



Spatial predictions of groundwater potential using automated machine learning (AutoML): a comparative study of feature selection and training sample size in Qinghai Province, China

Zitao Wang^{1,2,3} · Jianping Wang^{1,2} · Mengling Li^{1,2,3}

Received: 16 August 2023 / Accepted: 22 November 2023 / Published online: 1 December 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Predicting groundwater potential is crucial for identifying the spatial distribution of groundwater in a region. It serves as an essential guide for the development, utilization, and protection of groundwater resources. Previous studies have primarily emphasized finding the most accurate prediction model for groundwater potential while giving less attention to the selection of training features and sample sizes. This study aims to predict groundwater potential within Qinghai Province using automated machine learning technology and assess the influence of sample sizes and feature selection on prediction accuracy. Sixteen groundwater conditioning factors were categorized into categorical and numerical variables. Four feature selection modes were utilized as input in training the model. The results indicated that, except for correlations between evaporation and landforms (−0.8) and precipitation and normalized difference vegetation index (0.8), the Pearson correlation coefficients among the remaining sixteen factors were ≤ 0.5 or ≥ -0.5 . The models XGB-ALL, RF-Entropy, ET-CRITIC, and XGB-PCA yielded accuracy scores of 0.783, 0.685, 0.745, and 0.703, and area under curve (AUC) of 0.819, 0.724, 0.779, and 0.747, respectively. If enough samples are available with the tree model, an increased number of features can improve prediction accuracy. The principal component analysis method showed difficulty in reducing the dimensionality of the input space, while the Entropy method proved efficient. The accuracy and AUC value of the prediction model improved with an increasing number of samples. Training with 8 features and 200 data points achieved an accuracy of 0.745, sufficient to evaluate regional groundwater potential. As for training with 600 samples, the model's performance accuracy rose to 0.9, enabling precise groundwater potential prediction. The outputs of this research can provide decision-makers in groundwater resource management in Qinghai Province with crucial theoretical and practical support. The lessons learned can have future applications in similar situations.

Keywords Groundwater potential · Automated machine learning (AutoML) · Qinghai Province · Feature selection · Training sample size

Responsible Editor: Marcus Schulz

Highlights

- This study predicted the groundwater potential in Qinghai Province using automated machine learning technology. The influence of sample sizes and feature selection on prediction accuracy was assessed.
- The study collected 16 factors that affect groundwater potential and adopted four feature selection methods, including the “ALL” method which selected all 16 factors, and the remaining methods each included only 8 factors for training.
- The Entropy method proved efficient in reducing the dimensionality of the input space. as the number of samples increases, the accuracy and AUC value of the groundwater potential prediction model risen. Training with 8 factors and 200 samples results in 0.75 accuracy, sufficient to evaluate regional groundwater potential.

Extended author information available on the last page of the article

Introduction

Groundwater is an essential resource for supporting human existence and global socio-economic progress (Anand et al. 2021). It serves as a critical source of freshwater for drinking, farming, and industrial activities. Because of its wide availability and affordability, it acts as the major source of water for many suburban and rural settlements in developing countries (Wang et al. 2023a). Nevertheless, groundwater depletion presents an increasing worldwide concern since it is not a renewable source. Providing an uninterrupted groundwater supply to fulfill the upsurging demands of the community continues to be a daunting issue (Sun et al. 2019). Hence, forecasting groundwater potential is imperative for productive management, utilization, and preservation of this scarce asset.

Groundwater potential refers to either the availability of groundwater in a particular area or the volume of water that can be withdrawn from an aquifer without affecting the surrounding environment (Jhariya et al. 2021). The formation and change process are complex and determined by various environmental factors (Tegegne 2022). In the past, forecasting groundwater potential relied on traditional methods, such as drilling and geophysical surveys, which were both costly and time-consuming, and limited to specific areas. However, with the advancement in artificial intelligence algorithms and computer performance (Reichstein et al. 2019), researchers are increasingly turning to machine learning models to forecast groundwater potential. By collecting and quantifying environmental factors through remote sensing (Shamsudduha and Taylor 2020) and geographic information systems (Bera et al. 2021), researchers can employ machine learning algorithms to predict the spatial distribution of groundwater potential or examine the intricate relationships between various environmental factors and groundwater potential. Commonly, logistic regression (Rizeei et al. 2019), decision trees (Lee and Lee 2015), random forests (Golkarian et al. 2018), gradient boosting machines (Sachdeva and Kumar 2021), support vector machines (SVM) (Naghibi et al. 2017), deep neural networks (Wang et al. 2022), and convolutional neural networks (Panahi et al. 2020) were used to forecast groundwater potential. Additionally, hybrid models that combine machine learning with other evaluation methods like analytic hierarchy process (Ahmad et al. 2023), Technique for Order of Preference by Similarity to Ideal Solution (Mahnaz Zaree et al. 2019), rank sum ratio (Wang et al. 2023a), multi-criteria decision-making (Farhat et al. 2023), and genetic algorithms were also employed to improve predictions.

Machine learning algorithms provide efficient methods for evaluating regional groundwater potential. However, previous studies focused mainly on comparing and evaluating the performance of various models (Wang et al. 2023b). To predict groundwater potential in a particular area, researchers collected existing data, typically obtained from boreholes that pump groundwater (Arabameri et al. 2021), and used them as the training dataset for the model. Researchers often chose two or more separate models for training and compared their accuracy to identify the most precise one (Razandi et al. 2015). However, collecting sufficient training data is challenging due to economic and natural constraints that limit sample collection. Several studies have imposed limits on the sample size, which range from a few tens to two hundred (Bera et al. 2021), and may be inadequate for machine learning models to offer precise predictions. Furthermore, to achieve more credible model performance, multiple studies split the dataset into a training set of 70% of the samples and a 30% test set (Pham et al. 2021), resulting in a further one-third reduction in the original sample size. Additionally, the prediction

of groundwater potential can be affected by the choice of training features, as some factors may be more sensitive to regional groundwater potential than others (Thanh et al. 2022). Researchers often aim to use as many factors as possible in training to increase accuracy, but this approach can lead to multi-collinearity problems, particularly in commonly linear models, such as linear regression, logistic regression, SVM, and naive Bayes. Furthermore, selecting too many factors can cause the curse of dimensionality (Pedregosa et al. 2011), which limits the accuracy of the model, particularly due to the small number of borehole samples. Therefore, in the context of predicting groundwater potential, the selection and number of factors, as well as the sample size, can have a greater impact on the accuracy of the model than the quality of the model or selection of parameters.

In this study, we aim to enhance the accuracy of groundwater potential prediction by exploring the impact of feature selection methods and sample size on prediction performance. A total of 16 factors that affect groundwater potential in Qinghai Province, China were collected. Unlike previous studies that directly train all factors, this research incorporates four distinct methods: the “ALL” method, which utilizes all 16 factors affecting groundwater potential, and three techniques that reduce the dimension of factors — Principal Component Analysis (PCA) (Sun et al. 2021), Entropy (Naghibi et al. 2015b), and Criteria Importance Through InterCriteria Correlation (CRITIC) (Rostamzadeh et al. 2018). Each of these three methods includes only eight factors for training, providing a comprehensive comparison of their performance. Furthermore, an extensive analysis on the influence of sample size on prediction performance is conducted, a factor often overlooked in previous studies. As our research was not focused on finding the ideal model or parameter, the automated machine learning (AutoML) method (Wang et al. 2021) was used to identify and select the optimal model and parameter set. The predictive model was applied to determine groundwater potential across the entire Qinghai Province. By using trained models, the study was able to accurately assess the groundwater potential in Qinghai Province. The results of this study will be helpful for guiding future research on groundwater potential prediction by assisting with the selection of appropriate sample sizes and features.

Data and data processing

Description of the study area

Qinghai Province is situated in the northeast region of the Qinghai-Tibet Plateau in Northwest China, spanning from 89°35'E to 103°04'E for longitude and 31°9'N to 39°19'N for latitude (Liu et al. 2013). The province covers

an area of approximately 720,000 km² and has an altitude ranging from 2000 to 5000 m (Han et al. 2016). Its geological conditions chiefly arose from the Himalayan orogeny and the Qinghai-Tibet Plateau uplift. The region is endowed with a plethora of rock strata from various geological periods from the Paleozoic to the Late Cenozoic, along with some scattered volcanic and intrusive rocks. The study area experiences a plateau continental climate and its annual mean temperature oscillates between -5.1 and 9.0 °C, while the annual mean rainfall varies substantially across different locations, ranging from 18 to 780 mm (Han et al. 2021). The average annual evaporation is around 1012–3335 mm (Fig. 1). The region is the origin of numerous rivers and lakes, including those that flow outward into the ocean, such as the Yangtze River and the Yellow River (Cao et al. 2020), as well as inland rivers that pour into the Qaidam Basin (Wang et al. 2022). Qinghai Province is crucial to China's water conservancy and mining industries, as it is abundant in a broad range of non-ferrous metals, coal, oil, gas, and salt minerals (Kong et al. 2017). Despite the abundance of local water resources, the province still faces a notable disparity between water supply and demand. Therefore, it is necessary to provide an accurate assessment of groundwater resources in Qinghai Province to achieve sustainable development and efficient management, given the importance of water for industrial and agricultural activities in the region.

Sample dataset

The reliability of prediction results and the accuracy of machine learning algorithms used for groundwater potential prediction depend on the number of training samples (Panahi et al. 2020). However, the limited economic and natural conditions in the geosciences field pose several challenges in collecting sufficient samples for the prediction process. Therefore, many researchers use a limited number of samples to estimate the groundwater potential of various study areas. In this study, we investigated the influence of the number of samples on the prediction results by constructing different datasets involving various sample sizes. 800 samples were extracted randomly from the hydrogeological map of GeoCloud (<http://geoscience.cn>) (Wang et al. 2023b), and we categorized them into five groups based on the number of samples available: 50, 100, 200, 600, and 800 samples (Fig. 1). We also used 1500 validation samples to evaluate the model's performance. Based on groundwater abundance at the source location indicated on the hydrogeological map, the samples were divided into two categories: type 1 for enriched groundwater and type 0 for groundwater scarcity. Moreover, the entire study area was discretized, which had an accuracy of $1\text{ km} \times 1\text{ km}$, into 699,016 points to draw a spatial distribution of groundwater potential in the study area after completing the model training.

