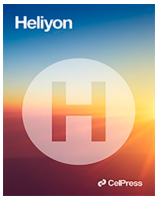
Heliyon 10 (2024) e27867

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Mapping specific groundwater nitrate concentrations from spatial data using machine learning: A case study of chongqing, China

Yuanyi Liang [a,*], Xingjun Zhang [a], Lin Gan [b], Si Chen [a], Shandao Zhao [b], Jihui Ding [b], Wulue Kang [a], Han Yang [a]

[a] *Observation and Research Station of Ecological Restoration for Chongqing Typical Mining Areas, Ministry of Natural Resources (Chongqing Institute of Geology and Mineral Resources) Chongqing, 401120, China*
[b] *Chongqing Institute of Geological Environment Monitoring, Chongqing, 401122, China*

ARTICLE INFO

Keywords:
Groundwater nitrate contamination
Machine learning models
GIS
Uncertainty assessment

ABSTRACT

Groundwater resources is not only important essential water resources but also imperative connectors within the intricate framework of the ecological environment. High nitrate concentrations in groundwater can exerting adverse impacts on human health. It is imperative to accurately delineate the distribution characteristics of groundwater nitrate concentrations. Four different machine learning models (Gradient Boosting Regression (GB), Random Forest Regression (RF), Extreme Gradient Boosting Regression (XG) and Adaptive Boosting Regression (AD)) which combine spatial environmental data and different radius contributing area was developed to predict the distribution of nitrate concentration in groundwater. The models use 595 groundwater samples and included topography, remote sensing, hydrogeological and hydrological, climate, nitrate input, and socio-economic predictor. Gradient Boosting Regression model outperforms the other models (R2 = 0.627, MAE = 0.529, RMSE = 0.705, PICP = 0.924 for test dataset) under 500 m radius contributing area. A high-resolution (1 km) groundwater nitrate concentration distribution map reveal in the majority of the study area, groundwater nitrate concentrations are below 1 mg/L and high nitrate concentration (>10 mg/L) proportion in southeast, northeast and central main urban area karst valley regions is 1.89%, 0.91%, and 0.38% respectively. In study area, hydrogeological conditions, soil parameters, nitrogen input factors, and percentage of arable land are among the most influential explanatory factors. This work, serving as the inaugural application of utilizing effective spatial methods for predicting groundwater nitrate concentrations in Chongqing city, furnish decision-making support for the prevention and control of groundwater pollution, particularly in areas primarily dependent on groundwater for water supply and holds profound significance as a milestone achievement.

1. Introduction

The issue of nitrate contamination in groundwater is a common concern shared by numerous agricultural regions across the globe. Nitrate is a naturally occurring form of nitrogen essential for plant growth. However, in recent decades, intensive agricultural activities in rural areas have led to an excessive leaching of nitrate from sources such as animal manure and synthetic fertilizers into

* Corresponding author.
E-mail address: liangyuanyi@cqdky.com (Y. Liang).

https://doi.org/10.1016/j.heliyon.2024.e27867
Received 12 November 2023; Received in revised form 10 February 2024; Accepted 7 March 2024
Available online 13 March 2024

groundwater [1]. This has resulted in elevated nitrate concentrations in groundwater, exceeding acceptable limits. When groundwater serves as a source of drinking water, nitrate nitrogen concentrations exceeding 10 mg/L can have adverse health effects, especially on subpopulations exhibiting higher endogenous nitrosation capacity, leading to conditions such as cancer and reproductive issues [2,3]. The accumulation of nitrates in groundwater poses significant negative impacts on both the ecology and lifetime health risk [4,5]. Numerous factors contribute to the quantity of nitrate that infiltrates groundwater. These factors encompass both present and past land use practices, historical and current nitrogen applications or deposition, soil type, the depth to the groundwater table, and the rate at which groundwater is replenished [6]. The intricate interplay of these variables can have a significant impact on the extent of nitrogen leaching into groundwater.

Groundwater is also an important resource for the city of Chongqing, China, particularly in areas that are remote from freshwater supplies (Northeast and Southeast of Chongqing) [7,8]. Since 1990, rapid population growth and extensive agricultural activities have had a significant impact on both the quality and availability of groundwater, especially in agricultural and urban areas. Furthermore, the region under study features numerous carbonate rock karst areas characterized by the presence of extensive karst channels, sinkholes, and collapses, providing primary pathways for groundwater nitrate contamination. The natural concentration of nitrate in groundwater typically remains relatively low, generally not exceeding 2.5 mg/L [9].

According to data from the Chongqing Municipal Groundwater Survey Report, there are approximately 25,000 private wells distributed throughout Chongqing Municipality. In addition, there are a total of 520 national and municipal groundwater monitoring stations in Chongqing. However, the private wells are not integrated into the water quality monitoring system. Since private wells are not subject to water quality regulations, it is of paramount importance to determine the extent of groundwater pollution in the vast areas where underground monitoring stations have not been established, especially in regions where groundwater serves as the primary water supply infrastructure. In previous annual assessments of groundwater characteristic ions in Chongqing, the methods employed included inverse distance weighting interpolation. However, these approaches did not take into account the influence of groundwater recharge and discharge on the characteristic ions. The above method can lead to significant bias in areas where groundwater monitoring wells are sparsely. In recent decades, a variety of modelling methods have been developed to simulate nitrogen input, turnover, and transport in the hydrosphere at different spatio-temporal scales, and to develop improved management plans. Existing models range from simple empirical conceptual approaches to complex process-based models [10–13].

Large-scale prediction and simulation of contaminant was used more and more as management plans by national and region scale. High spatial resolution map of the actual state of nitrate concentrations in groundwater was generally predicted by traditional geostatistical approaches such as kriging-based methods are widely used to interpolate point measurements [14]. Traditional groundwater assessment approaches such as DRASTIC [10] were extensively used to evaluate the groundwater vulnerability related to the hydrogeological parameters generate a groundwater vulnerability map linked with hydrogeological and land use characteristics with a subjective evaluation. In terms of physical models, they can simulate groundwater and contaminant transport effectively, but their requirements for detailed data and high computational costs make it challenging to replicate on a large scale [13,15]. In summary, these interpolation methods and weighted calculations may not effectively capture the external factors influencing the groundwater environment. Physical models based on groundwater flow fields are challenging to apply in large-scale groundwater simulation studies.

