, indicators that fluctuate over time or seasonally were represented using a yearly average approach (Chen et al. 2019; Panahi et al. 2020; Morsy and Othman 2021).

The hydrological process is controlled by landform and slope, which are significant surface factors for groundwater potential (Razandi et al. 2015). Powered by gravitational potential energy, groundwater and surface water flow from the high mountains to the Piedmont plain and eventually into salt lakes. The landform types in the Qaidam Basin were classified into four categories based on elevation: plain, plateau, mountain, and glacier (Fig. 2a). Slope refers to the ratio of the elevation difference and horizontal distance between a grid and its surroundings, which controls the rate of water flow. The locations with higher slope values have more rapid surface water flow rates, resulting in less infiltration into the ground. The slope for the study area was calculated by the digital elevation model (DEM) with an accuracy of 30 m from the Geospatial Data Cloud (https://www.gscloud.cn); it has continuous values between 0 and 71.62° (Fig. 2b).

The interactions between endorheic basins and external water sources are controlled by evapotranspiration and precipitation (Jia et al. 2011; Jin et al. 2013). In arid endorheic areas, the lack of precipitation, high levels of evaporation, and wide diurnal temperature variations regulate the exchange of energy and information between groundwater, which in turn influences the development and preservation of water resources. Data on rainfall and evapotranspiration were sourced from TPDC and WorldClim 2 (Fick and Hijmans 2017) respectively for this study. The range of average precipitation is 14 to 448 mm, whereas the range of average evapotranspiration is 1675 to 3232 mm, as shown in Fig. 2c and d. The majority of precipitation occurs in mountainous regions, where surface water flows originated. In contrast to rainfall, evapotranspiration is highest in the northwestern and central regions of the study area and gradually decreases toward the southern and eastern regions.

The main interfaces between surface water and groundwater are soil and geology. The pace at which surface water infiltrates groundwater and the overall volume of infiltration vary depending on the type of soil and geology (Shekhar et al. 2015). In the arid regions of Northwest China, the vertical hydrological exchange accounts for roughly 80% of the water balance (Cao et al. 2018). In this study, the categorization information for the soil and lithology factors in the Qaidam Basin were provided by the Resource and Environment Science and Data Center (https://www.resdc.cn). The soil was classified as ten categories: aridisols, desert soils, primarosols, saline soils, hydromorphic soils, high mountain soils, rocks, salt crust, frigid frozen soils, and cold calcic soils (Fig. 3a), and the geology factor was divided into seven categories by geologic time: intrusive rocks, Lower Proterozoic, Mesoproterozoic, Lower Paleozoic, Upper Paleozoic, Mesozoic, and Cenozoic (Fig. 3b).

In dry regions, surface runoff is the primary source of groundwater recharge. When rivers go from the mountains around the study area to the center of the Qaidam basin, they exchange with groundwater near river courses (Golkarian et al. 2018). Therefore, the likelihood that rivers will recharge groundwater increases with river concentration. We used the river density to evaluate the impact of the river indicator on groundwater potential in the study area. The ratio of the total number of main streams and tributaries to the raster area was applied to determine the river density. The river density in the Qaidam Basin ranged continuously from 0 to 0.11 km^{-1} (Fig. 3c). Faults are channels achieving the hydrological exchange. In regions adjacent to water-conducting faults, communication between surface water and groundwater is easier (Ahmad et al. 2021). The distance to faults, which indicates the distance from the nearest fault at any grid in the study area, was computed by buffer tool (Wang et al. 2020) from the Geographic Information Systems (GIS). It was classified into six groups: <2, 2–5, 5–10, 10–20, 20–40, >40 km (Fig. 3d).

The potential of groundwater is impacted by vegetation cover in both positive and negative ways. The vegetation can effectively reduce surface evaporation in arid areas where evaporation is extremely high (Han et al. 2021). On the other hand, plants themselves consume water for transpiration. In this study, FVC, which was derived from normalized difference vegetation index (NDVI), was used as a measure of vegetative cover (Han et al. 2021), that is:

$$\text{FVC} = \frac{NDVI - NDVI_s}{NDVI_v - NDVI_s} \tag{1}$$

where $NDVI_v$ and $NDVI_s$ indicate the values of pure vegetation and bare land, respectively, and the NDVI was extracted from MODIS images (https://glovis.usgs.gov/). Compared to NDVI, the FVC range is constant between 0 and 1 (Fig. 3e).

Among the nine selected indicators affecting the groundwater potential of the Qaidam Basin, slope, evapotranspiration, precipitation, river density, and FVC are continuous variables, whereas landform, soil, geology, and distance to faults belong to discrete variables. The continuous variables are rescaled from 0 to 1 depending on whether a factor has a positive or negative impact on the result using the min–max normalization (Milewski et al. 2020), corresponding equation goes here:

$$\begin{cases} X_{ij}^{*} = \frac{X_{ij} - X_{jmin}}{X_{jmax} - X_{jmin}}, X_j \text{ is positive} \\ X_{ij}^{*} = \frac{X_{jmax} - X_{ij}}{X_{jmax} - X_{jmin}}, X_j \text{ is negative} \end{cases} \tag{2}$$

where X_{ij} and X_{ij}^{*} represent the values of the continuous variables before and after normalization, respectively. If there is a clear quantitative relationship between types of a given variable, such as the distance to faults and landform, the discrete variable was preprocessed similarly to continuous variables; if there is no such relationship, such as soil and geology, they were numbered in decimal form.

Methodology

The DGCF including nine conditioning factors for estimating groundwater potential was created in the previous section. Two sets of point files were generated: one consisting of 85 classified borehole data points and the other comprised of 275,157 vector points, obtained by discretizing the study area with 1 km intervals. The "Extract multi values to point" tool was used to extract the DGCF values to these points based on their respective earth coordinates, resulting in the creation of the sample dataset (size: 85 × 10, containing groundwater potential types) and the database (size: 275,157 × 9). The PPR and RF models were trained on the borehole dataset, respectively, and used to predict the groundwater potential of the Qaidam Basin. The factor weights of the PPR and RF models were then combined with the RSR model for evaluation. As a result, four results of groundwater potential in the study area were acquired: PPR, RSR-PPR, RSR-RF, and RF. The flow chart of the paper was shown in Fig. 4.