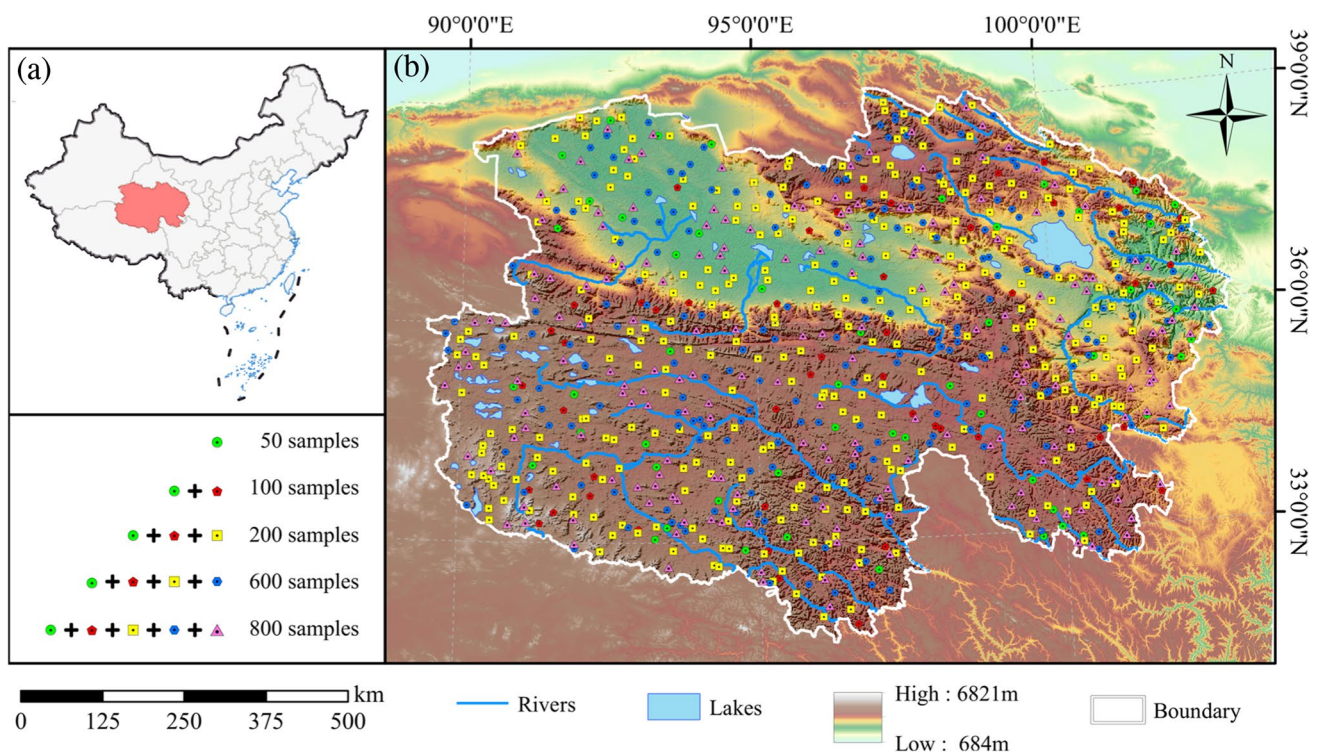


Fig. 1 a Location of Qinghai Province in China. b Topographic features and location of the samples in the study area

Database of conditioning factors

The generation, enrichment, migration, and discharge of regional groundwater are directly or indirectly controlled by groundwater conditioning factors (Mahnaz Zaree et al. 2019). These factors serve as features in the machine learning model used for predicting groundwater potential. Therefore, before proceeding with model training and predicting groundwater potential, it is necessary to collect the eigenvalues of all conditioning features that could potentially impact groundwater in the entire study area. Furthermore, the spatial resolution of the factor features must be no lower than that of the discrete units to ensure their accuracy. In our study, the resolution was less than $1\text{ km} \times 1\text{ km}$. We analyzed 16 characteristics in the Qinghai region that affect groundwater (Díaz-Alcaide and Martínez-Santos 2019), including precipitation, evapotranspiration, normalized difference vegetation index (NDVI), landcover, slope, topographic wetness index (TWI), slope aspect, curvature, distance to rivers, distance to roads, fault density, residential density, landform,

vegetation types, soil, and lithology (Figs. 2, 3, 4, and 5). We discussed each characteristic separately and its impact on predicting groundwater potential.

Precipitation and evaporation are the two critical sources and sinks of groundwater. During rainfall, some water infiltrates into the ground, fully replenishing groundwater (Jin et al. 2013). Precipitation also enhances the replenishment of surface runoff, indirectly influencing groundwater replenishment (Jia et al. 2011). On the other hand, evaporation is the process through which water is removed from soil and groundwater reservoirs, thereby reducing groundwater storage. For this study, we acquired precipitation data for Qinghai Province from WorldClim (Fick and Hijmans 2017), with values ranging between 18 and 780 mm, and evaporation data from <https://data.cma.cn>, with values ranging between 1012 and 3335 mm. The spatial variability of precipitation and evaporation across the study area was significant, as shown in Fig. 2 (a) and (b). The Qaidam Basin, situated in the northwest of Qinghai Province, experienced minimal precipitation, paired with high evaporation. Conversely, the eastern part of the study area received relatively abundant rainfall.

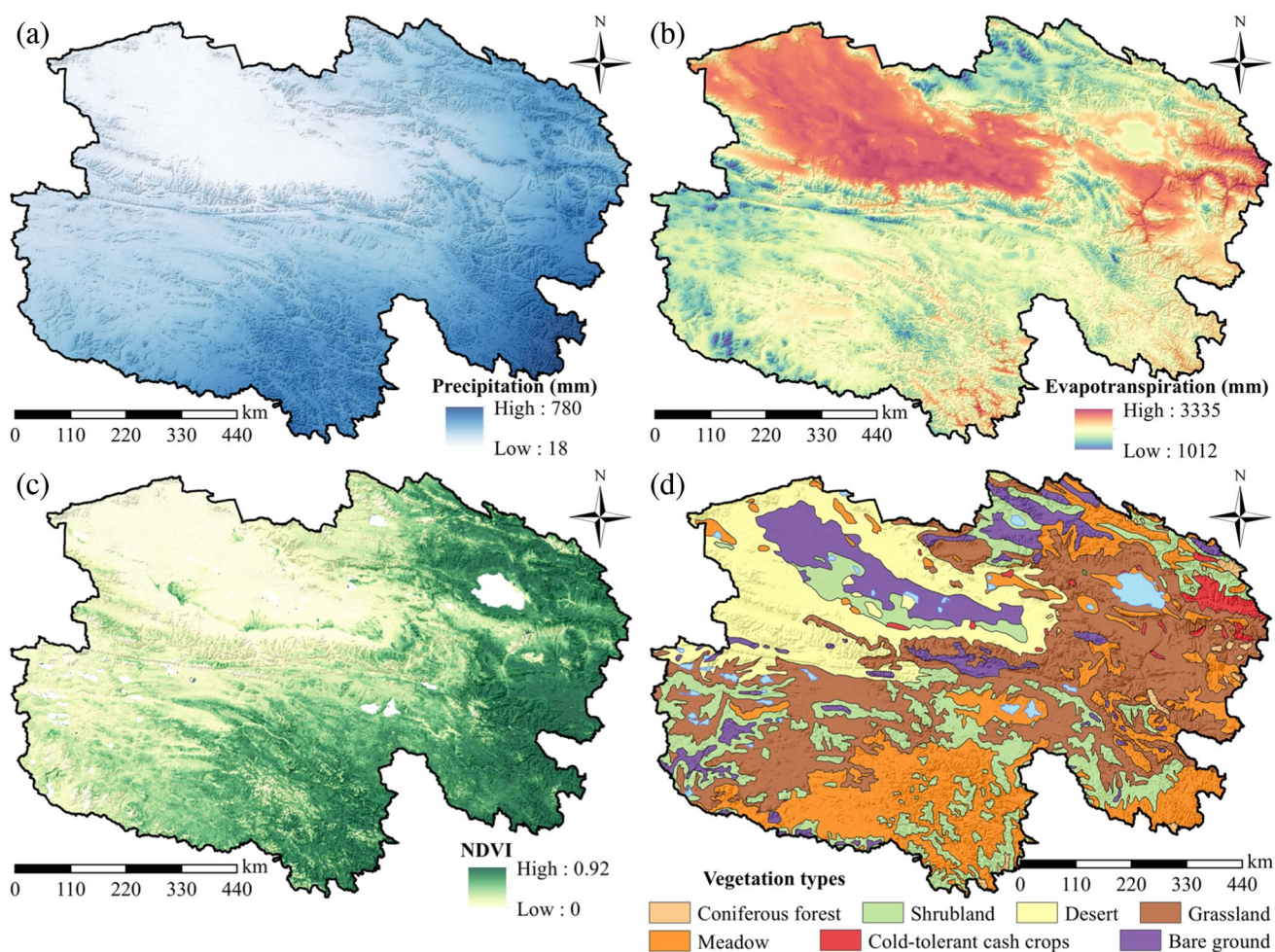


Fig. 2 Factors affecting groundwater potential: **a** precipitation (mm), **b** evapotranspiration (mm), **c** NDVI, and **d** vegetation types

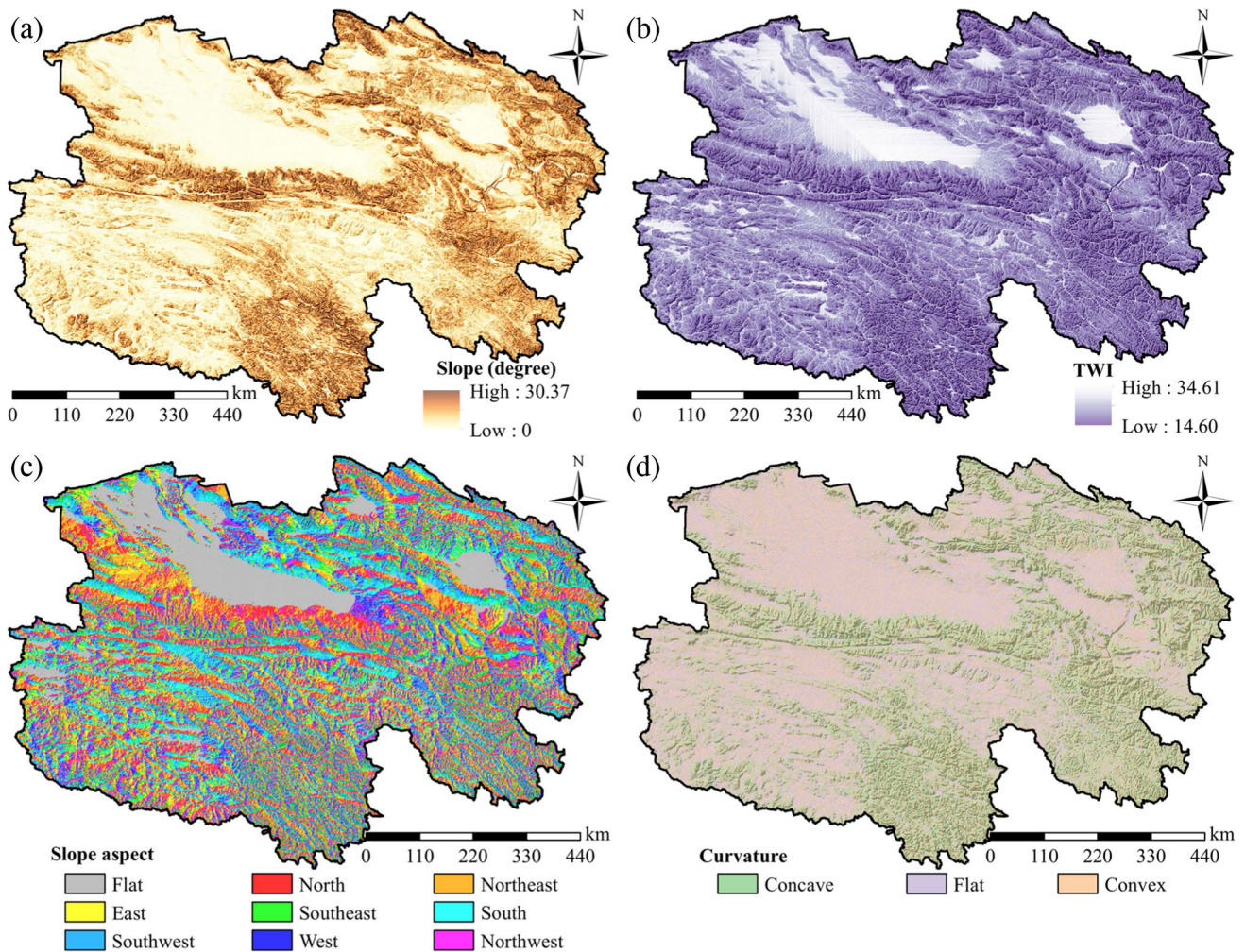


Fig. 3 Factors affecting groundwater potential: **a** slope (degree), **b** TWI, **c** slope aspect, and **d** curvature