Compared with complex traditional physical models and the lack of information interpolation method, a novel data-driven models combine with circle catchment area which is used in many in regional and national-scale studies on groundwater quality assessment to assess the contamination of groundwater and to map the pollutants in the water, such as Multiple Linear Regression, Random Forest, and other machine learning methods [16–18]. [19] confirmed that the random forest model outperforms other machine learning models in predicting groundwater nitrate concentration in southern region of Spain [20]. develop a hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. Result from the final machine learning models which include twenty-five variables show the higher accuracy compared to ordinary kriging, universal kriging, and multiple linear regression. In another study [21], uses the random forest model which outperforms the other models and the exclusively spatial available predictors to predict the nitrate in groundwater in Hesse. Result from the RF model which identify the hydrogeological units, the percentage of arable land and the N surplus on agricultural used area as the most relevant predictors for estimating the spatial distribution of nitrate concentration in groundwater. In Northern Atlantic Coastal Plain aquifer system, groundwater pH and redox conditions were predicted by boosted regression trees method (BRT) [17]. XGBoost model, combined with 187 predictive indicators, was employed to predict groundwater nitrate concentrations across the United States, and an estimation was conducted for the population at risk of unsafe drinking water [22].

To achieve predictions of nitrate concentrations in Chongqing's groundwater, identify drivers of predicted nitrate and provide comprehensive theoretical support for the management of groundwater resource quality. We use machine learning models that can systematically evaluate the nitrate content of groundwater at large scale in Chongqing especially in areas where groundwater monitoring wells have been not built. In this study, we used different designs of contributing area surrogate which provide a good solution to represent the hydrologic catchment area at large scale to distribute spatial data to groundwater monitoring wells [23–25]. Extreme Gradient Boosting (XG), Random Forest regression(RF), Gradient Boosting (GB), and Adaptive Boosting (AD) algorithms were used to model groundwater nitrate concentration in response to predictor variables representing land use, climate, soil parameters, nitrogen input, and hydrogeologic factors [20,26–29]. Last, XG, RF, GB and AD are used to establish 1 × 1 km resolution maps of groundwater nitrate concentration for Chongqing region. These algorithm models learns the relationship between the response and the predictor variables and does not rely on hypothesis testing assumptions about the data as do more traditional statistical methods [20,22]. Through these efforts, we aim to not only provide accurate spatial predictions of nitrate concentrations but also contribute valuable insights into the complexities of groundwater contamination dynamics in Chongqing.

2. Materials and method

2.1. Study area

The study area is located in Chongqing the city of 82402 km^2 in southwest of China. Chongqing is situated at the transitional area between the Tibetan Plateau and the plain on the middle and lower reaches of the Yangtze River in the sub-tropical climate zone often swept by moist monsoons. The topography of Chongqing is higher in the northeast and southeast than in the west, among which the highest elevation is in the northeast of Chongqing and the lowest elevation is in the west of Chongqing. The mean (2005–2020) annual air temperature is 18 °C and the mean annual precipitation is 1404 mm. Study area and groundwater sampling location were shown in Fig. 1.

Due to intricate geological structures and topographic conditions, spanning both the Yangtze Platform and the Qinling Fold System, the hydrogeological environment within the area is notably complex. Based on a combination of factors including groundwater occurrence conditions, hydraulic characteristics, and aquifer properties, groundwater in the region can be categorized into four major types: carbonate rock fractured-cave water (referred to as karst water), clastic rock porous-fractured water, bedrock fractured water, and unconsolidated rock porous water. The remaining area is characterized by Silurian mudstone, shale, and siltstone, forming a relatively impermeable layer. The respective proportions of these types in terms of the total area are represented by 35.90%, 38.09%, 18.15%, and 0.64%.

2.2. Groundwater nitrate data

Groundwater nitrite-plus-nitrate data were collected from 520 national and provincial groundwater monitoring stations, along with 75 spring samples that were obtained during local hydrogeological investigation work in Chongqing from 2019 to 2022. The sampling frequency is biannual, covering both the dry season and the wet season each year (the sampling locations are indicated in Fig. 1). The inclusion of data from a wide range of national and provincial monitoring stations ensures extensive spatial coverage and a comprehensive view of groundwater quality within Chongqing. Operation of a groundwater monitoring well networks which is constructed by Chongqing planning and Natural Resources Bureau (CQPNRB) starting in 2016. The samples of groundwater monitoring well and spring which is more than once, the average value of the samples data used as the groundwater nitrate data.

The Chongqing Institute of Geological Environment Monitoring (CQIGEM) operates the groundwater monitoring service in Chongqing, monitoring the groundwater quality of all hydrogeological units in the region. The monitoring network comprises of 520 monitoring well stations, and the 75 spring samples were selected from the hydrogeological investigation of the northeast region of Chongqing. Sampling and analysis are carried out by the Chongqing Geotechnical Engineering Testing Laboratory in accordance with

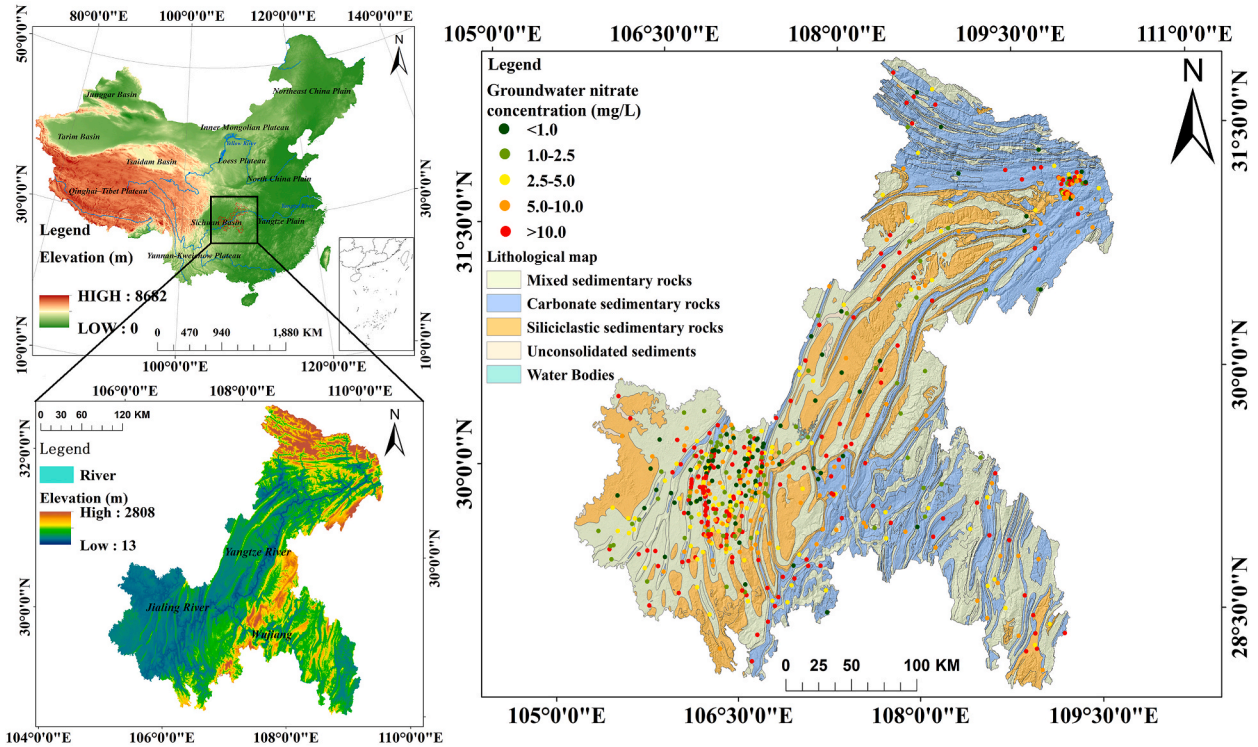


Fig. 1. Geographical location of the study area and distribution map of groundwater sampling points.