Rank sum ratio (RSR)

The rank sum ratio model, which combines nonparametric and traditional statistics, was first proposed by Tian (2002). The RSR approach involves transforming a dataset with *n* rows of samples and *m* columns of features into dimensionless RSR values, which are then used to sort and bin the samples (Wang et al. 2015). The RSR values comprise the

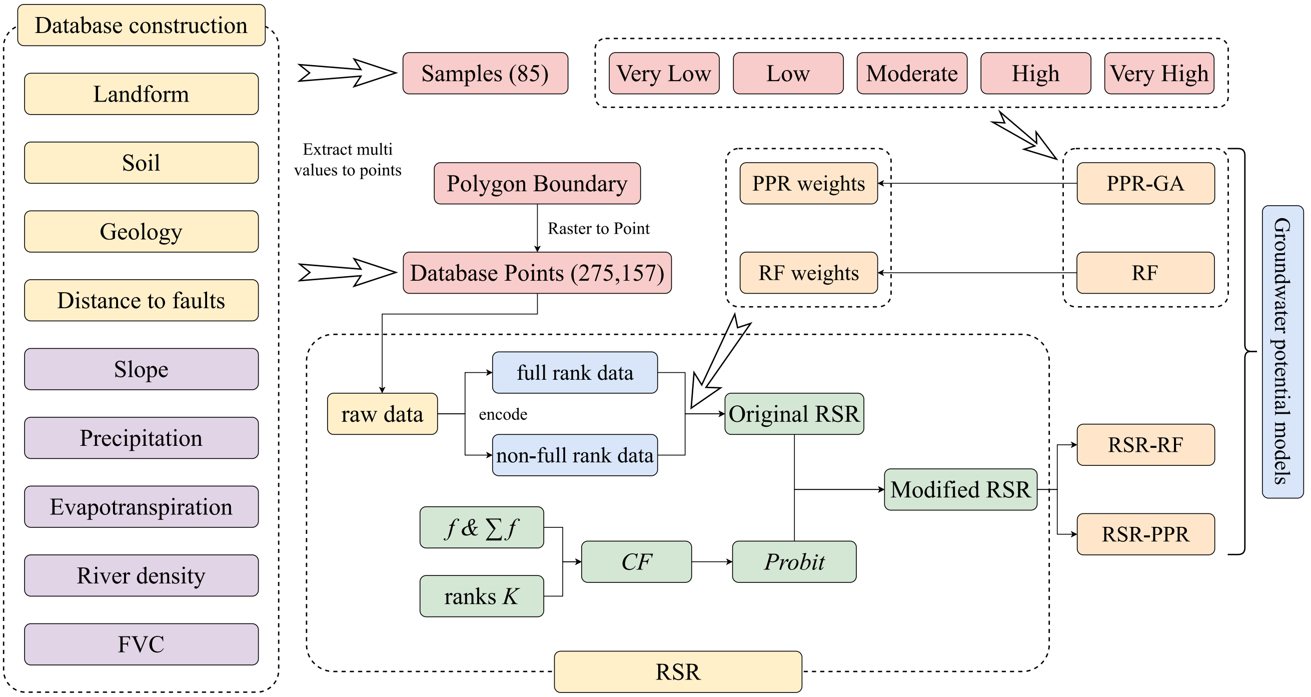

Fig. 4 Flowchart of the methodology

data for all evaluation indicators and represent their combined level, with a higher RSR value indicating a better outcome for decision makers.

The raw data can be encoded as the rank data using two methods: the full rank method, where positive indicators are ranked in descending order and negative indicators are ranked in ascending order, and the non-full rank method, which involves using the equation:

$$R_{ij} = 1 + (n-1) \times X_{ij}^{*} \tag{3}$$

where R_{ij} is the ranked data. Then, the *RSR* values were obtained by:

$$\begin{aligned} RSR_i &= \frac{1}{n}\sum_{j=1}^{m}\omega_j R_{ij} \\ s.t.\ &\sum_{j=1}^{m}\omega_j = 1 \end{aligned} \tag{4}$$

where ω_j represents weights. After the above process was finished, the RSR values are corrected by *Probit* regression. There are four steps to using the *Probit* model (Wang et al. 2015):

Step 1 is to rank the RSR values in order from the smallest to largest, and to list the frequencies *f* with the same RSR values. Step 2 is to determine the average rank $\bar{}R$ at each *f*. Step 3 is to calculate the cumulative frequencies *CF*, that is:

$$\begin{cases} CF_i = \frac{\overline{R}}{n} \times 100\%, i \in (1, n-1), \\ CF_n = \left(1 - \frac{1}{4n}\right) \times 100\% \end{cases} \tag{5}$$

The final step is to convert the *CF* into probability units, *Probit*, which is the standard normal deviation *u* of the *CF* plus five. We can establish a linear regression equation by the modified *RSR* values and *Probit*:

$$RSR_i = a + b \times Probit \tag{6}$$

where the *a* and *b* are undetermined parameters. The least square method was employed to fit the *a* and *b*, and the *RSR* regression values were assessed, replacing the initial *RSR* values. Finally, the modified RSR values were categorized into various classes based on appropriate thresholds for evaluation. In this study, the standardized factor database ($X_{ij}{}^{*}$) was used and the modified RSR values obtained represented the desired groundwater potential values (size: 275,157×1). The mapping of groundwater potential was finished by converting these values using geographic coordinates into two-dimensional pictures.

RSR is sensitive to tiny data gaps since it only evaluates the relative sizes of factors rather than themselves (Yu 2021). Unlike machine learning models, RSR models do not require training sample data. This makes RSR models an ideal choice for evaluating groundwater potential in areas with limited or no sample data.

Projection pursuit regression (PPR)

Projection pursuit regression is a statistical algorithm (Friedman and Stuetzle 1981) which projects the feature data from high-dimensional space to low-dimensional space (1–3 dimension) that reveals the most details about the structure of the dataset (Friedman 1985). This algorithm can be used for various ML tasks, such as classification, clustering, and regression.

Before using the PPR model, the dataset must be uniformed in accordance with Eq. (2) in order to eliminate any negative effects caused by the inconsistent directions and scales of the features. Then, assume that we have a set of directions corresponding to the j features, so the projection process can be explicitly expressed by (Jia et al. 2019):

$$z_i = \sum_{j=1}^{m} a_j X_{ij}^* \tag{7}$$

where z_i is the projection value of i-th sample, and a_j represents the direction of the j-th feature. The size of z_i is $n \times 1$ since the projected groundwater potential is a one-dimensional data. Therefore, we are supposed to excavate the best directions to acquire the projection values that substitute for initial features as far as possible. For the regression problem, the z_i was required to extract more information from the initial features, namely, to get the larger value of standard deviation δ_z:

$$\delta_z = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(z_i - \frac{1}{n}\sum_{i=1}^{n} z_i\right)^2} \tag{8}$$

Meanwhile, the maximum correlation, which quantifies the association between projection values z_i and labels y_i, was calculated by the Pearson's coefficient $P(y, z)$. We define then the fitness function $Q(a)$ (Zhang and Dong 2009):

$$\begin{aligned} &maxQ(a) = \delta_z \times P(y, z) \\ &\quad s.t. \sum_{j=1}^{m} a_j^2 = 1 \end{aligned} \tag{9}$$

In this study, the problem was solved by a genetic algorithm to obtain the best projection direction a. Finally, the groundwater potential values for the entire study area were calculated by substituting a and DGCF into Eq. (7).