Vegetation type and coverage are critical factors in regulating and using groundwater. Vegetation can impact soil water-holding capacity and the rate of evaporation, which leads to changes in groundwater recharge and the water table level (Orellana et al. 2012). In areas with ample vegetation, plants can reduce the evaporation rate of groundwater, and the root system can increase infiltration capacity. Consequently, this can decrease surface water runoff and improve groundwater recharge. However, vegetation also absorbs more groundwater and subsequently releases it into the atmosphere through transpiration, which can lead to groundwater depletion. To assess the impact of vegetation on groundwater potential, this study considers vegetation types and the normalized difference vegetation index (*NDVI*) as factors. We divided the vegetation in Qinghai Province into six types: coniferous forests, shrublands, deserts, grasslands, meadows, and bare grounds. *NDVI* is calculated using multispectral remote sensing data using the following formula (Han et al. 2021):

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad (1)$$

where *NIR* represents the near-infrared band's reflectance value and *RED* represents the reflectance value of the red band. Both *NIR* and *RED* are reflectance values ranging from 0 to 1. The resulting *NDVI* values range from 0 to 1 (Fig. 2(c)), with higher values indicating greater vegetation growth and coverage. We calculated the *NDVI* index of Qinghai Province using MODIS images (accessed from <https://glovis.usgs.gov>), which ranged from 0 to 0.92 (Fig. 2(d)).

Slope is the measure of the degree of rise or fall in the vertical direction as the surface moves a particular distance in the horizontal direction, typically expressed in angles. On steeper slopes, surface runoff flows more quickly and infiltrates less into the groundwater. In contrast, in areas with gentle slopes, surface water is more likely to recharge the groundwater (Naghibi et al. 2015a).

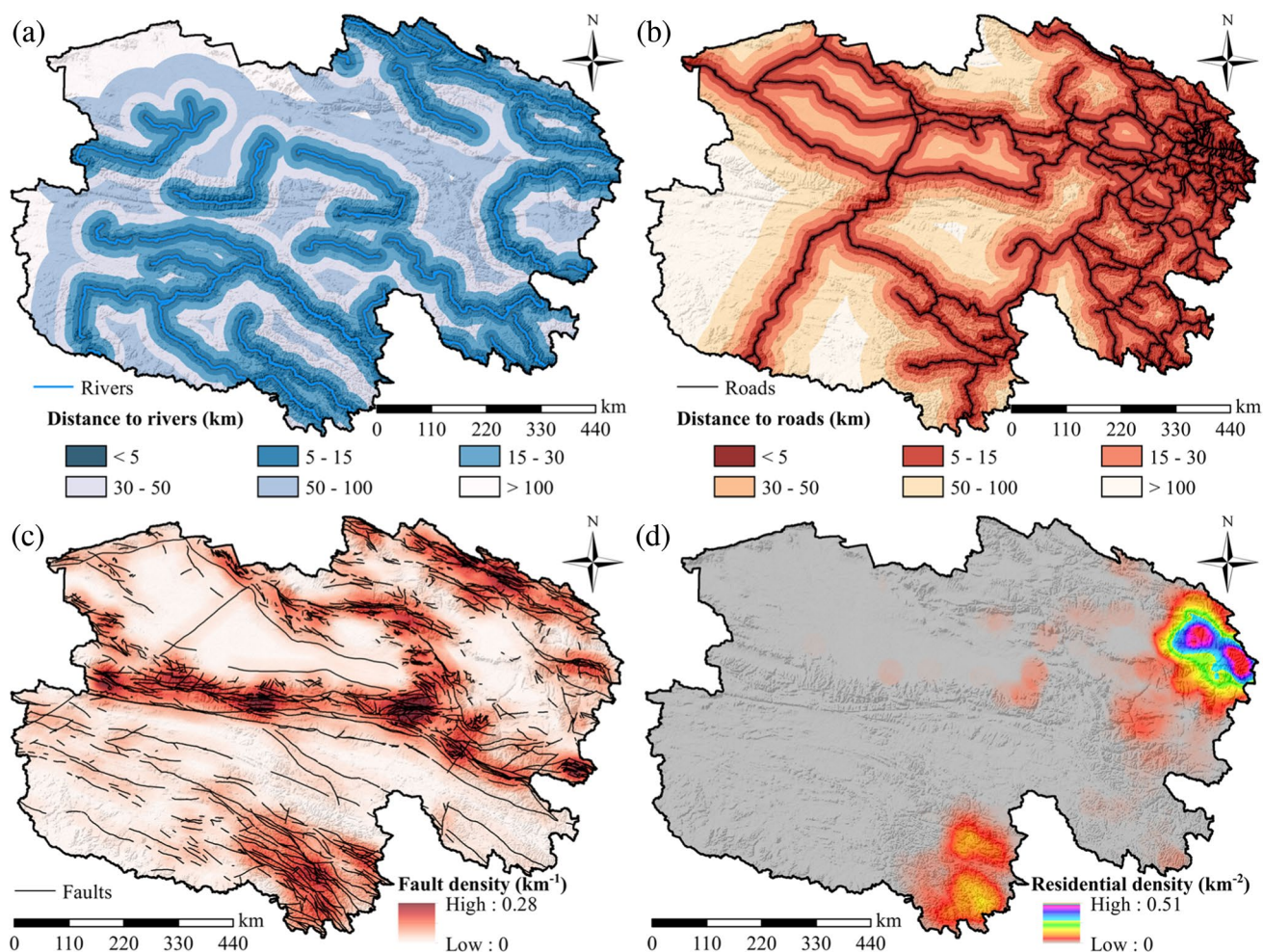


Fig. 4 Factors affecting groundwater potential: **a** distance to rivers (km), **b** distance to roads (km), **c** fault density (km^{-1}), and **d** residential density (km^{-2})

The study area's slope data can be extracted from the digital elevation model (DEM; obtained from <https://gscloud.cn>), which is calculated based on the elevation difference and horizontal distance from one DEM grid to the surrounding eight DEM grids. The slope range of Qinghai Province is $0\text{--}30.37^\circ$ (Fig. 3(a)). Slope aspect is the surface slope's orientation, specifically, the direction with the steepest slope expressed as an angle relative to the north direction. Slope aspect affects the direction of water flow in surface runoff, leading to variations in groundwater recharge across space (Singh et al. 2019). The slope aspect data were obtained by extracting the direction with the most significant elevation difference of the surrounding grid values within the DEM. Based on the aspect angle, the slope aspect of the study area can be categorized into nine categories: flat, north, northeast, east, southeast, south, southwest, west, and northwest (Fig. 3(b)). Curvature is a measure used to describe the topography of a surface. This metric has a direct

impact on the flow and infiltration of surface runoff. In regions with lower curvature, concave areas can develop, which are prone to retaining and accumulating surface water (Arabameri et al. 2019). These areas facilitate the complete recharge of groundwater. In contrast, regions with higher curvature are convex, allowing for the swift flow and flooding of surface runoff and resulting in a decreased supply of groundwater. To determine the curvature data for the study area, we extracted the values from the DEM data and sorted them into three categories based on magnitude: concave (< 0), flat (0), and convex (> 0) (Fig. 3(c)).

TWI is a metric that quantifies the potential for water retention in soil and vegetation within a specific region. It is calculated based on DEM using the following formula:

$$TWI = \ln \left(\frac{\alpha}{\tan \beta} \right) \quad (2)$$

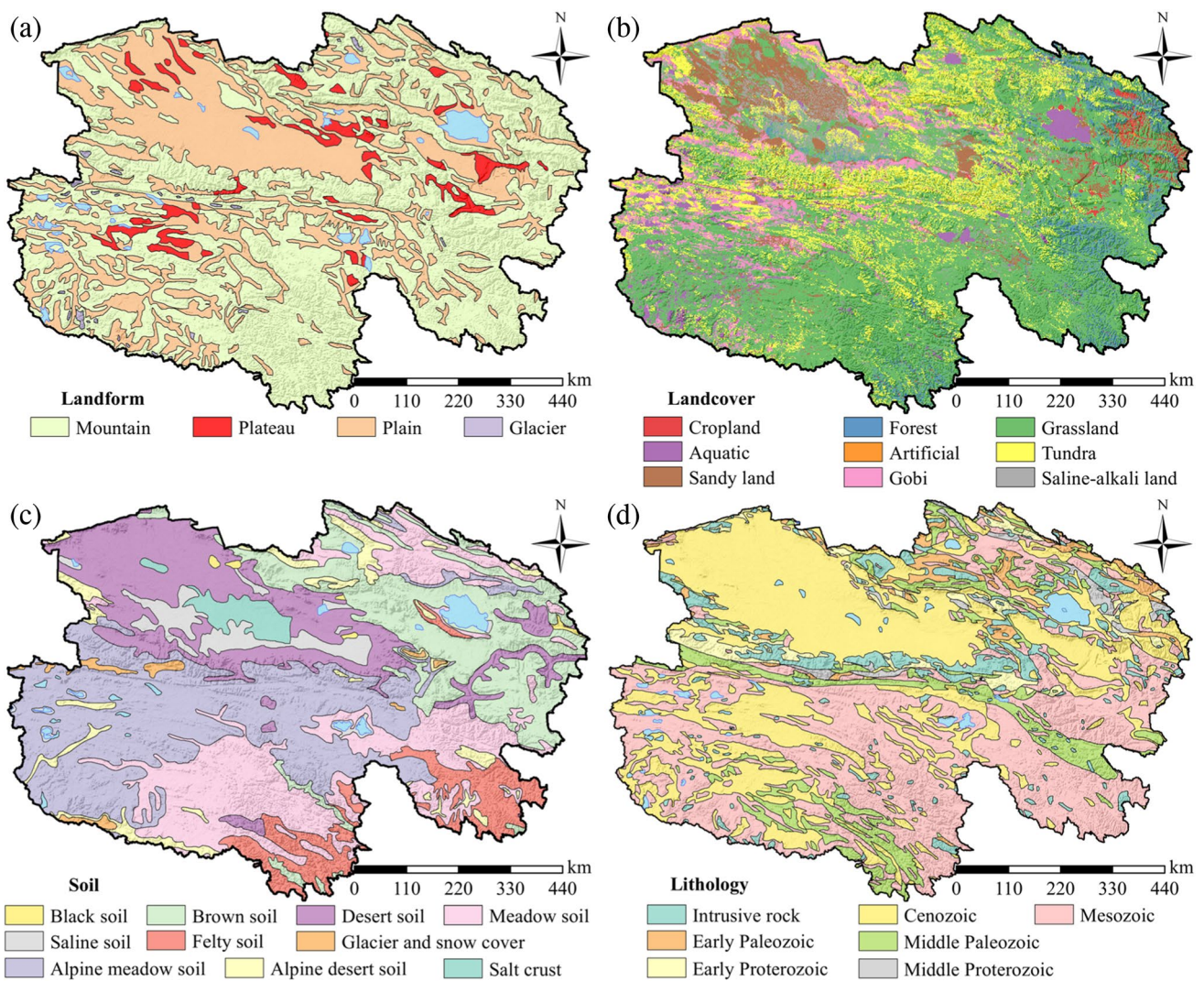


Fig. 5 Factors affecting groundwater potential: **a** landform, **b** landcover, **c** soil, and **d** lithology

$$\alpha = (FA + 1) \times S \tag{3}$$

where α denotes the catchment area, which can range from 0 to the total area of the catchment. β represents the slope's tangent, which can range from 0 (for a flat surface) to infinity (for a vertical surface). FA is the flow accumulation value, which can range from 0 (for areas with no inflow) to a large number representing the total inflow into a point. S corresponds to the grid area, which is a fixed value based on your grid resolution. A higher TWI value indicates poor drainage and longer water retention times, which can contribute to maintaining soil moisture and increasing groundwater supply (Rinderer et al. 2014). Conversely, a lower TWI value indicates good drainage, lower soil moisture content, and relatively lower groundwater supply. In Qinghai Province, the TWI was calculated to be between 14.60 and 34.61 (Fig. 3(d)).