GBT 14848-2017. To assess the nitrate concentration in shallow groundwater using machine learning models, a total of 595 samples were selected in accordance with the model's requirements.

2.3. Predictor parameters

Many factors influence on groundwater quality and quantity. They can be classified into five categories: topology, remote sensing, hydrogeological and hydrological data, climate and nitrogen input. For predicting groundwater nitrate content, 34 predictor parameters (Table 1) compose the dataset include monitoring well construction data, Remote sensing data and existing research results of models, including monitoring well location, land cover, soil parameters, NDVI value, estimated value of nitrogen input, accumulate nitrogen input data by land cover change, nitrogen input intensity (e.g. fertilizer,manure, deposition, biological nitrogen fixation) to the land surface, hydrological data, climate and groundwater table. In previous research of predict groundwater nitrate content we found land cover, nitrate input, climate, aquifer and soil properties are the most relevant parameters [1,9,18,21,22,30,31].

2.3.1. Topography

Groundwater tables are often conceptualized as subdued replicas of topography which most strongly control groundwater fluxes [47]. Therefore, we used the global digital surface model with horizontal resolution of approximately 30 m which is produced by Panchromatic Remote-sensing Instrument for Stereo Mapping (PRISM) aboard the Advanced Land Observing Satellite (ALOS) [32] from 2006 to 2011 to extract the topography parameters information. In order to describe the nonlinear relationships between topography and water table depth and control of groundwater flow field by topography, six variables (mean elevation (DEM), standard deviation of elevation (DEM_std), mean slope (Slope), standard deviation of slope (Slope_std), Topographic Position Index (TPI), and standard deviation of TPI (TPI_std)) were calculated from DEM data.

Table 1
Summary of the spatial predictor variables.

Predictor	Variable	Unit	Domain	Resolution	Data type	Data source
Topography						
DEM	DEM	m	Global	30 m	Numercial	[32]
Standard deviation of elevation	DEM_std	m	Global	30 m	Numercial	[32]
Slope	Slope	°	Global	30 m	Numercial	[32]
Standard deviation of slope	Slope_std	°	Global	30 m	Numercial	[32]
Topographic Position Index	TPI	–	Global	30 m	Numercial	[32]
Standard deviation of Topographic Position Index	TPI_std	–	Global	30 m	Numercial	[32]
Remote sensing						
Percentage of Cropland	Cropland	%	Global	10 m	Numercial	[33]
Percentage of Forest	Forest	%	Global	10 m	Numercial	[33]
Percentage of Grassland	Grassland	%	Global	10 m	Numercial	[33]
Percentage of Shrubland	Shrubland	%	Global	10 m	Numercial	[33]
Percentage of Wetland	Wetland	%	Global	10 m	Numercial	[33]
Percentage of Water	Water	%	Global	10 m	Numercial	[33]
Percentage of urban land	Urban land	%	Global	10 m	Numercial	[33]
Percentage of Bareland	Bareland	%	Global	10 m	Numercial	[33]
Mean value of NDVI	NDVI_mean	–	Global	500 m	Numercial	[34]
Maximum value of NDVI	NDVI_max	–	Global	500 m	Numercial	[34]
Standard deviation value of NDVI	NDVI_std	–	Global	500 m	Numercial	[34]
Hydrogeological and hydrological data						
Topsoil K	Soil_K	–	Global	1 km	Numercial	[35]
Soil nitrogen	Soil_Nr	cg/kg	Global	1 km	Numercial	[35]
Content of sand	Sand	%	China	1 km	Numercial	[35]
Content of slit	Slit	%	China	1 km	Numercial	[35]
Content of clay	Clay	%	China	1 km	Numercial	[35]
Lithology	lithology	–	Global	Polygons	Category	[36]
Conductivity of aquifer	Cond_aq	m/d	Local	Polygons	Numercial	[36]
Groundwater Table Depth	WTD	m	Global	1 km	Numercial	[37]
Climate						
Precipitation	PRE	mm	Global	1 km	Numercial	[38]
Potential Evapotranspiration	PET	mm	Global	1 km	Numercial	[39]
Nitrate input						
Fertilizer Input (county level)	F_in	kg/ha/y	Local	Polygons	Numercial	[40]
Nitrogenous Fertilizer Input	Nrf_in	tg/yr	China	5 km	Numercial	[41]
Nitrogen Leakage	Nr_leakage	N/ha	Local	1 km	Numercial	[1]
Nitrogen Cumulative Input	Nr_cum_in	N/ha	Local	1 km	Numercial	[1,42,43]
Distance of Point Source Pollution	PPD	m	Local	250 m	Numercial	[44]
Social economy						
Population Density	POP	pop/km^2	China	1 km	Numercial	[45]
GDP pre 1 km grid	GDP	10000yuan/km^2	China	1 km	Numercial	[46]

2.3.2. Remote sensing data

The remote sensing data include land cover, normalized difference vegetation index. Land cover can have a significant impact on the movement of water and the fate of nitrate in the environment [1,42]. The influence of each land use type on groundwater quality was represented by the percentage of each land use type in the contributing area (Cropland, Forest, Grassland, Shrubland, Wetland, Water, Urban land and Bareland). NDVI time series were include in predictive parameters to provide indications of agroecosystem dynamics [27]. NDVI was provided by The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Vegetation Indices (MOD13Q1) Version 6.1 [34] product with 250 m 16 days pixel reliability. Spanning one growing season, maximum level of photosynthetic activity in the canopy (NDVI_max), standard deviation of NDVI (NDVI_std) and average mean of NDVI (NDVI_mean) were used to indirectly contemplate nitrogen loss from crop removal, and/or nitrogen leaching to groundwater due to nitrogen fertiliser and irrigation management practices [27].