Random forest (RF)

Random forest is an ensemble learning algorithm presented by Breiman (2001) that integrates multiple DTs in a Bagging way. From the initial training set of N samples, n samples were randomly sampled with replacement, and they were then trained using a DT (Fig. 5). A total of m DT models were created by repeating this procedure m times, and they were then integrated into a RF model. The RF result was voted on by m DTs. Therefore, RF is considered as an improvement over the DT algorithm (Golkarian et al. 2018).

In machine learning, random forest is one of the most popular and accurate algorithms, especially when used to large datasets (Naghibi et al. 2017; Sajedi-Hosseini et al. 2018; Wang et al. 2020). The unbiased estimates of the generated errors were obtained internally by the RF when building the model (Paul et al. 2018). Thus, RF can handle the input samples containing high-dimensional features without dimensionality reduction. The importance of each feature can also be produced from the RF and utilized as coupling parameters for the RSR model.

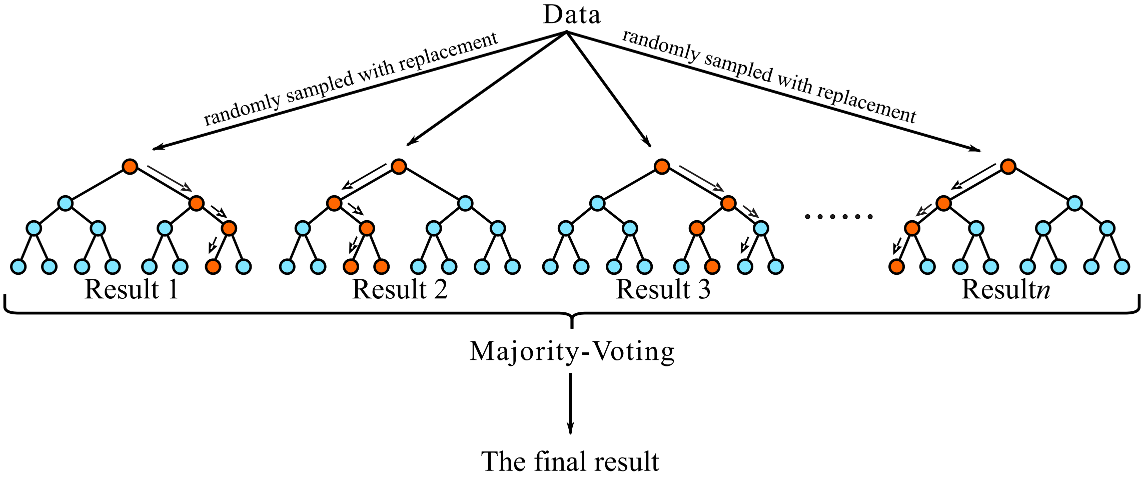

Fig. 5 The structure of a RF algorithm

In this study, using the three approaches mentioned above, we created four groundwater potential prediction models: PPR, RSR-PPR, RSR-RF, and RF. The RSR-PPR and RSR-RF were combinations of the RSR with the calculated weights by PPR and RF respectively, and they were compared with the PPR and RF. All the calculating work on the computer was carried out using Python 3 with its 3-party modules, including Numpy (Harris et al. 2020), Scipy (Virtanen et al. 2020), Sklearn (Pedregosa et al. 2011), NetCDF4, and so on.

Results and discussion

The distribution of the groundwater potential

The Qaidam Basin contains 275,157 sets of groundwater potential values determined using the PPR, RSR-PPR, RSR-RF and RF models. The predicted results of the PPR and RF regression models were real numbers ranging from 1 to 5, which are inconsistent with the magnitude of the RSR results. To facilitate comparison of the individual models, we normalized all predicted values to a range of 0 to 1, and shown in Fig. 6a. It can be seen that the RSR-RF curve is smoother than that of the RSR-PPR, which both reflect a Gaussian distribution. The density curves of PPR and RF, however, show an irregular distribution. Figure 6b displays the weights of the nine factors for the RF and PPR models with 85 samples, where the RF weights represent the feature importance of the DT model outputs, and the PPR weights are the square of the projection directions. In descending order, the RF weights are landform (0.294), evapotranspiration (0.225), river density (0.145), FVC (0.096), slope (0.069), distance to faults (0.055), precipitation (0.052), soil (0.045), and lithology (0.018). The PPR weights are as evapotranspiration (0.246), landform (0.176), precipitation (0.152), river density (0.134), FVC (0.089), lithology (0.083), slope (0.081), soil (0.035), and distance to faults (0.003). Both regression methods reveal that landform and evapotranspiration are the key elements controlling the groundwater potential in the Qaidam Basin, which is consistent with the results of previous research using a geographical detector (Wang et al. 2022). The differences in the weights, as reflected by the two methods, are the distance to faults (Fig. 3d) and geology (Fig. 3b), which may be due to the dense faulting and complicated geology types in the Qaidam Basin, where there are relatively few drill samples. Note that different ML techniques can produce different results when there is a lack of sample references, their accuracy may not always be guaranteed.

The 275,157 points of the PPR, RSR-PPR, RSR-RF, and RF predicted spatial distribution of the groundwater potential in the Qaidam Basin were projected in WGS1984_46N coordinates (Han et al. 2021) and then converted into the rasters, as shown in Fig. 7. The 275,157 results were divided into the five categories: very low, low, moderate, high, and very high using the natural breakpoint method (Das 2017). The four methods generally reveal the same pattern: the southern and northeastern mountain regions of the Qaidam Basin have high groundwater potential, whereas the center and northwest of the basin, characterized by low slope, low rainfall, sparse vegetation and rivers, high evaporation, and uniform lithology, have lower groundwater potential.