The connection between rivers and groundwater is clear. During periods of high water flow, water infiltrates into the groundwater, boosting groundwater recharge. Conversely, during droughts, groundwater sustains the river's flow (Golokarian et al. 2018). Groundwater reservoirs located near rivers are typically permeable and mobile, allowing for faster water flow and spreading. In Qinghai Province, rivers are categorized into six classes based on their distance from the river: < 5, 5–15, 15–30, 30–50, 50–100, and > 100 km (Fig. 4(a)). Road construction can interrupt soil connectivity, hindering groundwater flow and drainage as well as changing the direction of groundwater flow (Velis et al. 2017). In this study, we classified the distances from roads in Qinghai Province into six categories: < 5, 5–15, 15–30, 30–50, 50–100, and > 100 km (Fig. 4(b)).

Varied fault densities can have diverse effects on groundwater distribution and flow. Faults can be influenced by

vertical stresses and are susceptible to deformation and rupture, which, in turn, facilitate the flow and penetration of groundwater, resulting in abundant underground water supply in those areas (Ahmad et al. 2021). However, high-density subsurface faults can also interfere with water flow during certain times. The fault density in Qinghai Province has been calculated to range from 0 to 0.28 km^{-1} (Fig. 4(c)). In areas with high population density, there is an inherent rise in water demand which results in the extraction of groundwater and a resultant drop in the groundwater table. Furthermore, urban areas with dense populations have increased conversion rates between surface water and groundwater, for example, rainwater seeping into sewers and underground pipes. The study area houses most of its population in the eastern region, where the resident density ranges from 0 to 0.51 km^{-2} (Fig. 4(d)).

The undulations within the terrain dictate the height and rate of groundwater flow, as well as its direction. In comparison to flat terrain, areas with tortuous topography and significant fluctuations encounter more fluctuations in groundwater levels and experience complex water flow dynamics. In steep and irregular areas, ground rainfall quickly gathers to form rivers and streams, leading to water loss, while mountainous areas with pitted terrain are more favorable to groundwater accumulation and retention (Subba Rao 2006). Additionally, solid precipitation, like snow, can influence hydrological processes through subsurface processes, in varying geomorphological regions. The study area is characterized by four basic landform types: mountain, plateau, plain, and glacier (Fig. 5(a)). Differential land use and land management practices have a profound impact on recharge rates and aquifer storage capacity. For instance, urban expansion contributes to enlarged impervious surfaces such as buildings and roads that reduce infiltration, cause surges in stormwater runoff, and thus cause a decline in groundwater recharge rates. Conversely, intensive agricultural practices, such as irrigation, can exhaust or deplete groundwater reservoirs. In Qinghai Province, the types of land cover (data obtained from <http://globallandcover.com>) can be categorized into nine types: cropland, forest, grassland, aquatic, artificial, tundra, sandy land, Gobi, and saline-alkali land (Fig. 5(b)).

The interaction between surface water and groundwater and the connection between them is reflected in soil and lithology, as they both play a crucial role in this process. Different types of soil and lithology exhibit varying levels of permeability, water storage, and drainage capabilities, and thus, influence the hydrodynamic attributes of groundwater (Wang et al. 2022). The study area encompasses ten soil categories: black soil, brown soil, desert soil, meadow soil, saline soil, felty soil, glacier and snow cover, alpine meadow soil, alpine desert soil, and salt crust (Fig. 5(c)). Lithology,

on the other hand, is divided into seven distinct categories: intrusive rock, Cenozoic, Mesozoic, early Paleozoic, middle Paleozoic, early Proterozoic, and middle Proterozoic (Fig. 5(d)).

Methodology

Figure 6 illustrates the systematic flow chart of this study, which is composed of four key steps:

- (1) The 16 conditioning factors were meticulously categorized into 5 categorical variables and 11 numerical variables. A one-hot encoding technique was employed to transform the categorical variables into 26 categories, while the numerical variables were used directly as continuous data and standardized.
- (2) During the machine learning training phase, four distinct methods were utilized to select different factors as sample features. The ALL feature subset incorporated all the factors, while the PCA method projected the factors onto 8-dimensional space. The Entropy and CRITIC methods were used as weighting techniques to quantify the weight values of the 16 factors were quantified, and the top 8 factors with higher weights were selected, respectively. This step is crucial for reducing dimensionality and focusing on the most influential features.
- (3) A range of training samples were sequentially input into the AutoML models sequentially, and the model yielding the highest accuracy was chosen. This step allows for an unbiased and automated selection of the best performing model. Furthermore, we conducted an experiment by incrementally increasing the sample size from 50 to 800 to assess the accuracy of the AutoML test set after each training iteration.
- (4) Finally, the study area was divided into 699,016 points which were used to create raster data to map the groundwater potential of the entire study area. By comparing the accuracy and generalization of model predictions using different model factors and sample sizes, an evaluation was conducted on the effect of selected samples and factors on prediction.

One hot encoding and Principal Component Analysis

One hot encoding is a technique used for encoding discrete features. It involves encoding a discrete factor into binary form, such that each feature's binary encoding is unique and distinct. To implement one hot encoding in the context of a groundwater potential analysis, discrete types are first assigned unique integers (Bai et al. 2022). These

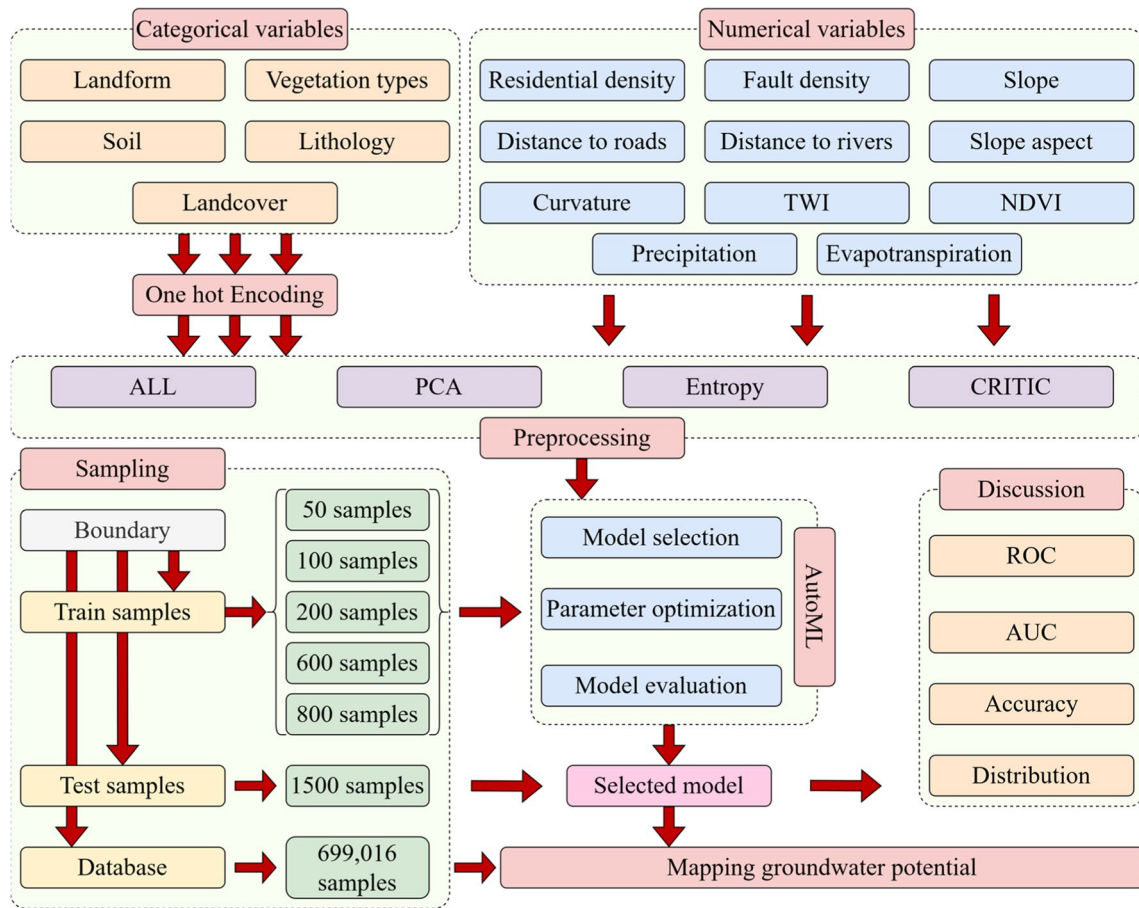


Fig. 6 Flowchart of the methodology

integers are then converted to binary numbers, as depicted in Fig. 7. The resulting matrix has rows that represent individual samples and columns that represent various discrete types of encoded bits. Each code bit can only have a value of 0 or 1, indicating whether a specific sample belongs to the corresponding discrete type (Pedregosa et al. 2011).

One hot encoding ensures that each variable type of input is equal in the model. If integer labels are utilized to encode different types, the machine learning model will learn that the size of the encoded value between different types has a quantitative relationship, potentially leading to inaccurate predictions by the model. One hot encoding changes each type into a binary classification, thereby increasing the interpretability of the model’s predictions. For this study, we utilized one hot encoding to encode five features, including landscape, vegetation types, soil, lithology, and landcover. These features cannot be expressed numerically; therefore, one hot encoding was used to encode them accurately.