2.3.3. Hydrogeological and hydrological data

Hydrogeological parameters play a crucial role in controlling the movement, distribution, and quality of groundwater. The groundwater table plays a crucial role as a controlling factor in the infiltration process of groundwater pollutants, making it a vital indicator in pollution mitigation. To extract hydrogeological parameters (Topsoil K (Soil_K), Soil nitrogen (Soil_Nr), Content of sand (Sand), Content of slit (Slit) and Content of clay (Clay)) of topsoil, have significant influence on the transition between surface water and unconfined surficial aquifer, we used Regridded Harmonized World Soil Database V. 1.2 [35] which is regridded from HWSD soil dataset that is the most comprehensive and detailed database of soil characteristics, with 1 km pixel. Aquifer (Lithology), a body of porous rock or sediment saturated with groundwater and have direct influence on groundwater flow was provided by new global lithological map database GLiM [36]. We used different permeability empirical value of rocks which is derived from 1:25,0000 scale hydrogeological maps and formation lithology to express the conductivity of aquifer (Cond_aq). Groundwater table depth (WTD) was obtained in Global patterns of groundwater table depth dataset [37]. To sum up, hydrogeological data collection consist of hydraulic conductivity of soil, soil parameters, lithology and groundwater table depth.

2.3.4. Climate data

Climate parameters, such as precipitation (PRE) and evapotranspiration (PET), can have a significant impact on the recharge, discharge, and quality of groundwater. WorldClim is a database of high spatial resolution global weather and climate data with high resolution (approximately 1 km), aggregated across a temporal range of 1970–2000 [38]. The WorldClim Vision2.0 was produced by interpolating observations from 9000 to 60,000 stations and incorporating other covariates. Precipitation data was downloaded from WorldClim Vision2.0. In this approach, we also consider the influence of potential evapotranspiration on water cycle and topsoil salinity. The potential evapotranspiration data (approximately 1 km pixel) which were downloaded from Global Aridity Index and Potential Evapotranspiration Climate Database v3 which is modeled based on the implementation of the FA0-56 Penman-Monteith Reference Evapotranspiration (ET0) equation [39].

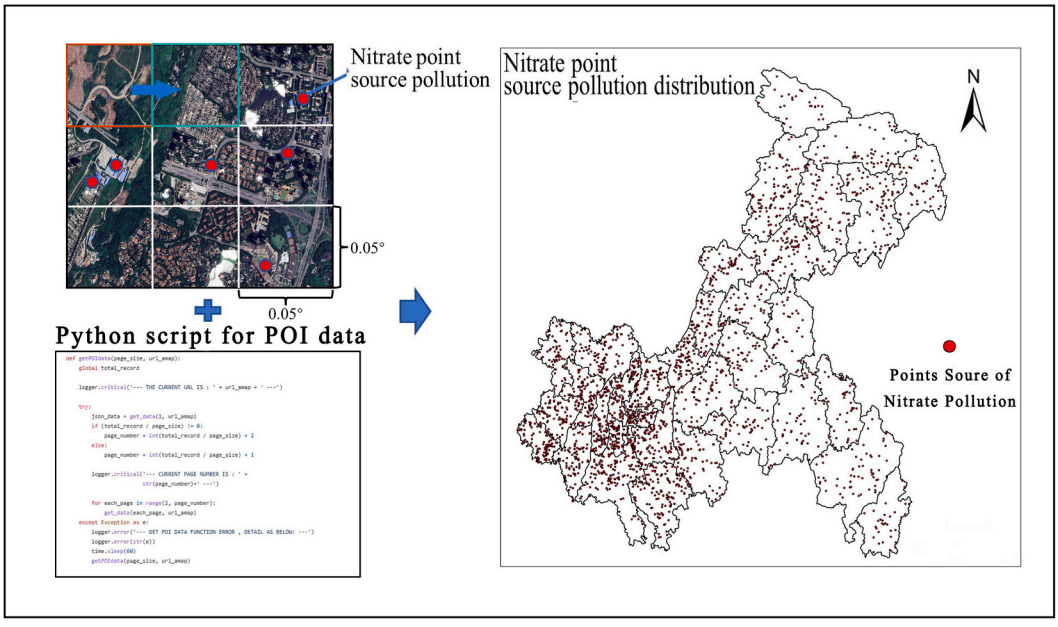


Fig. 2. Study area grid split and Python script code for nitrate point source pollution POI data.

2.3.5. Nitrogen input

The sources of nitrate in groundwater can be categorized as nonpoint and point sources. Land cover data can effective represent the contributing area of nitrate input such as agricultural activity and urbanization. The presence of nitrates in groundwater due to the use of chemical fertilizers, nitrogenous fertilizer input (F_in) and nitrogen leakage such as nitrogenous fertilizer input (Nrf_in), nitrogen leakage (Nr_leakage), and nitrogen cumulative input (Nr_cum_in). The fertilizer input is based on the average data from Chongqing Statistical Yearbook for the years 2011–2018. Nitrogenous fertilizer input data was resampled to 5 km × 5 km resolution, covering the period from 1952 to 2018 by integrating improved cropland maps [41]. To reflect the impact of land use changes on the nitrate concentration in groundwater, we combine 2011–2018 land cover which were provided by Terra and Aqua combined Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type (MCD12Q1) Version 6 data product [43], its six land-use types (Cropland, Forest, Grassland, Water, Built-up area and unused land) nitrogen input estimate data [42] and nitrogen leakage data of year 2017 were compiled by using the CHANS model [1,42]. However, it cannot consider the timeliness of the point source of nitrate data. Therefore, we use web service database which provide frequently updated information on businesses and points of interest can be applied on larger scales and data scarce regions where data availability is limited [48]. For this study, the place web service API which is launched by Gaode Map [44] was used to get the location of point data of nitrate emission (poi of agriculture, forestry, animal, and husbandry) and thereby obtain the point source pollution distance predictor (PPD). To get the location data more efficient, we use python script which can search the different type point of interesting location in 0.05° polygon region (Fig. 2).

2.3.6. Socio-economic predictor

The escalation of groundwater pollution presents a multifaceted challenge shaped by the intricate interplay of population growth and socio-economic influences. The increased water access for domestic, agricultural, and industrial use can lead to over-extraction, lowering water tables and facilitating the infiltration of contaminants. Strained water resources can intensify the risk of pollution, particularly in regions with inadequate infrastructure to manage increased water demand effectively [49]. Urban expansion and intensified industrial and agricultural production have exacerbated groundwater pollution [42]. Therefore, we used kilometer grid population data (POP) and GDP data (GDP) as input indicators for population and socio-economic factors.

2.4. Statistical model framework

In this study, we developed four nolinear regression models including random forest regression (RF) [50], extreme Gradient Boosting regression (XG) [51], Adaptive Boosting regression (AD) [24], Gradient Boosting regression (GB) [52] for prediction of groundwater nitrate concentration. Ensemble learning algorithm is combination of multiple base models. Compared to single algorithm, ensemble learning algorithm have observable improvement and also apply in many discipline and area with advantages over traditional algorithms [29,53,54].