The groundwater potential values by the PPR exhibit a strong binary nature (Fig. 7a), i.e., the most regions either extremely high or very low, making it difficult to further differentiate the low potential areas and providing little guidance for local drilling programs. The RF model (Fig. 7d) showed that the central part of the basin generally has low to very low groundwater potential. Conversely, the mountainous regions surrounding the basin have a striped pattern of very high groundwater potential, which is dependent on the presence of samples with very high groundwater potential. This distribution pattern contradicts conventional hydrogeological knowledge. The mapping of the groundwater potential of the entire study area by the RF model indicates that only areas with factor characteristics similar to those of samples with very high groundwater potential will be predicted

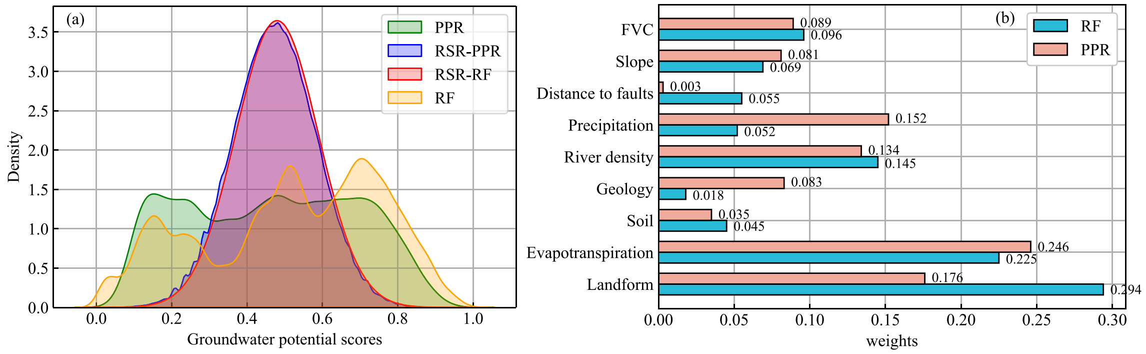

Fig. 6 The density distribution and factor weights of the models

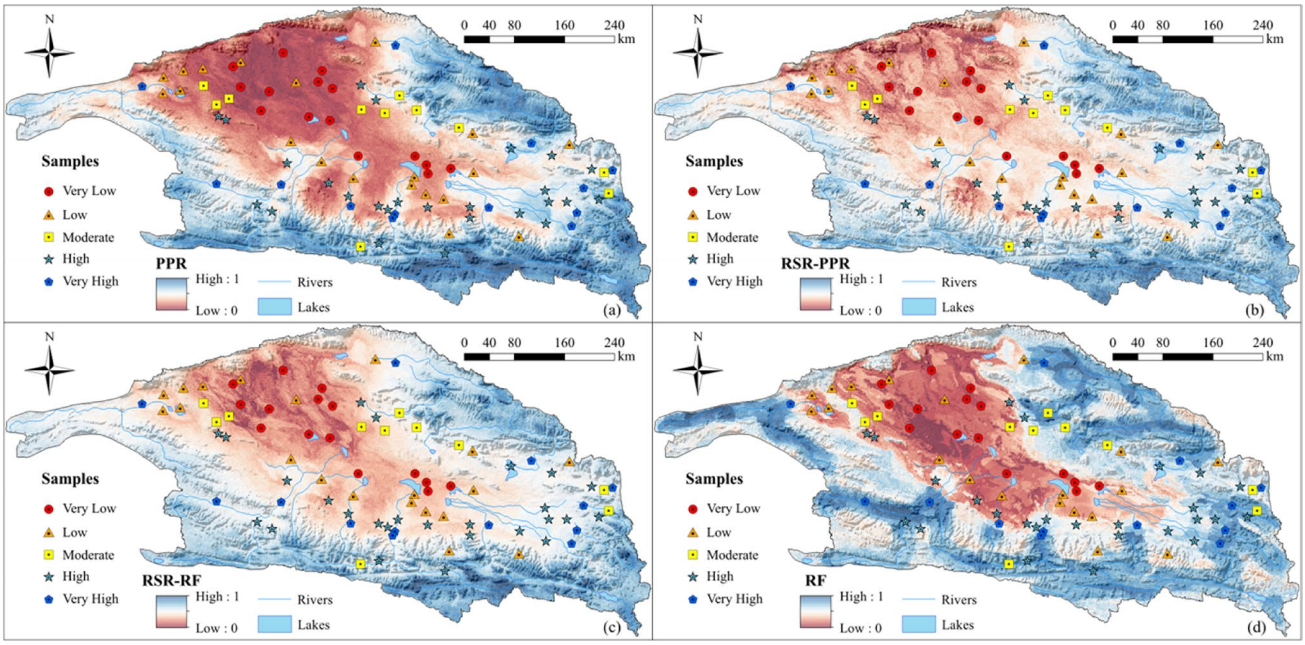

Fig. 7 The groundwater potential results of the four models: (a) PPR, (b) RSR-PPR, (c) RSR-RF, (d) RF

to have very high potential as well. This is due to overfitting of the RF model, which is a result of insufficient examples or samples that cannot cover the various types of each factor.

The RSR-PPR (Fig. 7b) and RSR-RF (Fig. 7c) accurately assessed the spatial distribution of the groundwater potential in the Qaidam Basin. Both methods indicated that the northwestern part of the basin has very low groundwater potential, while the central part primarily has low to moderate groundwater potential. Field observations support this, as the northwestern area is dominated by arid salt flats and lacks rivers, whereas the central part contains multiple salt lakes. These results provide valuable reference for future drilling activities, as the salt lake industries and facilities are located in the central and northwestern regions of the basin.

In addition, high groundwater potential was found to be maintained near rivers, which is consistent with the previous discussion of the river density. In short, the spatial distributions projected by RSR-PPR and RSR-RF outperformed those projected by RF and PPR, providing a more detailed subdivision of local areas.

Model performance

In this study, the performance of the four models was evaluated by computing the groundwater potential at the sample sites (Fig. 8). The predictions of the samples are displayed on the horizontal axis, while the vertical axis depicts the different groundwater grades of the samples (Jin et al. 2001). If the prediction results of a model for the samples show an obvious ladder-like structure in Fig. 8, with each type of sample highly concentrated, it can be concluded that the model accurately predicted the groundwater potential of the 85 borehole samples. However, it should be noted that the 85 borehole samples do not represent the entire study area. The evaluation results of the 85 samples were extracted from the 275,157 sets of results based on the sample coordinates. Therefore, the accuracy of the RSR model in predicting the groundwater potential of the Qaidam Basin demonstrated if it reflects the results of the 85 samples accurately. The RF model demonstrated the highest accuracy for 85 samples. However, examination of Figs. 7d and 8d reveals that the RF model exhibits significant overfitting. When predicting the entire study area, the RF model shows significant distortion in the predicted groundwater potential in the mountainous regions surrounding the Qaidam basin.