PCA is a data dimensionality reduction method that transforms high-dimensional data into a lower dimensional space (Pan et al. 2016). Specifically, it performs a linear

transformation of the original data to a new coordinate system, finding the direction that maximizes the variance of the data in the new coordinate system, referred to as the first principal component. The second principal component is then found, which is orthogonal to the first principal component, and subsequent principal components are found successively, until the first k principal components are generated (Helena et al. 2000). These principal components comprise a new, lower dimensional space and have certain explanatory properties that can aid in understanding the data distribution. For instance, assuming that there is a groundwater potential assessment data set x containing m samples and n features, where $m, n \in \mathbb{N}$ (set of natural numbers). This data must first be standardized into y :

$$y_{ij} = \frac{x_{ij} - \mu_j}{\delta_j}, (i = 1, 2, 3, \dots, n; j = 1, 2, 3, \dots, m) \quad (4)$$

where δ_j represents the standard deviation of feature j and μ_j denotes its mean. Both δ_j and μ_j are real numbers ($\delta_j, \mu_j \in \mathbb{R}$). Once y has been obtained, the covariance matrix C can be computed. For any pair of features j and k , the covariance calculation expression is as follows:

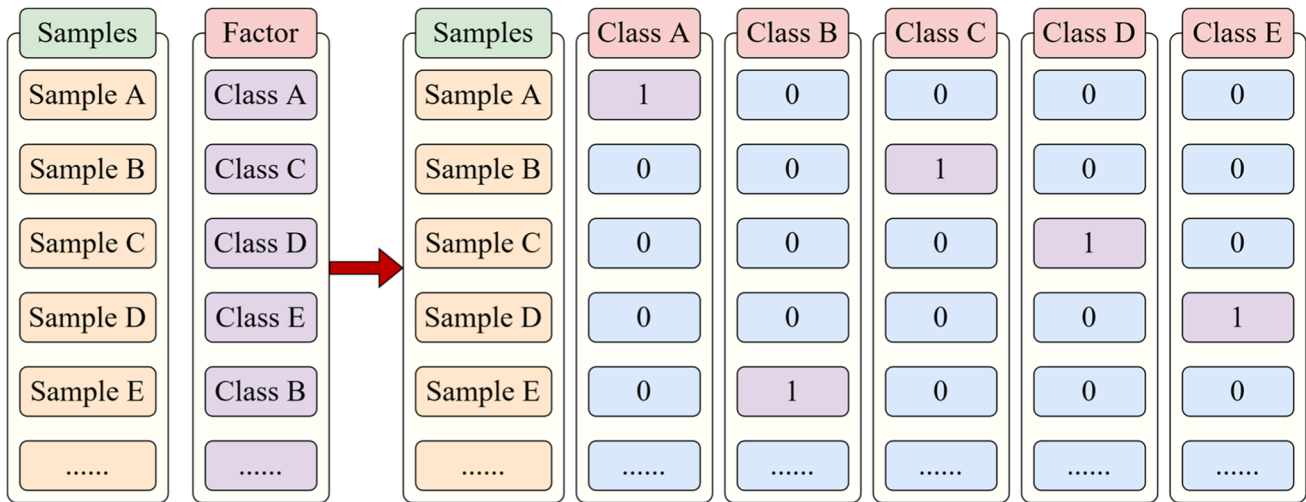


Fig. 7 One hot encoding

$$C_{jk} = \frac{1}{n-1} \sum_{i=1}^n (y_{ij} - \mu_j)(y_{ik} - \mu_k) \quad (5)$$

The eigenvalue decomposition of the covariance matrix C produces m eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_m$ ($\lambda_i \in \mathbb{R}$), and corresponding eigenvectors v_1, v_2, \dots, v_m . Each element of the eigenvector v_i represents the weight of the corresponding feature in the new coordinate system, which is also the direction in the new coordinate system. The eigenvalue λ_i represents the variance of the corresponding feature in the new coordinate system, used to measure the degree of dispersion of the data. The eigenvectors are sorted according to the eigenvalues, from largest to smallest, and the top k eigenvectors corresponding to the largest eigenvalues are selected as the principal components. The data is then projected onto the principal components to obtain the dimensionally reduced data matrix Z :

$$Z = V_k Y \quad (6)$$

Here, V_k represents the matrix composed of the first k principal components ($k \in \mathbb{N}$ and $k \leq m$), and Y is the normalized original data matrix. PCA is a commonly used technique, and its strengths include the ability to compress data dimension while retaining maximum information. During groundwater potential prediction, too many features may lead to the curse of dimensionality, due to the lack of training samples. As such, we reduced the 16 factors to eight features after applying PCA processing.

Entropy weight method and criteria importance through intercriteria correlation

The Entropy Weight Method (EWM) utilizes fuzzy mathematics theory and information entropy theory to calculate the weight of indicators (Zhang et al. 2021b). A matrix

N of m indicators and n samples can be formed for data that has been standardized or normalized, where $m, n \in \mathbb{N}$. The proportion P of the j th index in the i th sample can be obtained, which reflects the variation of the index, such that (Li et al. 2019)

$$P_{ij} = \frac{N_{ij}}{\sum_{i=1}^n N_{ij}} \quad (7)$$

Using P , the information entropy of the j th index is calculated as follows:

$$E_j = -\frac{1}{\ln n} \sum_{i=1}^n P_{ij} \ln P_{ij}, (j = 1, 2, \dots, m) \quad (8)$$

By using the above formula, the information entropy of each indicator can be calculated. Based on this, the weight γ_j of indicator j can be obtained through the information entropy of all indicators:

$$P_{ij} = \frac{1 - E_{ij}}{\sum_{i=1}^n (1 - E_{ij})} \quad (9)$$

The information entropy is used to reflect the degree of difference between evaluation indicators. The weight decreases as information entropy increases, reflecting a greater difference between evaluated indicators. Conversely, the weight increases as information entropy decreases, indicating a smaller difference between evaluation indicators.

CRITIC calculates the correlation coefficient between indicators to determine the degree of mutual influence between indicators, and subsequently, calculates their weight. Unlike the EMW, the CRITIC method requires

normalization, not standardization, before processing (Giao et al. 2023). This is because the weight evaluation criterion is based on the standard deviation. Different indicators require different normalization methods (Zhang et al. 2021a). For instance, indices that have a positive impact on groundwater potential enrichment are calculated using the expression:

$$y_{ij} = \frac{x_j - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

For indicators that have a negative impact on groundwater potential enrichment, y_{ij} can be calculated using

$$y_{ij} = \frac{x_{\max} - x_j}{x_{\max} - x_{\min}} \tag{11}$$

Using this, the amount of information C of the j th index can be calculated, as follows:

$$C_j = \sigma_j \sum_{k=1}^m (1 - r_{jk}) \tag{12}$$

where δ_j represents the standard deviation of index j ($\sigma_j \in \mathbb{R}$) and r_{jk} denotes the correlation coefficient between index j and k ($r_{jk} \in \mathbb{R}$). All C values are calculated based on the above formula, and ultimately, the weight vector for each indicator is obtained as follows:

$$W_j = \frac{C_j}{\sum_{k=1}^m C_k} \tag{13}$$

The CRITIC method pays attention to the relationship between indicators, unlike the EWM, but requires prior knowledge of the correlation between indicators. Through the EWM and CRITIC methods, the weights of the 16 factors were obtained. The eight factors with larger weights were then selected to predict groundwater potential, thereby reducing the model’s complexity and making groundwater prediction more interpretable.

Automated machine learning

Machine learning, a subfield of artificial intelligence, constructs general paradigms for predicting or classifying new data by recognizing patterns and rules in known datasets. However, the multitude of machine learning models available, each with varying effectiveness for different problems, makes choosing the best model a challenge. Existing machine learning models, whether simple single models like decision trees or complex ensemble models, all contain a wealth of hyperparameters. Therefore, the traditional process of building machine learning models involves algorithm selection and manual adjustment of hyperparameters, which requires a significant amount of time and effort.

AutoML is a process that automates these time-consuming iterative tasks in machine learning model development. It simplifies the application of machine learning by automatically selecting models, adjusting hyperparameters, and optimizing model performance (Feurer et al. 2015). It can quickly build high-quality machine learning models without the need for laborious manual tuning. The process of AutoML usually includes data preprocessing, feature engineering, model selection, hyperparameter tuning, and post-evaluation. This study mainly involves model selection and hyperparameter optimization. Hyperparameter tuning can be represented by the following formula:

$$x^* = \arg \min_x f(x) \tag{14}$$

In this formula, x^* represents the best model parameters and $f(x)$ represents the model’s loss function. We use Mean Squared Error (MSE) to represent it:

$$f(x) = \frac{1}{n} \sum_{i=1}^n (y_i - z_i)^2 \tag{15}$$

Here, n is the total number of training samples ($n \in \mathbb{N}$), y_i is the true value of the i th sample ($y_i \in \mathbb{R}$), and z_i is the prediction value of the i th sample by machine learning ($z_i \in \mathbb{R}$). Traditional hyperparameter optimization includes grid search and random search. Both methods exhaustively or randomly search possible parameter combinations in the parameter space to find the optimal solution. However, for parameters with higher dimensions, these two methods consume too much time and result in unreliable parameter selection results. This study adjusts machine learning hyperparameters by establishing a probability model of the loss function using Bayesian optimization. The specific steps are as follows:

- (1) Select some initial hyperparameter samples \mathbf{x} , and calculate their target function values $f(\mathbf{x})$.
- (2) Based on the existing samples \mathbf{x} and their corresponding target function values $f(\mathbf{x})$, establish a surrogate model $p(\cdot)$ for the target function $f(\cdot)$, usually using Gaussian Process.
- (3) Based on the surrogate model and observed data points, use an acquisition function to determine the next query point. Common acquisition functions include Expected Improvement (EI) and Probability of Improvement. This study uses EI , whose formula is as follows:

$$EI(x) = E[\max(f(\mathbf{x}) - f(x_i^*), 0)] \tag{16}$$

Here, x_i^* represents a batch of candidate points generated in the i th iteration. We calculate the EI values of all candidate points and select the point $x_{i^* \max}$ with the maximum

EI value as the next query point. We then calculate its corresponding target function value $f(x_{i^* \max})$ and add this data point to the sample set \mathbf{x} .

(4) Repeat steps 2 and 3 until a stopping condition is met, such as when the number of iterations reaches a preset value or when the target function value is below a certain threshold. In this way, we can achieve a balance between exploration (searching for unassessed areas) and exploitation (searching for known information), thereby effectively finding a global optimal solution.

The emergence of AutoML has considerably reduced the difficulty of machine learning modeling, making it more efficient and user-friendly. In this study, we selected five ensemble models for predicting groundwater potential. These models include Extra Trees (ET) (Geurts et al. 2006), Light Gradient Boosting Machine (LGBM) (Fan et al. 2019), L1-Regularized Logistic Regression (LRL1), Random Forest (RF) (Breiman 2001), Extreme Gradient Boosting (XGB) (Chen and Guestrin 2016), and XGB limit depth (XGBLD). Compared to other single models, these ensemble models can achieve superior performance and more robust generalization results. We utilized FLAML (Wang et al. 2021), a Python AutoML framework, to automate the process of model selection and hyperparameter tuning. This allowed us to select and optimize the most effective machine learning models for predicting groundwater potential under varying characteristics and sample sizes.