2.4.1. Random forest regression (RF)

A random forest algorithm [55] as a kind of predictive model base on classification and regression is a collection of randomized and independent decision trees. Bootstrap sampling is used to generate different subsets and decision trees, and then averaging the predictions of the individual trees to make a final prediction in regression. This machine learning algorithm improves the model's predictive performance by variance reduction and results in a more robust prediction. Random forests which is widely apply in many areas [37,54,56,57] known for their good performance and ability to handle high dimensional data and measure variable importance.

2.4.2. Extreme Gradient Boosting Regression (XG)

Extreme Gradient Boosting (XG), is a highly effective machine learning algorithm that has gained widespread adoption in both academia and industry [22,58,59]. Meanwhile, XGBoost is an implementation of gradient boosting that is designed to be efficient, scalable, and flexible. One of the key advantages of XGBoost is its ability to handle missing values and imbalanced data. In addition, it offers a range of hyperparameters that can be fine-tuned to further enhance model performance. The XGB method is considered an optimized form of gradient boosting due to two important distinctions: the use of a regularized loss function which penalizes the complexity of the model by setting an error reduction threshold and column subsampling of the predictor variables to introduce randomness, both of which aid to avoid overfitting [51].

2.4.3. Gradient Boosting Regression (GB)

Gradient Boosting Regression (GB) [60] is a sophisticated machine learning algorithm which is also base on the CART extensively employed in regression tasks. Distinguishing from RF, GB is trained by minimizing the residuals of the preceding decision trees [61]. Belonging to the ensemble learning family, it represents an extension of the renowned Gradient Boosting Machine (GB) algorithm, designed to capitalize on the collective strengths of weak learners in order to construct a robust predictive model.

2.4.4. Adaptive Boosting Regression (AD)

AdaBoost, also known as Adaptive Boosting, is a highly influential ensemble learning algorithm widely utilized in both classification and regression tasks. It was originally proposed by Yoav Freund and Robert Schapire in 1996 [24], representing a seminal advancement in the field of machine learning. The algorithm's primary objective is to construct a robust and accurate ensemble model by combining the predictions of multiple weak learners, such as decision trees with limited depth or stumps. This adaptively changing weighting mechanism ensures that subsequent weak learners focus on learning from the challenging data points, effectively improving

their accuracy on these instances in the following iterations.

2.4.5. Hyperparameter tuning method

Machine learning algorithms have some a set of hyperparameters that affect the generalization ability of the models generally. Inappropriate hyperparameters may lead to underfitting or overfitting predictions. Random search [62], Grid search [63], Bayesian optimization [64], etc. were proposed in past approaches to get optimal model hyperparameters for algorithms. In this study, grid search and random search were used to optimize the hyperparameters of Random Forest Regression (RF), XGBoost Regression (XG), Adaptive Boosting Regression (AD), Gradient Boosting Regression (GB) respectively. In optimization process, R2 (coefficient of determination), MAE (mean absolute error), RMSE (root mean square error) was used to evaluate the regression prediction performance. Briefly, R2 measures the proportion of the variance in observed data explained by the model; R2 = 1 indicates a perfect fit between observed and estimated values. MAE is the average absolute difference between observations and model predictions. RMSE is the standard deviation of the differences between observed and estimated values.

In order to check generalization, guarding against overfitting and evaluate the performance of a model better. We use 5-Fold Cross-Validation and four statistical methods to assess the difference between predicted and observed values for evaluating the performance of above models. Coefficient of determination (R^2), mean absolute error (MAE) and mean-square-error (RMSE) were used to express the average model prediction error in the unit of natural logarithm of NO_3–N concentration.

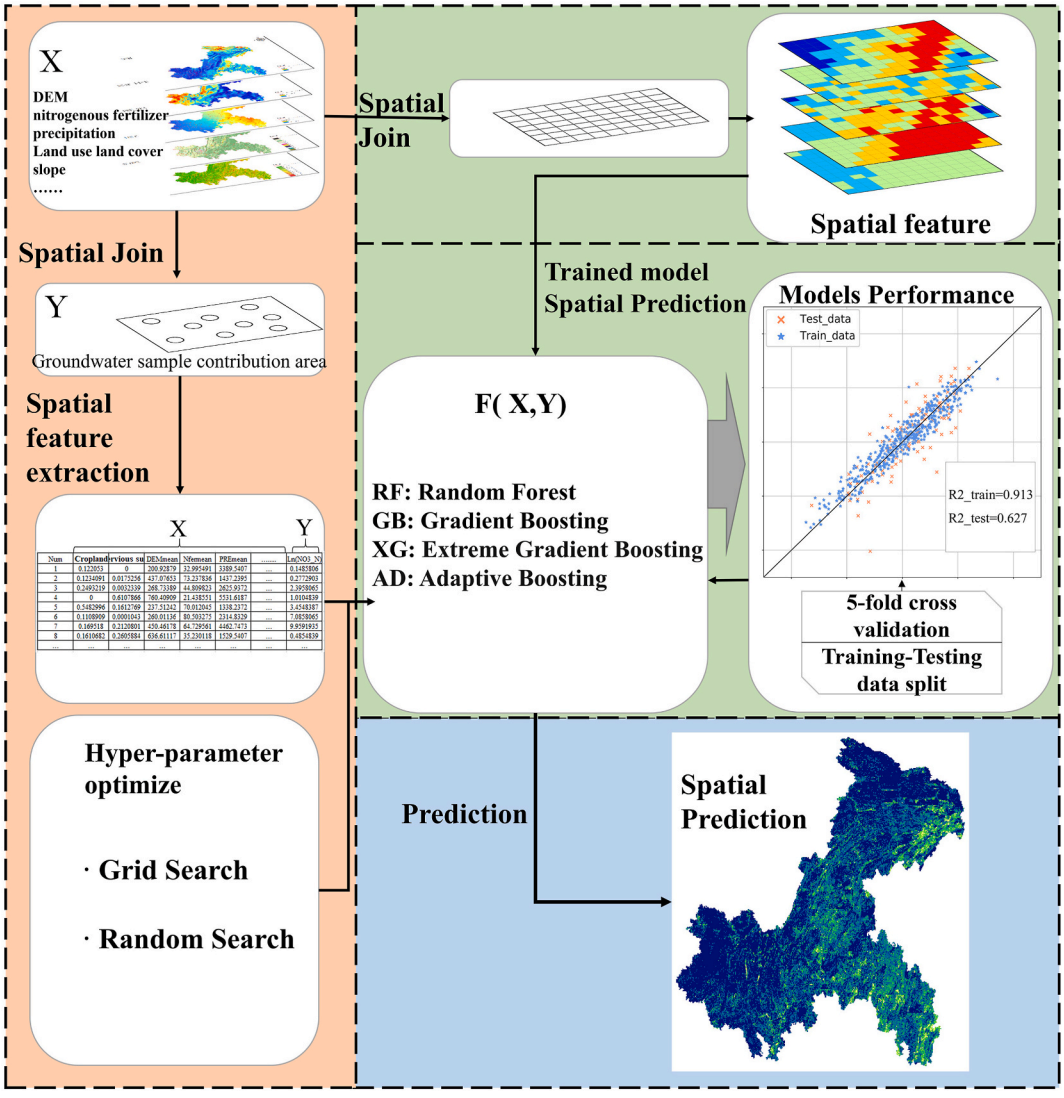


Fig. 3. The workflow for predicting groundwater nitrate concentration based on GIS and machine learning models includes data preprocessing, model evaluation, and spatial prediction.