The RSR-PPR and RSR-RF models also display the clear step-like distribution. The results of the RSR-PPR are not as compact as those of the RSR-PPR, suggesting that the weights obtained from the RF model are more appropriate than those from the PPR model. The major differences between the factor weights obtained from the RF model and the PPR model, as shown by the comparison in Fig. 6b, are associated with distance to faults, precipitation, geology, and landform. In arid regions, the role of landform and faults in regulating groundwater flow is critical, while precipitation is scarce. This could lead to the PPR model overestimating the importance of precipitation and undervaluing the significance of landform and distance to faults. Consequently, the RSR-PPR model may not accurately predict the groundwater potential of the 85 samples as accurately as the

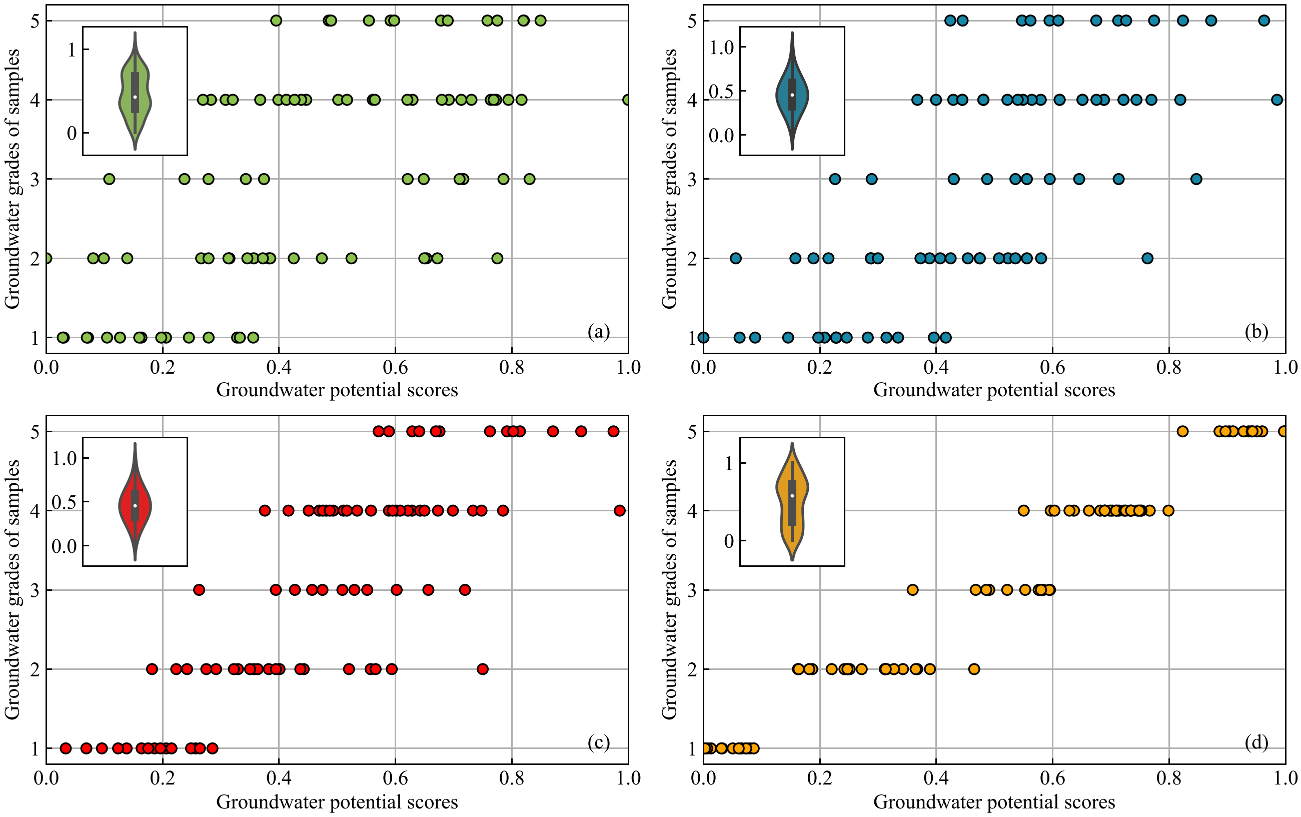

Fig. 8 The scores and distributions of the samples with four models: (a) PPR, (b) RSR-PPR, (c) RSR-RF, (d) RF

RSR-RF model. A violin plot displaying the distribution of the 85 samples is located in the upper left corner of each subplot. The distribution of samples is consistent with that of the study area; specifically, the RSR-PPR and RSR-RF approaches produce Gaussian distributions whereas RF and PPR have irregular distributions. Overall, although the RSR-RF algorithm was not trained on borehole data, it still classified them into different groundwater potential types well.

We analyzed the groundwater potential zones of the DGCF points and 85 samples through histograms to more accurately measure the impact of each model on the prediction (Fig. 9). The intermittent points for PPR were 0.2235, 0.3843, 0.5490, and 0.7020; for RSR-PPR, they were 0.3333, 0.4353, 0.5216, 0.6196; for RSR-RF, they were 0.3451, 0.4392, 0.5216, 0.6157; and for RF, they were 0.2039, 0.3882, 0.5804, and 0.7451. Moreover, we divided the 85 samples into five parts at 0.2 intervals. The red histogram shows the ratio of each groundwater potential class to all DCGFs. The yellow histogram shows the percentage of the groundwater potential types that match the initial classification value, with a higher value indicating better prediction for this class. The blue and green histograms show the proportions of water-rich and water-poor samples in a specific groundwater potential type, respectively.

The yellow histogram shows that for the four models, the ratios of the same samples with the very low potential class were RSR-RF (0.87) > RF (0.80) > PPR (0.73) > RSR-PPR (0.60). But in the very high potential class, the ratios are RF (0.46) > PPR (0.08) and RSR-PPR (0.08) > RSR (0.00). The ratios of the same class samples from very low to very high, using the RSR-RF model as an example, were 0.87, 0.45, 0.36, 0.19, and 0.00. These characteristics suggest that sample effectiveness decreases as the groundwater potential class increases from low to high. Low potential samples, concentrated in the central and northwestern part of the basin, accurately reflect the local features. However, high potential samples are few in the Piedmont basin and underrepresented in high mountain regions, limiting their availability. Unlike the RSR models, the RF model ratio reached 0.462 in the very high groundwater potential class, indicating overfitting due to heavy dependence on the 85 samples. Because the RSR-RF model is largely based on DGCF, it is more accurate than the RF model when there are few samples.