Results and discussion

Factor correlation and importance

The study utilized Pearson’s correlation coefficient to determine the correlations between potential groundwater influencing factors (Chen et al. 2018) in Qinghai Province (Fig. 8). The curvature feature did not pass the null hypothesis rejection test for vegetation types (0.9986), soil (0.1684), lithology (0.1374), residential density (0.9214), fault density (0.7576), distance to roads (0.8059), and distance to rivers (0.9462), indicating that curvature has no statistical correlation with these factors. Except for distance to roads (0.6287) and *TWI* and slope and soil (0.7254), the *p*-values of all other factors were less than 0.01, indicating a correlation between most factors. Furthermore, a significant negative correlation exists between evaporation and landform, reflected by a correlation coefficient of -0.8 . This can be explained by the high altitude of the Qinghai Plateau, to which Qinghai Province belongs, compared to the Qaidam Basin located in the northwest of Qinghai Province that has a lower altitude and a higher annual evaporation rate of over 3000 mm per annum (Fig. 2(b)). As such, landform and evaporation showed a significant negative correlation. The correlation coefficient between precipitation and *NDVI* is 0.8, which can be attributed to the little precipitation in the plateau desert climate of the study area. Increased precipitation leads to more vegetation growth resulting in an increase in *NDVI*

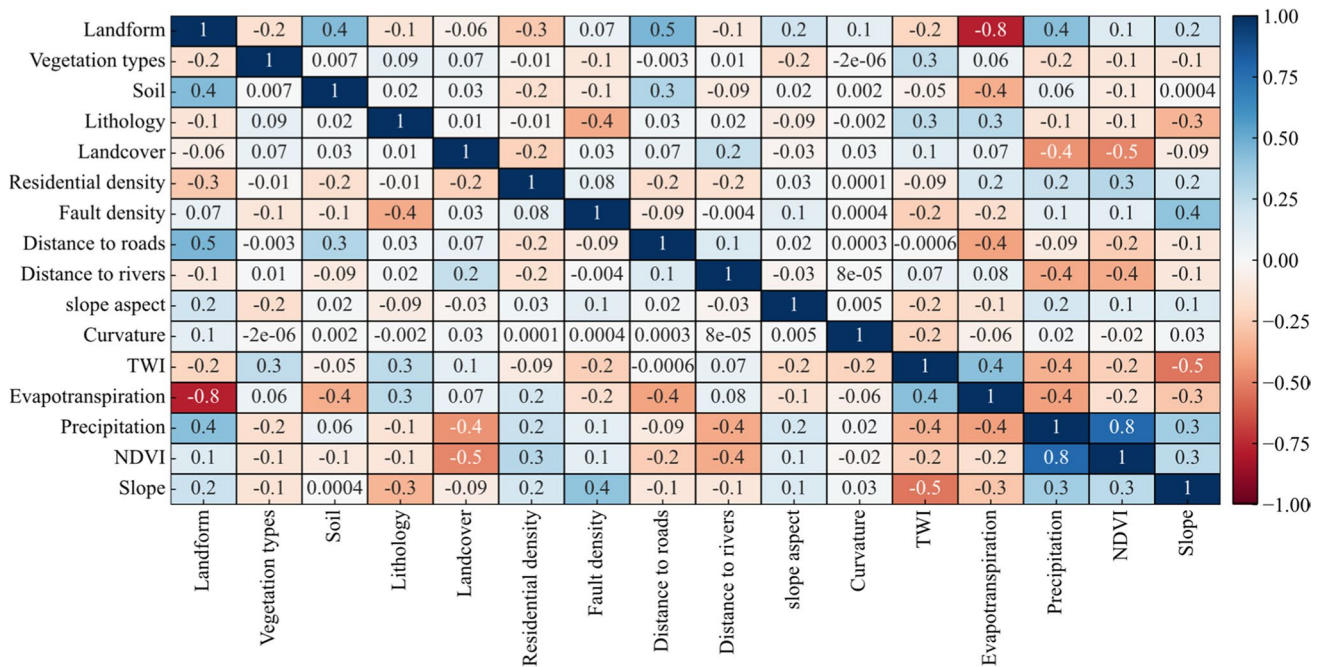


Fig. 8 Heatmap of factor correlations

value. Moreover, although some factors, such as slope, slope aspect, and *TWI*, were extracted from DEM data, their linear correlations between each other were relatively low, with positive correlation coefficients less than or equal to 0.5, and negative correlation coefficients greater than or equal to -0.5 . In conclusion, the generally low Pearson correlation coefficients among the various factors suggest a weak correlation between the factors. Each factor shows a high degree of independence, which allows them to perform their respective roles in predicting the groundwater potential of the study area effectively.

Figure 9 displays the weights of all factors calculated through both the CRITIC and EWM methods. The EWM approach determined the weights of the 16 factors from largest to smallest as follows: evapotranspiration (0.390), landform (0.325), curvature (0.107), soil (0.051), lithology (0.041), fault density (0.021), distance to rivers (0.021), *NDVI* (0.011), precipitation (0.010), distance to roads (0.006), residential density (0.005), slope (0.004), land cover (0.004), *TWI* (0.002), slope aspect (0.002), and vegetation types (0.002). The CRITIC method determined their weights in the following descending order: landform (0.472), evapotranspiration (0.248), precipitation (0.102), slope aspect (0.0640), distance to roads (0.050), distance to rivers (0.0270), soil (0.0140), land cover (0.009), lithology (0.007), vegetation (0.003), slope (0.002), *TWI* (0.002), *NDVI* (<0.001), curvature (<0.001), fault density (<0.001), and residential density (<0.001). Due to differences in the distribution of weights and decision-making objectives, the weights assigned to some indicators are inconsistent. Nevertheless, both methods illustrate that landform and evapotranspiration are critical factors in groundwater enrichment in Qinghai. The two factors that exhibit the greatest difference between the two methods are curvature and precipitation. Based on the weight values calculated by the methods, the

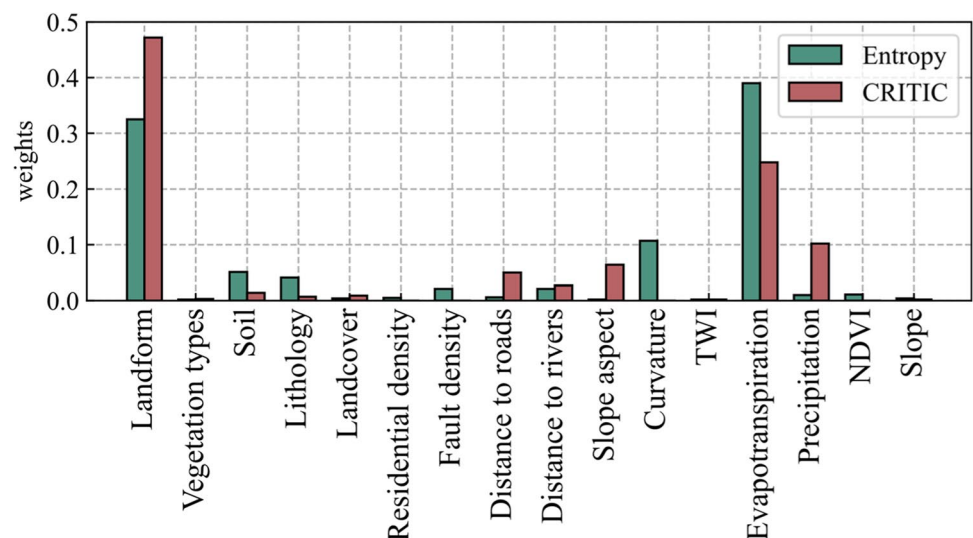
first eight factors with the highest weights are selected for machine learning training features.

Influence of factor selection and sample quantity on groundwater potential prediction

Using AutoML, we employed four distinct feature selection methods to train machine learning models with a sample size of 200, focusing on groundwater conditioning factors. The machine learning models, selected by AutoML corresponding to the four factor selection methods, were XGB, XGB, RF, and ET. Figure 10 (a) and (b) illustrate the scores of the prediction model trained by four different factor selection modes using 1500 test samples. The accuracy scores were 0.783, 0.685, 0.745, and 0.703, respectively, with an area under curve (AUC) of 0.819, 0.724, 0.779, and 0.747. For a more comprehensive performance assessment, Table 1 provides detailed metrics such as accuracy, precision, AUC, recall, and F_1 score. Notably, apart from precision, XGB-ALL outperforms the other models, followed by RF Entropy, ET-CRITIC, and XGB-PCA.

The ALL factor selection mode yielded the highest accuracy, indicating that all factor choices are reasonably valid. This is because, while reducing some factors, the model's accuracy decreases as well. Furthermore, the primary utilization of a tree-based ensemble model in the prediction model, as opposed to a linear model, ensured that our predictions did not encounter issues with multicollinearity. Linear models quite often face multicollinearity problems since their trained features exhibit a linear relationship with each other. In contrast, tree-based models are designed so that each node in the tree depends on a single optimal feature to divide the data, meaning that each node only utilizes a specific feature, thereby minimizing the intricacy of the connections between features (Paul et al. 2018). Additionally, with

Fig. 9 Factor weights calculated using the EWM and CRITIC



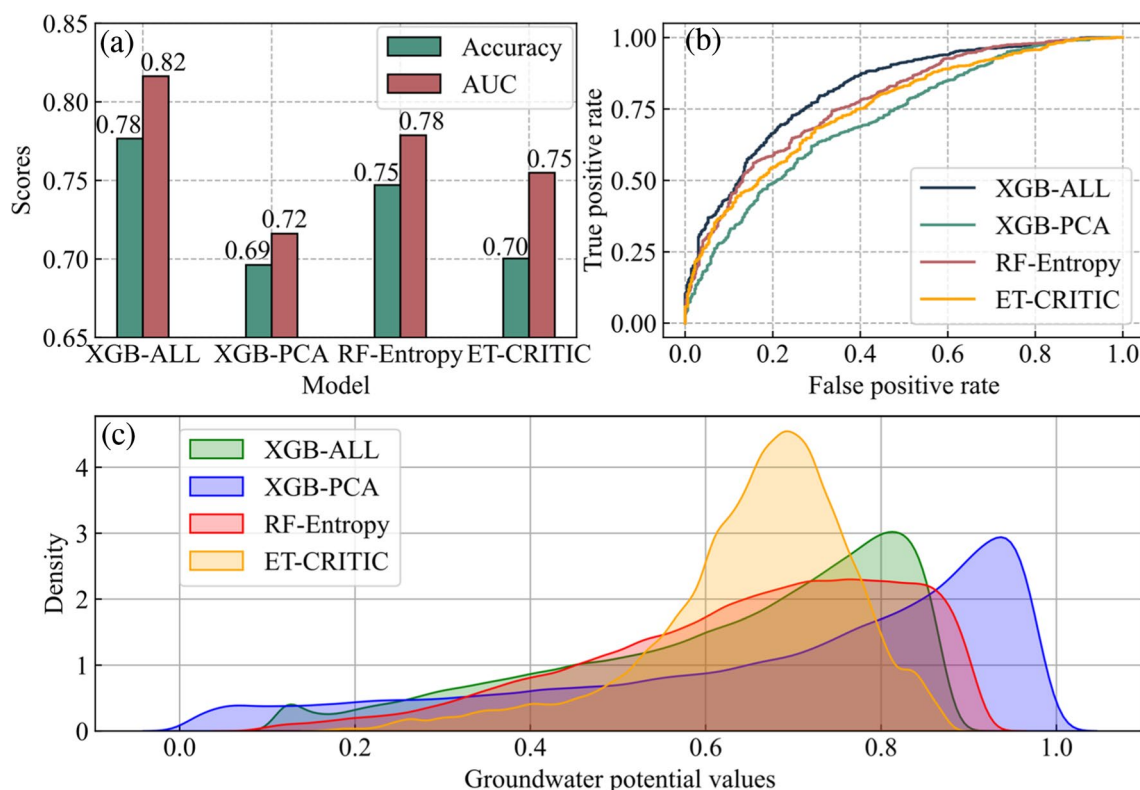


Fig. 10 **a** Accuracy and AUC values of the models. **b** ROC curves of the models. **c** Density distribution of groundwater potential in the study area

Table 1 The performance of the four models under various assessment criteria

	Accuracy	Precision	AUC	Recall	F_1 score
XGB-ALL	0.783	0.748	0.819	0.942	0.834
XGB-PCA	0.685	0.730	0.724	0.830	0.777
RF-Entropy	0.745	0.752	0.779	0.916	0.826
ET-CRITIC	0.703	0.767	0.747	0.856	0.809

ensemble models such as RF or XGB, the characteristics employed in each tree are diverse, further mitigating collinearity issue among them. Consequently, using a tree-based ensemble model with an appropriate number of samples can ensure higher accuracy in predicting models involving multiple features.