2.5. Data pre-processing

To assess the nitrate concentration of groundwater in poor flow condition on large scale, the catchment area which is proposed to simply replace physical hydrological model and showed a good performance. Circular buffer area of monitoring well can represent the catchment area that called contributing area [23,25].Before confirming the radius of the contributing area, the optimal distance between the location of monitoring well and the boundary of contributing area must be calculated. Therefore, we define the four different radius (500 m, 1000 m, 1500 m and 2000 m) of contributing area under the condition that the influence of land cover and other parameters of sampling location surrounding area is homogenous.

Predictors have two different data types: numerical and categorical. For this study, we adopt the percentage of different subtype and one hot encode method to quantize the categorical spatial data. Therefore, for raster data, we use mean value to statistics the pixels in the buffer scope. In this approach, we use QGIS Desktop [65] and QGIS Python console to process above statistical work.

To predict groundwater nitrate concentration in total study area. We create 1 km grid for study area and use the above method to perform the corresponding join processes with all predictors (Fig. 3).

2.6. Bias correction and uncertainty assessment

2.6.1. Bias correction

Ensemble-tree machine learning regression models can effectively fit nonlinear functions and generally unbiased in the sense that the sum of the errors (observed values compared to estimated) is close to zero [66]. However, the estimate value of ML models can underestimate large values and overestimate small values. In this study, Empirical distribution matching (EDM) method which was used to correct the bias between observed values and estimated values involves transforming one dataset to match the empirical distribution of another dataset. The EDM model offers improved bias correction between predicted and observed values. We employed the EDM model to perform bias correction on the predictions of both the training and validation datasets. Subsequently, the well-fitted EDM model was applied to perform bias correction on the predictions of the testing dataset. This approach ensured that the predicted values across all datasets were adjusted appropriately to account for any systematic biases, enhancing the accuracy and reliability of the model's predictions.

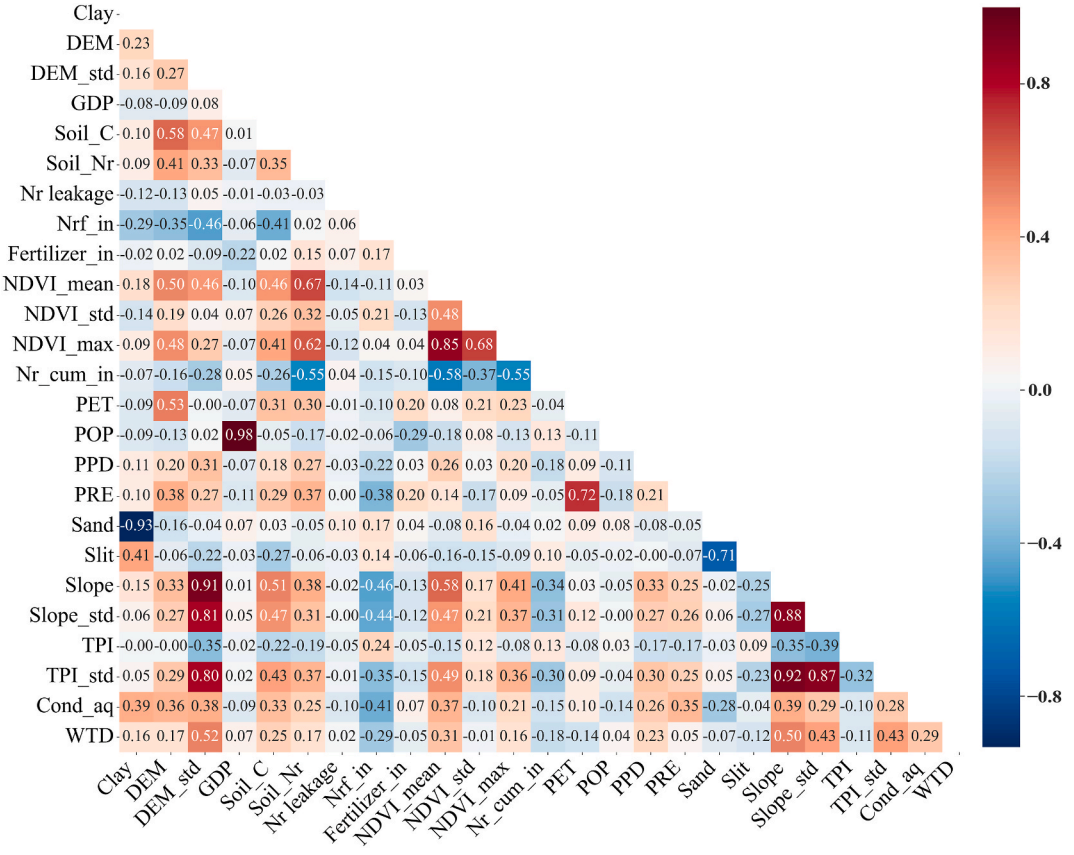


Fig. 4. Pearson correlation matrix of independent variables (categories and land use-related indicators have been removed). Abbreviations for all variables are the same as those in Table 1.

2.6.2. Uncertainty assessment

Quantile regression (QR) was used to evaluate the predictive uncertainty of the models. Unlike classic methods, such as Monte Carlo-based approaches, which typically focus on a single source of uncertainty, these methods assess model residuals and take into account all potential sources of uncertainty [18,67,68].