The water-rich samples are primarily in the three groundwater potential classifications of low, moderate, and high. The high potential samples in the Piedmont basin are unrepresentative since the southern and northeastern margins of the basin were predicted to be high potential areas. The ratio

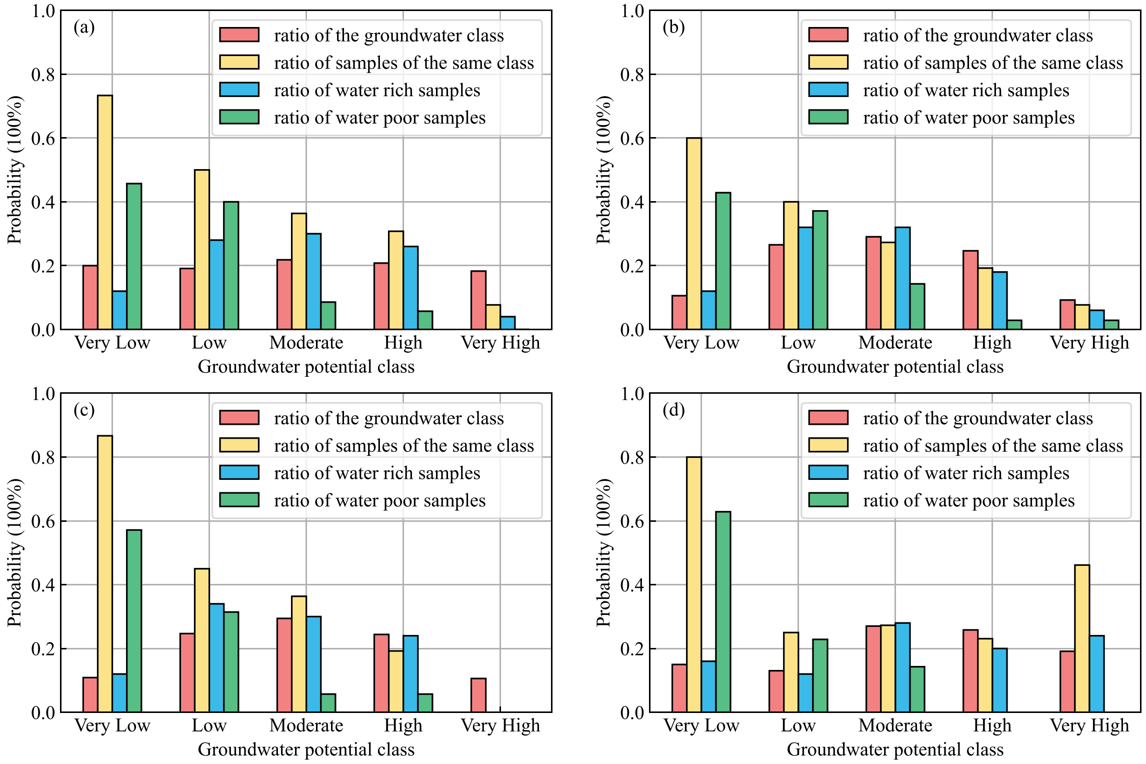

Fig. 9 The ratios of the study area and samples: (a) PPR, (b) RSR-PPR, (c) RSR-RF, (d) RF

of water-poor samples decreases from very low to high class, as shown by the yellow histogram. The ratios of water-poor samples were 0.057, 0.058, 0.057, and 0 for the high and very high potential classes, and 0.857, 0.572, 0.885, and 0.858 for the very low and low potential classes. The RSR-RF model provides the most valid outcomes for the water-poor samples.

Unlike prior studies on the groundwater potential of the Qaidam Basin (Wang et al. 2022), the focus of this study is not solely to find the most accurate prediction method. Rather, we aim to find ways to predict groundwater potential using sample data and reducing the effects of having limited samples. As RSR is an evaluation model that generates 275,157 samples in one run, no sample training is required. By combining the weights generated by the RF and PPR models, we found that the RSR model outperforms pure machine learning models. This is likely because the RSR model evaluates the relative importance of factors affecting groundwater potential, thus being sensitive to small differences in data and effectively projecting the nine factors into a one-dimensional space. The results of RSR-RF were found to be better than RSR-PPR, indicating that despite the overfitting of the RF model, the factor weights generated still have some reference value.

Conclusions

In arid endorheic basins, the use of ML or DL algorithms to forecast groundwater potential can result in incorrect or overfitting findings due to the scarcity of drill samples. In addition, large-scale drilling in these areas is often challenging because of budgetary constraints. This study applied a combination of RSR and ML algorithms to map the groundwater potential of the Qaidam Basin for the first time. Nine factors were selected and transformed into a DGCFs with a size of 275,157×9. A reference dataset of 85 known borehole samples was gathered and divided into five groups based on hydraulic discharge: very low, low, moderate, high, and very high. The samples were trained using the PPR-GA and RF algorithms, and their weights were then integrated with the RSR approach. Four results were obtained: PPR, RSR-PPR, RSR-RF, and RF. The results showed that the groundwater potential is highest in the mountainous regions surrounding the Qaidam Basin and gradually decreases toward the central and northwestern regions, where most industries and facilities are located. Landform (0.176, 0.294) and evapotranspiration (0.246, 0.225) were found to be the two main determinants of groundwater potential, followed by the river density (0.134, 0.145). The four models were ranked in efficacy in predicting the samples: RF > RSR-RF > RSR-PPR > PPR. However, the RF model showed susceptibility to overfitting, particularly in high groundwater potential regions with fewer samples, limiting its applicability. The accuracies of the four models in the low groundwater potential area were 0.73, 0.60, 0.87, and 0.80, respectively, and the ratios of water-poor samples for the low and very low groundwater potential classes were 0.857, 0.572, 0.885, and 0.858. The RSR model did not require training on samples and is effectively evaluated against the DGCF, reducing the risk of overfitting. The combination of the RSR model and the weight value generated by the RF model accurately divides and verifies the drilling samples, ensuring the accuracy of the results. In general, the RSR-RF method proved to be a reliable tool for predicting groundwater potential in the Qaidam Basin. The method offers improved groundwater potential evaluation for the mountainous areas around the basin with limited samples, and more refined groundwater potential zoning for the central and northwestern parts of the basin where the salt lake industry is concentrated. This study exposes the spatial distribution of groundwater potential in the Qaidam Basin, providing a foundation for cost-saving targeted drilling activities. We believe that this method can provide a valuable reference for groundwater potential prediction in regions with few samples.

Author contribution All authors contributed to the study conception and design. Material preparation and data collection were performed by Dongmei Yu, and Kai Chen. Project administration and supervision were carried out by Jianping Wang. Formal analysis and visualization were performed by Zitao Wang. The first draft of the manuscript was written by Zitao Wang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The study was supported cooperatively by the Second Tibetan Plateau Scientific Expedition and Research Program (2019QZKK0805), the key deployment projects of the Chinese academy of sciences (ZDRW-ZS-2020–3), the Qinghai provincial science and technology innovation platform (2020-ZJ-T03), and the National Natural Science Foundation of China (U20A2088).

Data availability The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare no competing interests.