The XGB-PCA method shows the lowest accuracy in predicting groundwater potential models. This is because of the inadequate correlation among the various factors, especially after applying one hot encoding, leading to a lack of correlation between each encoding category. But PCA works well when variables possess a robust correlation and loses more information when the correlation is weak. Therefore, when the 16 factors are reduced to eight dimensions, XGB-PCA's forecast is inaccurate. The RF-Entropy and ET-CRITIC

methods exhibit moderate effects on predicting groundwater potential in Qinghai Province. These methods utilize their calculated weights to pick only the eight factors with the most significant weights for training. It is noteworthy that RF-Entropy exhibits greater accuracy than ET-CRITIC, which indicates that information entropy is a better weight-assignment technique for determining factors influencing groundwater potential in Qinghai Province.

The spatial distribution of groundwater potential in Qinghai Province, as drawn under the four factor selection modes, is represented in Fig. 11, while its density distribution is shown in Fig. 10(c). The prediction results have been stratified into five separate categories using the natural break-point method, namely, very low, low, moderate, high, and very high, in order to differentiate the types of groundwater potential. The values of groundwater potential in the region are generally higher than 0.5, and the areas with high and very high potentials are primarily concentrated in the southwest and southeast regions of the study area, which are the primary source of three critical rivers in China. Conversely, the low groundwater potential areas are concentrated in the north-western part of the study area, which is an arid region with major salt lake industries (Wang et al. 2022). Although the overall distribution of the four models is similar, the density distribution of groundwater potential values varies.

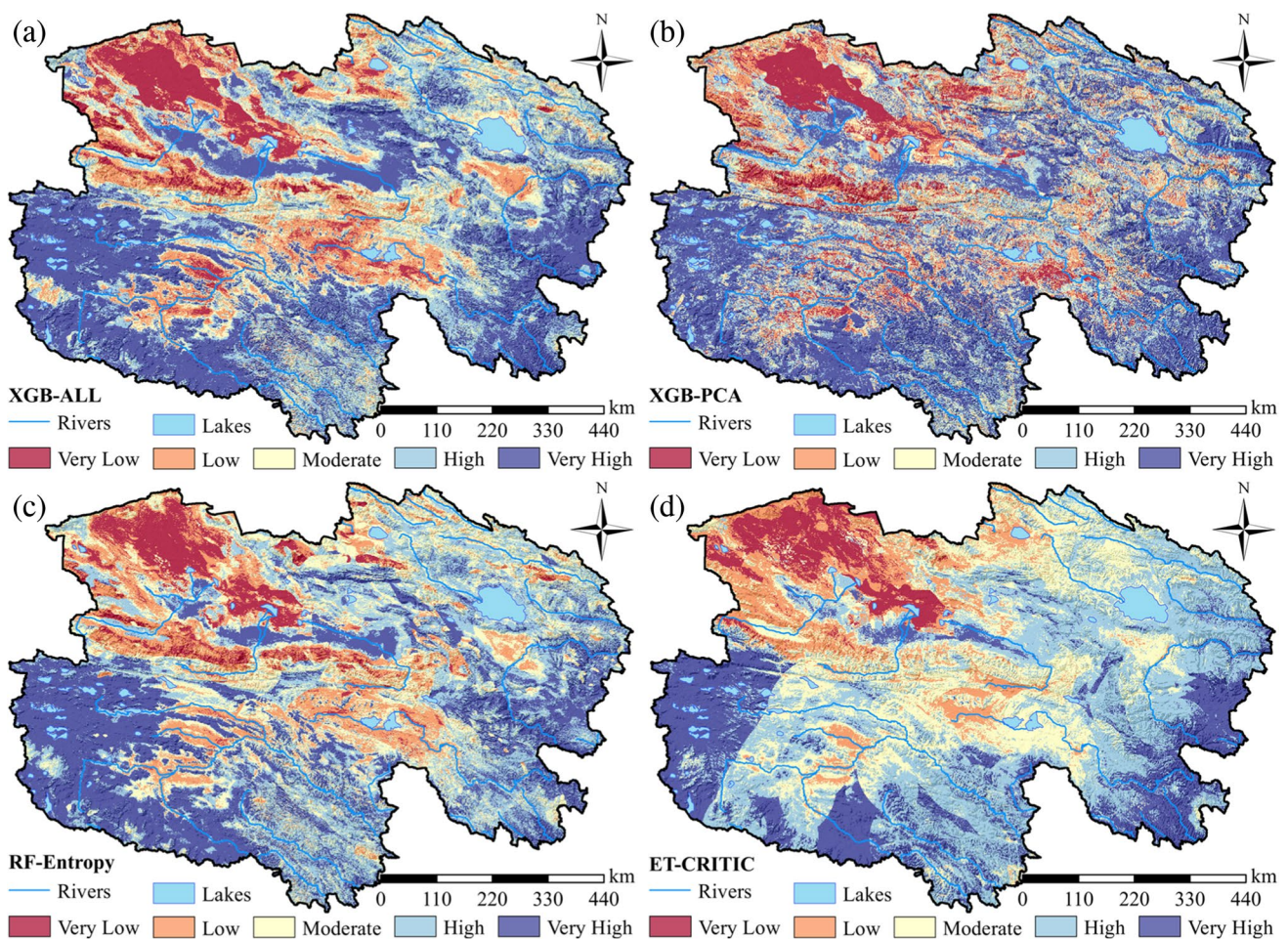


Fig. 11 Spatial distribution of groundwater potential: **a** XGB-ALL, **b** XGB-PCA, **c** RF-Entropy, and **d** ET-CRITIC

The prediction outcomes of the ET-CRITIC are more concentrated around 0.7, while those of XGB-PCA are between 0.9 and 1.0. In comparison, the results of RF-Entropy and XGB-ALL model are somewhat analogous (around 0.8). This demonstrates that utilizing the Entropy method to screen factors with higher weights for predicting groundwater potential can bring about effective dimensionality reduction without sacrificing accuracy.

Besides feature selection, the number of samples significantly affects the model results. Figure 12 depicts the performance of the model in predicting groundwater potential in Qinghai Province by varying sample sizes. As the number of samples increases under the AutoML training framework, the model's accuracy displays an upward trend with fluctuations. The model accuracy fluctuated around 0.7 but did not increase significantly as the sample size increased from 50 to 200. The accuracy and AUC improved when the number of samples exceeded 200, reaching an accuracy of approximately 0.9 after training with 600 samples, and then gradually stabilizing. In many

cases, the number of samples was limited due to certain conditions. Therefore, to attain a model accuracy and AUC of 0.75 and above, the assessment of regional groundwater potential should include a minimum of 200 samples. For a more detailed characterization of regional groundwater potential, the sample size must exceed 600.

Due to AutoML being re-run each time the number of samples is modified, different model types were selected on each occasion. Among the 6 used models, XGB was the most popular, with 44 executions, followed by the ET and RF models making an appearance 12 times. As for the LGBM, LRL1, and XGBLD models, they were included no more than five times. Thus, when predicting groundwater potential, priority can be given to the XGB model.

In conclusion, the primary hurdle in accurately assessing the groundwater potential of a specific area is not the performance of machine learning algorithms, but rather the scarcity of available samples. While various machine learning models, particularly well-established ensemble learning models, may yield different results across various research

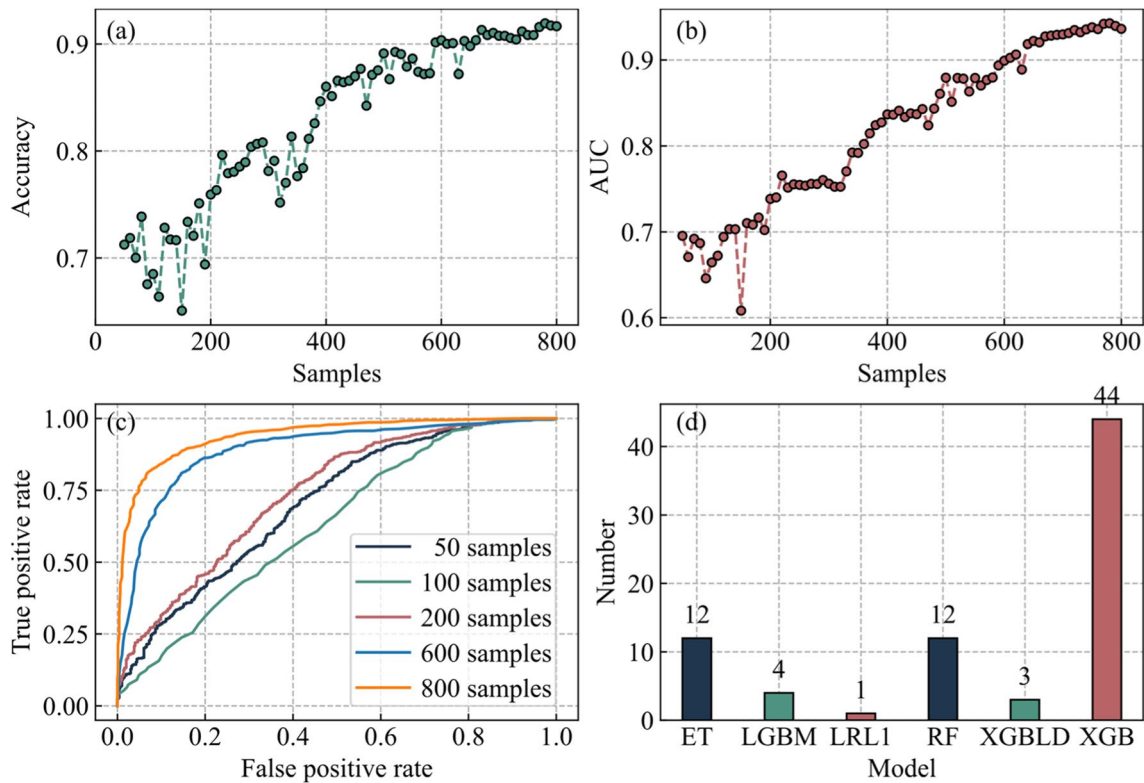


Fig. 12 Comparison of model performance with different sample sizes

areas and feature sets, these differences are typically minor. On the other hand, collecting a sufficient number of samples in a vast research area is a daunting task that requires significant time and financial resources. This challenge contradicts our initial intention of using machine learning algorithms, which is to achieve the most accurate predictions at the lowest cost. Furthermore, the lack of samples limits our ability to incorporate a large number of features for machine learning evaluation, leading to the curse of dimensionality. Therefore, the primary challenge in predicting groundwater potential lies in finding a balance between sample quantity and feature selection to achieve the most accurate results.

In this study, we utilized AutoML with the aim of streamlining the process of machine learning model selection and hyperparameter tuning, allowing us to concentrate on the samples and features themselves. The prediction of groundwater potential under various sample sizes and feature selection methods was carried out using AutoML, thereby minimizing potential biases arising from manual model and hyperparameter choices. When the sample size permits, the groundwater potential prediction model should include as many factors as possible to enhance accuracy. However, this approach depends on the use of an ensemble model based on tree models, such as XGBoost, as excessive multicollinearity among numerous factors may negatively impact the model's predictive performance, which tree models can

alleviate. In situations with limited sample sizes, it is advisable to limit the number of input features in the machine learning model. We observed that using the entropy method to evaluate the importance of all factors and selecting those with high weights for training can maximize the accuracy of groundwater potential prediction. While PCA (Principal Component Analysis) can reduce the number of factors and linear correlations between them during dimension reduction, if the correlations are weak, PCA may result in a loss of information and ultimately lead to less accurate predictions. Moreover, while increasing the sample size improves the accuracy of groundwater potential prediction, this improvement tends to plateau after reaching a certain scale. To effectively address the challenge of limited sample availability, we recommend that in the research area of Qinghai Province, a minimum sample size of 200 is necessary to achieve an accuracy level of 0.75. However, for higher precision requirements, a sample size of approximately 600 is needed.