The quantile regression (QR) was developed to assess the distributional error. To evaluate uncertainty using QR, each individual machine learning (ML) method was trained with the input-output pairs of the training dataset. Subsequently, the models were utilized to predict nitrate concentration values for all samples in the study area. For each desired quantile (e.g., 0.05 and 0.95), the predicted nitrate values from each ML model were used as input, with observed values as output, to calibrate the QR model. Finally, leveraging the predicted nitrate values from each ML model, the calibrated QR model was applied to compute the expected quantiles (0.05 and 0.95) of nitrate values, serving as inputs for the entire study area.To quantify the uncertainty of the models, the Mean Prediction Interval (MPI) (Eq. (1)) and Prediction Interval Coverage Probability (PICP) metrics (Eq. (2)) were used [69]. MPI, which represents the average width of the prediction interval, and PICP, indicating the probability of observed values falling within the prediction interval, are two essential metrics for assessing uncertainty. PICP is particularly crucial as it signifies the number of observations falling within the estimated interval, providing a measure of prediction accuracy. Meanwhile, MPI serves as a complementary measure, suggesting that among models with similar PICP values, the one with a lower MPI is considered a better-performing model.

$$MPI = \frac{1}{n}\sum_{t=1}^{n}\left(P_t^{upper} - P_t^{lower}\right) \tag{1}$$

$$PICP = \frac{1}{n}\sum_{t=1}^{n} C, C = \begin{cases} 1, if\ (P_t^{upper} < y_t < P_t^{lower}) \\ 0, else \end{cases} \tag{2}$$

3. Results and discussion

3.1. Independent variables correlation analysis

In this research, 34 predictive variables which include topography, climate, land cover, soil fraction and nitrogen input source were collected to predict the nitrate concentration of groundwater. Independent variables correlation analysis can identify the most important features for predicting the target variable and can be useful for feature selection. Pearson correlation coefficient of each two independent variables is shown in Fig. 4 (to enhance clarity, categories and land use-related indicators have been removed). As the result shown, the climate variables are mostly correlated elevation, the population density per square kilometer is positively correlated with the GDP. There is a high correlation between DEM data, its derived data (such as slope, regional elevation variance, and regional slope variance), and soil permeability. For soil parameters, sand, silt, and clay in soil texture also exhibit high correlation.

3.2. Model performance and uncertainty assessment

The models' performance was assessed using R2, MEA, and RMSE values for each model. The dataset was split into training and testing sets at a 5:1 ratio. Fig. 5 illustrates the distribution of target values in both sets. In this study, four distinct buffer radii (500 m, 1000 m, 1500 m, 2000 m) were utilized as contributing areas for monitoring wells. These buffer radii, coupled with 5-fold cross-validation, were applied across all statistical models during the hyperparameter tuning process to predict groundwater nitrate concentrations. Fig. 6 presents the outcomes of various buffer radii and models. Models based on circular buffers with a radius of 500 m exhibited higher R2 and lower RMSE and MAE compared to other buffer zones. Notably, RF and GB models using a 500 m buffer area

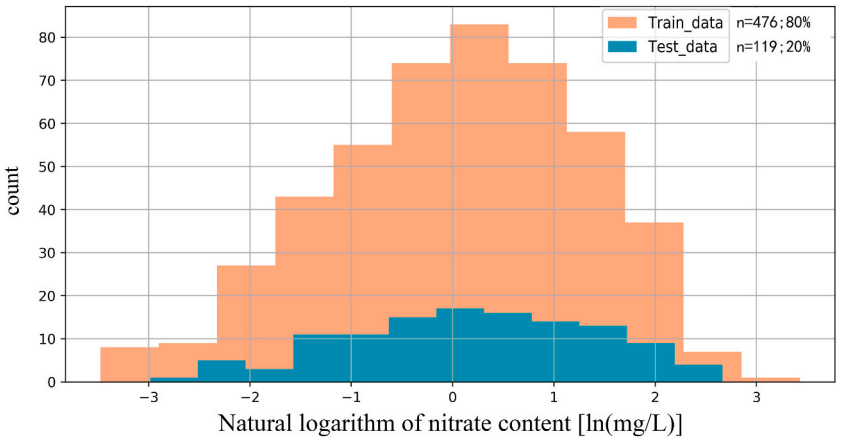


Fig. 5. The distribution of natural logarithm of groundwater nitrate content (ln(mg/l)) (mean 2019–2022) in training and testing datasets.

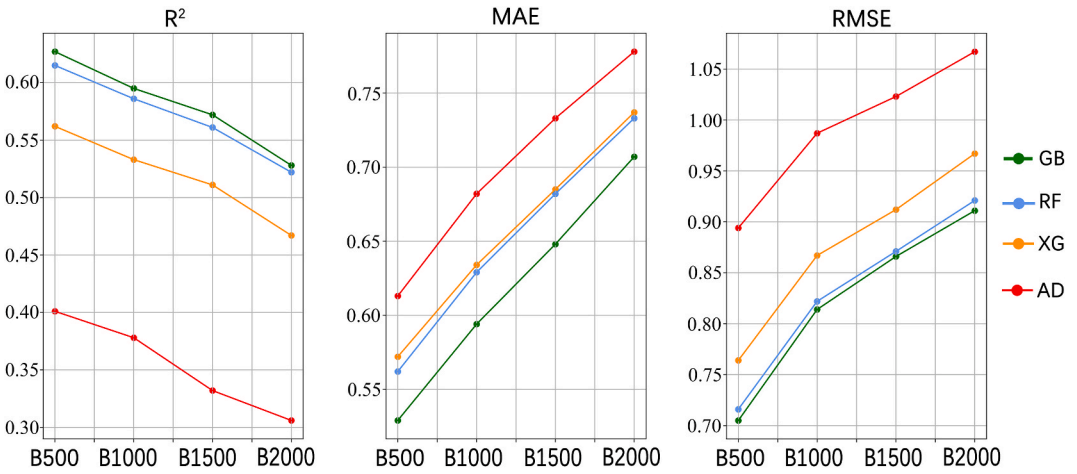


Fig. 6. The test dataset performance R^2 (coefficient of determination), MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) results from 5 fold cross-validation for GB (Gradient Boosting Regression), RF (Random Forest Regression), XG (Extreme Gradient Boosting) and AD (Adaptive Boosting Regression) based on circular buffer (500 m, 1000 m, 1500 m, 2000 m) respectively.

demonstrated superior performance compared to AD and XG models, achieving R2 values exceeding 0.88 and 0.6 on the training and testing sets, respectively. Moreover, MSE and MAE values were lower than those of the AD and XG models, suggesting that the buffer zone effectively represents the catchment area of groundwater monitoring wells, effectively extracting and substituting actual catchment area process information.