References

Ahmad I, Dar MA, Fenta A et al (2021) Spatial configuration of groundwater potential zones using OLS regression method. J Afr Earth Sci 177:104147. https://doi.org/10.1016/j.jafrearsci.2021.104147

Ahmed A, Alrajhi A, Alquwaizany AS (2021) Identification of groundwater potential recharge zones in flinders ranges, South Australia using remote sensing, GIS, and MIF techniques. Water 13:2571. https://doi.org/10.3390/w13182571

Akhtar J, Sana A, Tauseef SM et al (2022) Evaluating the groundwater potential of Wadi Al-Jizi, Sultanate of Oman, by integrating remote sensing and GIS techniques. Environ Sci Pollut Res 29:72332–72343. https://doi.org/10.1007/s11356-021-17848-x

Al-Abadi AM, Pourghasemi HR, Shahid S, Ghalib HB (2017) Spatial mapping of groundwater potential using entropy weighted linear aggregate novel approach and GIS. Arab J Sci Eng 42:1185–1199. https://doi.org/10.1007/s13369-016-2374-1

Anand B, Karunanidhi D, Subramani T (2021) Promoting artificial recharge to enhance groundwater potential in the lower Bhavani River basin of South India using geospatial techniques. Environ Sci Pollut Res 28:18437–18456. https://doi.org/10.1007/s11356-020-09019-1

Arabameri A, Pal SC, Rezaie F et al (2021) Modeling groundwater potential using novel GIS-based machine-learning ensemble techniques. J Hydrol Reg Stud 36:100848. https://doi.org/10.1016/j.ejrh.2021.100848

Arabameri A, Rezaei K, Cerda A et al (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

Arulbalaji P, Padmalal D, Sreelash K (2019) GIS and AHP techniques based delineation of groundwater potential zones: a case study from Southern Western Ghats, India. Sci Rep 9:2082. https://doi.org/10.1038/s41598-019-38567-x

Band SS, Heggy E, Bateni SM et al (2021) Groundwater level prediction in arid areas using wavelet analysis and Gaussian process regression. Eng Appl Comput Fluid Mech 15:1147–1158. https://doi.org/10.1080/19942060.2021.1944913

Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Cao Y, Nan Z, Cheng G, Zhang L (2018) Hydrological variability in the arid region of Northwest China from 2002 to 2013. Adv Meteorol 2018:e1502472. https://doi.org/10.1155/2018/1502472

Chen M (1986) Regional characteristics and assessment of groundwater resource in China. J Nat Resour 1:18–27. https://doi.org/10.11849/zrzyxb.1986.01.004

Chen W, Panahi M, Khosravi K et al (2019) Spatial prediction of groundwater potentiality using ANFIS ensembled with teaching-learning-based and biogeography-based optimization. J Hydrol 572:435–448. https://doi.org/10.1016/j.jhydrol.2019.03.013

Chen Y, Fan L, Qing F (2020) Affected situation of Chinese national standards based on non-integer rank sum ratio method. In: 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC). pp 1209–1213

Cui Y, Shao J (2005) The role of ground water in arid/semiarid ecosystems, Northwest China. Groundwater 43:471–477. https://doi.org/10.1111/j.1745-6584.2005.0063.x

Das S (2017) Delineation of groundwater potential zone in hard rock terrain in Gangajalghati block, Bankura district, India using remote sensing and GIS techniques. Model Earth Syst Environ 3:1589–1599. https://doi.org/10.1007/s40808-017-0396-7

Díaz-Alcaide S, Martínez-Santos P (2019) Review: advances in groundwater potential mapping. Hydrogeol J 27:2307–2324. https://doi.org/10.1007/s10040-019-02001-3

Doke AB, Zolekar RB, Patel H, Das S (2021) Geospatial mapping of groundwater potential zones using multi-criteria decision-making AHP approach in a hardrock basaltic terrain in India. Ecol Indic 127:107685. https://doi.org/10.1016/j.ecolind.2021.107685

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol 37:4302–4315. https://doi.org/10.1002/joc.5086

Friedman JH (1985) Classification and multiple regression through projection pursuit. Stanf Univ Lab Comput Stat 34. https://doi.org/10.2172/1447844

Friedman JH, Stuetzle W (1981) Projection pursuit regression. J Am Stat Assoc 76:817–823. https://doi.org/10.1080/01621459.1981.10477729

Golkarian A, Naghibi SA, Kalantar B, Pradhan B (2018) Groundwater potential mapping using C5.0, random forest, and multivariate adaptive regression spline models in GIS. Environ Monit Assess 190:149. https://doi.org/10.1007/s10661-018-6507-8

Granata F, Saroli M, de Marinis G, Gargano R (2018) Machine learning models for spring discharge forecasting. Geofluids 2018:8328167. https://doi.org/10.1155/2018/8328167

Han J, Wang J, Chen L et al (2021) Driving factors of desertification in Qaidam Basin, China: an 18-year analysis using the geographic detector model. Ecol Indic 124:107404. https://doi.org/10.1016/j.ecolind.2021.107404

Harris CR, Millman KJ, van der Walt SJ et al (2020) Array programming with NumPy. Nature 585:357–362. https://doi.org/10.1038/s41586-020-2649-2

Huang J, Jiang Y (2017) Influence of climate change on desertification in Qaidam Basin. In: 2017 2nd international conference on civil, transportation and environmental engineering (ICCTE 2017). pp 6–10

Jansen J (2019) Drone based geophysical surveys for groundwater applications. In: 2019 groundwater week

Jia S, Zhu W, Lű A, Yan T (2011) A statistical spatial downscaling algorithm of TRMM precipitation based on NDVI and DEM in the Qaidam Basin of China. Remote Sens Environ 115:3069–3079. https://doi.org/10.1016/j.rse.2011.06.009

Jia Z, Bian J, Wang Y et al (2019) Assessment and validation of groundwater vulnerability to nitrate in porous aquifers based on a DRASTIC method modified by projection pursuit dynamic clustering model. J Contam Hydrol 226:103522. https://doi.org/10.1016/j.jconhyd.2019.103522

Jin J, Wei Y, Ding J (2001) Projection pursuit model for comprehensive evaluation of water quality. Acta Sci Circumstantiae 21:431–434

Jin X, Guo R, Xia W (2013) Distribution of actual evapotranspiration over Qaidam Basin, an arid area in China. Remote Sens 5:6976–6996. https://doi.org/10.3390/rs5126976

Jin X, Liu J, Wang S, Xia W (2016) Vegetation dynamics and their response to groundwater and climate variables in Qaidam Basin, China. Int J Remote Sens 37:710–728

Lee S, Lee C-W (2015) Application of decision-tree model to groundwater productivity-potential mapping. Sustainability 07:13416–13432. https://doi.org/10.3390/su71013416

Li M, Sun H, Singh VP et al (2019) Agricultural water resources management using maximum entropy and entropy-weight-based TOPSIS methods. Entropy 21:364. https://doi.org/10.3390/e21040364

Liu D, Li H, Wang W, Dong Y (2012) Constructivism scenario evolutionary analysis of zero emission regional planning: a case of Qaidam Circular Economy Pilot Area in China. Int J Prod Econ 140:341–356. https://doi.org/10.1016/j.ijpe.2011.04.008