This research outcome presents a fresh perspective on how to approach the issue of groundwater potential prediction and offers a novel method for tackling problems related to limited sample quantity and feature selection. We believe that this will contribute to the advancement of groundwater potential prediction for future research. Nevertheless, due to variations in geological, geographical, and human activity conditions across different regions, it is essential to use the

methods detailed in this study to reevaluate and determine the optimal sample size and feature selection approach when assessing groundwater potential, rather than directly applying the recommended values from this research.

Conclusions

This study leveraged AutoML technology to predict groundwater potential in Qinghai Province, with a particular focus on analyzing the influence of the feature selections and sample sizes on the predictions. The models were trained using 50 to 800 samples, while an additional 1500 were used for model evaluation. Sixteen groundwater conditioning factors in Qinghai Province were classified into categorical and numerical variables based on feature types. Categorical variables underwent one hot encoding to prevent the model from being misled by the quantitative relationship of integer classifications. Four different feature selection modes, including ALL, PCA, Entropy, and CRITIC, were employed to train the model. Upon training completion, the entire research area was discretized into 699,016 points and fitted into the trained model. The output results were subsequently transformed into maps of groundwater potential in Qinghai Province. The study revealed that despite the general statistical correlation among 16 groundwater conditioning factors, their Pearson correlation coefficients were low. This implies that when using the tree model to predict the groundwater potential, a larger number of features can be utilized as long as there are sufficient samples, thereby enhancing the accuracy of the model. Due to the weak linear correlation between factors, the PCA method struggled to effectively reduce model dimensionality, negatively impacting prediction performance. Conversely, using the Entropy method to screen factors with higher weights ensured better accuracy while also reducing dimensionality, thus circumventing the potential curse of dimensionality. Results from model training revealed that as the number of samples increases, so does the accuracy and AUC value of the groundwater potential prediction model. Training with 8 factors and 200 samples resulted in an accuracy of 0.745, sufficient for evaluating regional groundwater potential. On the other hand, training with 600 samples led to a model accuracy performance of 0.9, thus realizing accurate prediction of groundwater potential. In summary, when dealing with the small sample sizes and low degrees of linear correlation between factors, we recommend using the Entropy method to screen factors with higher weights based on sample size and employing the XGB model for groundwater potential prediction. This study provides both theoretical and practical support for decision-makers dealing with groundwater resource management in the Qinghai Province. The findings underscore the importance of feature selection and sample size in machine learning models for

groundwater prediction. Furthermore, the model and methodology developed in this research can also be applied for predicting groundwater potential in other regions.

Author contribution All authors contributed to the study conception and design. Material preparation and data collection were performed by Mengling Li. 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 National Natural Science Foundation of China (U20A2088), and the Qinghai Provincial Science and Technology Innovation Platform (2020-ZJ-T03).

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

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

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>
- Ahmad I, Hasan H, Jilani MM, Ahmed SI (2023) Mapping potential groundwater accumulation zones for Karachi City using GIS and AHP techniques. *Environ Monit Assess* 195:381. <https://doi.org/10.1007/s10661-023-10971-x>
- 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, 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>
- 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>
- Bai Z, Liu Q, Liu Y (2022) Groundwater potential mapping in Hubei region of China using machine learning, ensemble learning, deep learning and AutoML methods. *Nat Resour Res* 31:2549–2569. <https://doi.org/10.1007/s11053-022-10100-4>

- Bera A, Mukhopadhyay BP, Chowdhury P et al (2021) Groundwater vulnerability assessment using GIS-based DRASTIC model in Nangasai River Basin, India with special emphasis on agricultural contamination. *Ecotoxicol Environ Saf* 214:112085. <https://doi.org/10.1016/j.ecoenv.2021.112085>
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Cao W, Wu D, Huang L, Liu L (2020) Spatial and temporal variations and significance identification of ecosystem services in the Sanjiangyuan National Park. *China Sci Rep* 10:6151. <https://doi.org/10.1038/s41598-020-63137-x>
- Chen W, Li H, Hou E et al (2018) GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. *Sci Total Environ* 634:853–867. <https://doi.org/10.1016/j.scitotenv.2018.04.055>
- Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, 785–794
- 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>
- Fan J, Ma X, Wu L et al (2019) Light Gradient Boosting Machine: an efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agric Water Manag* 225:105758. <https://doi.org/10.1016/j.agwat.2019.105758>
- Farhat B, Souissi D, Mahfoudhi R et al (2023) GIS-based multi-criteria decision-making techniques and analytical hierarchical process for delineation of groundwater potential. *Environ Monit Assess* 195:285. <https://doi.org/10.1007/s10661-022-10845-8>
- Feurer M, Klein A, Eggenberger, Katharina Springenberg J, et al (2015) Efficient and robust automated machine learning. In: Advances in neural information processing systems. 2962–2970
- Fick SE, Hijmans RJ (2017) WorldClim2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol* 37:4302–4315. <https://doi.org/10.1002/joc.5086>
- Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Mach Learn* 63:3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Giao NT, Nhien HTH, Anh PK, Thupitdang P (2023) Groundwater quality assessment for drinking purposes: a case study in the Mekong Delta. *Vietnam Sci Rep* 13:4380. <https://doi.org/10.1038/s41598-023-31621-9>
- 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>
- Han Z, Song W, Deng X (2016) Responses of ecosystem service to land use change in Qinghai Province. *Energies* 9:303. <https://doi.org/10.3390/en9040303>
- 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>
- Helena B, Pardo R, Vega M et al (2000) Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Res* 34:807–816. [https://doi.org/10.1016/S0043-1354\(99\)00225-0](https://doi.org/10.1016/S0043-1354(99)00225-0)
- Jhariya DC, Khan R, Mondal KC et al (2021) Assessment of groundwater potential zone using GIS-based multi-influencing factor (MIF), multi-criteria decision analysis (MCDA) and electrical resistivity survey techniques in Raipur City, Chhattisgarh, India. *J Water Supply Res Technol-Aqua* 70:375–400. <https://doi.org/10.2166/aqua.2021.129>
- Jia S, Zhu W, Lú 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>
- 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>
- Kong R, Xue F, Wang J et al (2017) Research on mineral resources and environment of salt lakes in Qinghai Province based on system dynamics theory. *Resour Policy* 52:19–28. <https://doi.org/10.1016/j.resourpol.2017.01.006>
- 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 Z, Zhou P, Zhang F et al (2013) Spatiotemporal characteristics of dryness/wetness conditions across Qinghai Province, Northwest China. *Agric for Meteorol* 182–183:101–108. <https://doi.org/10.1016/j.agrformet.2013.05.013>
- Naghbi SA, Pourghasemi HR, Dixon B (2015a) 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>
- Naghbi SA, Pourghasemi HR, Pourtaghi ZS, Rezaei A (2015b) Groundwater qanat potential mapping using frequency ratio and Shannon's entropy models in the Moghan watershed. *Iran Earth Sci Inform* 8:171–186. <https://doi.org/10.1007/s12145-014-0145-7>
- Naghbi 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>
- Orellana F, Verma P, Loheide II SP, Daly E (2012) Monitoring and modeling water-vegetation interactions in groundwater-dependent ecosystems. *Rev Geophys* 50 <https://doi.org/10.1029/2011RG000383>
- 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
- 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>
- Reichstein M, Camps-Valls G, Stevens B et al (2019) Deep learning and process understanding for data-driven Earth system science. *Nature* 566:195–204. <https://doi.org/10.1038/s41586-019-0912-1>

- Rinderer M, van Meerveld HJ, Seibert J (2014) Topographic controls on shallow groundwater levels in a steep, prealpine catchment: when are the TWI assumptions valid? *Water Resour Res* 50:6067–6080. <https://doi.org/10.1002/2013WR015009>
- Rizeei HM, Pradhan B, Saharkhiz MA, Lee S (2019) Groundwater aquifer potential modeling using an ensemble multi-adoptive boosting logistic regression technique. *J Hydrol* 579:124172. <https://doi.org/10.1016/j.jhydrol.2019.124172>
- Rostamzadeh R, Ghorabae MK, Govindan K et al (2018) Evaluation of sustainable supply chain risk management using an integrated fuzzy TOPSIS-CRITIC approach. *J Clean Prod* 175:651–669. <https://doi.org/10.1016/j.jclepro.2017.12.071>
- 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>
- 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>
- Singh SK, Zeddies M, Shankar U, Griffiths GA (2019) Potential groundwater recharge zones within New Zealand. *Geosci Front* 10:1065–1072. <https://doi.org/10.1016/j.gsf.2018.05.018>
- Subba Rao N (2006) Groundwater potential index in a crystalline terrain using remote sensing data. *Environ Geol* 50:1067–1076. <https://doi.org/10.1007/s00254-006-0280-7>
- 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>
- Sun X, Zhou Y, Yuan L et al (2021) Integrated decision-making model for groundwater potential evaluation in mining areas using the cusp catastrophe model and principal component analysis. *J Hydrol Reg Stud* 37:100891
- 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>
- Thanh NN, Thunyawatcharakul P, Ngu NH, Chotpanarat S (2022) Global review of groundwater potential models in the last decade: parameters, model techniques, and validation. *J Hydrol* 614:128501. <https://doi.org/10.1016/j.jhydrol.2022.128501>
- Velis M, Conti KI, Biermann F (2017) Groundwater and human development: synergies and trade-offs within the context of the sustainable development goals. *Sustain Sci* 12:1007–1017. <https://doi.org/10.1007/s11625-017-0490-9>
- 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>
- Wang Z, Wang J, Yu D, Chen K (2023a) The potential evaluation of groundwater by integrating rank sum ratio (RSR) and machine learning algorithms in the Qaidam Basin. *Environ Sci Pollut Res* 30:63991. <https://doi.org/10.1007/s11356-023-26961-y>
- Wang Z, Wang J, Yu D, Chen K (2023b) Groundwater potential assessment using GIS-based ensemble learning models in Guanzhong Basin. *China Environ Monit Assess* 195:690. <https://doi.org/10.1007/s10661-023-11388-2>
- Wang C, Wu Q, Weimer M, Zhu E (2021) FLAML: a fast and lightweight AutoML library. In: Fourth conference on machine learning and systems 3:434–447. <https://doi.org/10.48550/arXiv.1911.04706>
- 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>
- Zhang Q, Qian H, Xu P et al (2021a) Groundwater quality assessment using a new integrated-weight water quality index (IWQI) and driver analysis in the Jiaokou Irrigation District. *China Ecotoxicol Environ Saf* 212:111992. <https://doi.org/10.1016/j.ecoenv.2021.111992>
- Zhang Y, Jia R, Wu J et al (2021b) 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.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law. Authors and Affiliations

Authors and Affiliations

Zitao Wang^{1,2,3} · Jianping Wang^{1,2}  · Mengling Li^{1,2,3}

✉ 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