Zaree M, Javadi S, Neshat A (2019) Potential detection of water resources in karst formations using APLIS model and modification with AHP and TOPSIS. J Earth Syst Sci 128:76. https://doi.org/10.1007/s12040-019-1119-4

Mandal T, Saha S, Das J, Sarkar A (2021) Groundwater depletion susceptibility zonation using TOPSIS model in Bhagirathi river basin. Model Earth Syst Environ, India. https://doi.org/10.1007/s40808-021-01176-7

Milewski A, Lezzaik K, Rotz R (2020) Sensitivity analysis of the Groundwater Risk Index in the Middle East and North Africa Region. Environ Process 7:53–71. https://doi.org/10.1007/s40710-019-00421-7

Morsy EA, Othman A (2021) Delineation of shallow groundwater potential zones using integrated hydrogeophysical and topographic analyses, western Saudi Arabia. J King Saud Univ - Sci 33:101559. https://doi.org/10.1016/j.jksus.2021.101559

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

Naghibi SA, Pourghasemi HR, Dixon B (2015) 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

Pan Y, Song W, Xv Y (2016) Research and analysis on market value management in China based on method of rank-sum ratio and principal component analysis. Int J Econ Finance 8:124–124. https://doi.org/10.5539/ijef.v8n11p124

Panahi M, Sadhasivam N, Pourghasemi HR et al (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

Paul A, Mukherjee DP, Das P et al (2018) Improved random forest for classification. IEEE Trans Image Process 27:4012–4024. https://doi.org/10.1109/TIP.2018.2834830

Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Pham BT, Jaafari A, Phong TV et al (2021) Naïve Bayes ensemble models for groundwater potential mapping. Ecol Inform 64:101389

Pradhan AMS, Kim Y-T, Shrestha S et al (2021) Application of deep neural network to capture groundwater potential zone in mountainous terrain, Nepal Himalaya. Environ Sci Pollut Res 28:18501–18517. https://doi.org/10.1007/s11356-020-10646-x

Rateb A, Scanlon BR, Pool DR et al (2020) Comparison of groundwater storage changes from GRACE satellites with monitoring and modeling of major U.S. aquifers. Water Resour Res 56:e2020WR027556. https://doi.org/10.1029/2020WR027556

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

Sachdeva S, Kumar B (2021) Comparison of gradient boosted decision trees and random forest for groundwater potential mapping in Dholpur (Rajasthan), India. Stoch Environ Res Risk Assess 35:287–306. https://doi.org/10.1007/s00477-020-01891-0

Sajedi-Hosseini F, Malekian A, Choubin B et al (2018) A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci Total Environ 644:954–962. https://doi.org/10.1016/j.scitotenv.2018.07.054

Shamsudduha M, Taylor RG (2020) Groundwater storage dynamics in the world's large aquifer systems from GRACE: uncertainty and role of extreme precipitation. Earth Syst Dyn 11:755–774. https://doi.org/10.5194/esd-11-755-2020

Shankar MNR, Mohan G (2006) Assessment of the groundwater potential and quality in Bhatsa and Kalu river basins of Thane district, western Deccan Volcanic Province of India. Environ Geol 49:990–998. https://doi.org/10.1007/s00254-005-0137-5

Shekhar S, Pandey AC, Tirkey AS (2015) A GIS-based DRASTIC model for assessing groundwater vulnerability in hard rock granitic aquifer. Arab J Geosci 8:1385–1401. https://doi.org/10.1007/s12517-014-1285-2

Sun AY, Scanlon BR, Zhang Z et al (2019) Combining physically based modeling and deep learning for fusing GRACE satellite data: can we learn from mismatch? Water Resour Res 55:1179–1195. https://doi.org/10.1029/2018WR023333

Tegegne AM (2022) Applications of convolutional neural network for classification of land cover and groundwater potentiality zones. J Eng 2022:6372089. https://doi.org/10.1155/2022/6372089

Tian F (2002) Rank Sum ratio method and its application. Chin Med J (engl) 4:115–119

Virtanen P, Gommers R, Oliphant TE et al (2020) SciPy 1.0: Fundamental algorithms for scientific computing in python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang Y, Guo H, Li J et al (2008) Investigation and assessment of groundwater resources and their environmental issues in the Qaidam Basin. Geology Press, Beijing

Wang Z, Chen H, Li F (2019) Identifying spatial heterogeneity of groundwater and its response to anthropogenic activities. Environ Sci Pollut Res 26:29435–29448. https://doi.org/10.1007/s11356-019-06121-x

Wang Z, Dang S, Xing Y et al (2015) Applying rank sum ratio (RSR) to the evaluation of feeding practices behaviors, and its associations with infant health risk in rural Lhasa, Tibet. Int J Environ Res Public Health 12:15173–15181. https://doi.org/10.3390/ijerph121214976

Wang Z, Liu Q, Liu Y (2020) Mapping landslide susceptibility using machine learning algorithms and GIS: a case study in Shexian County, Anhui Province, China. Symmetry-Basel 12:1954. https://doi.org/10.3390/sym12121954

Wang Z, Wang J, Han J (2022) Spatial prediction of groundwater potential and driving factor analysis based on deep learning and geographical detector in an arid endorheic basin. Ecol Indic 142:109256. https://doi.org/10.1016/j.ecolind.2022.109256

Wu X, Shen S (2019) Comprehensive evaluation of medical service efficiency in TCM hospitals based on data envelopment method and rank sum ratio method. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp 2486–2492

Xiao Y, Shao J, Frape SK et al (2018) Groundwater origin, flow regime and geochemical evolution in arid endorheic watersheds: a case study from the Qaidam Basin, northwestern China. Hydrol Earth Syst Sci 22:4381–4400. https://doi.org/10.5194/hess-22-4381-2018

Yu B (2021) Computer dynamic forecast model with adaptability through the method of rank-sum ratio. In: Journal of Physics: Conference Series. IOP Publishing, p 012017

Zamani MG, Moridi A, Yazdi J (2022) Groundwater management in arid and semi-arid regions. Arab J Geosci 15:362. https://doi.org/10.1007/s12517-022-09546-w

Zhang C, Dong S (2009) A new water quality assessment model based on projection pursuit technique. J Environ Sci 21:S154–S157. https://doi.org/10.1016/S1001-0742(09)60062-0

Zhang P (1987) Salt Lakes of the Qaidam Basin. Science Press

Zhang Y, Jia R, Wu J et al (2021) Evaluation of groundwater using an integrated approach of entropy weight and stochastic simulation: a case study in east region of Beijing. Int J Environ Res Public Health 18:7703. https://doi.org/10.3390/ijerph18147703

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