The modeling results provided are based on a 500-m circular buffer zone, encompassing model testing and bias correction using the EDM-correction method (Fig. 7). The GB model exhibited remarkable performance, attaining the highest average R2 value of 0.627, coupled with the lowest MAE (0.529) and RMSE (0.705). The overall best-performing model even achieved an R2 of up to 0.68. RF demonstrated only slightly inferior performance with an R2 of 0.615, while AD and XG exhibited lower predictive capabilities with R2 values of 0.401 and 0.562, respectively. This consistent pattern of performance similarity between GB and RF, outperforming AD and XG, was evident across all estimated target functions. The predicted nitrate concentrations were plotted against the observed nitrate concentrations in Fig. 8. The predictive performance values for the training dataset exhibited significantly better results than those for all models on the testing data. The difference in predictive performance metrics between the training (GB: R2 = 0.913) and testing data

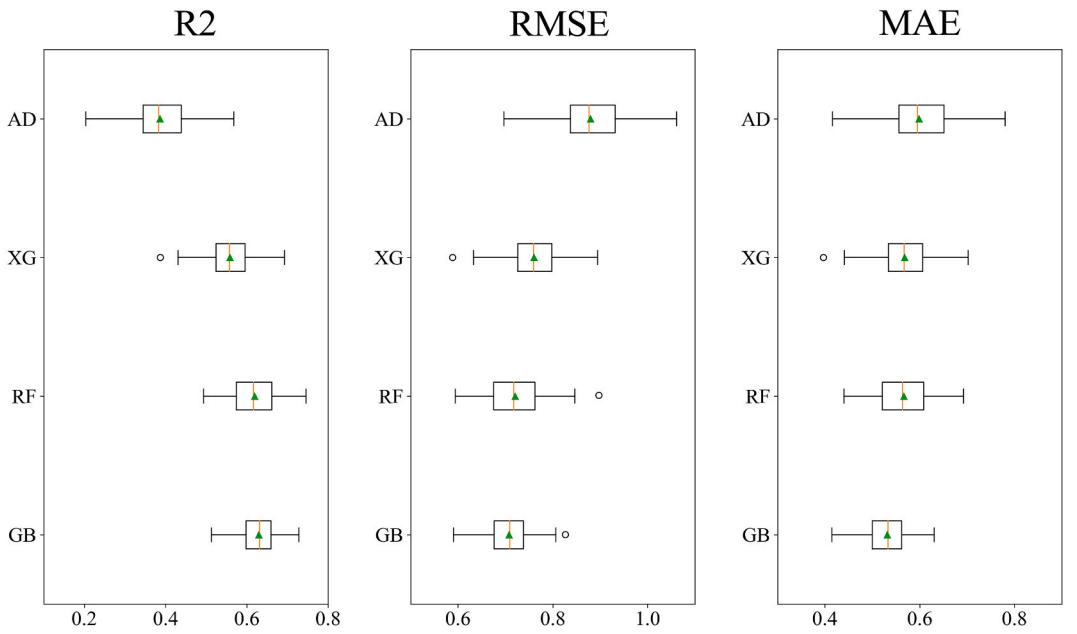
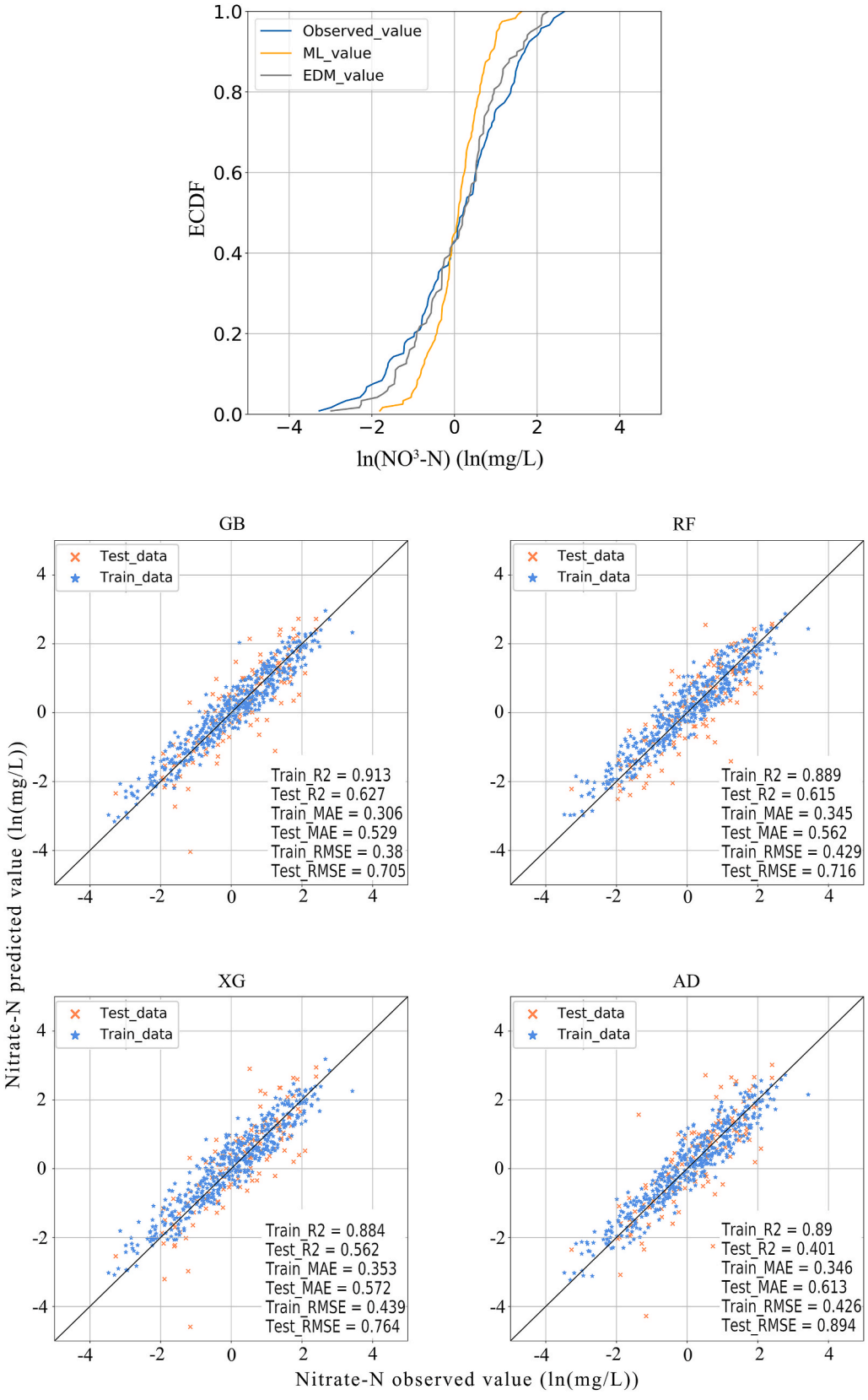


Fig. 7. Comparison of predictive performance R2, RMSE, MAE of the 4 models resulting from the 5 times repeated 5-fold cross-validation for GB (Gradient Boosting Regression), RF (Random Forest), XG (Extreme Gradient Boosting Regression) and AD (Adaptive Boosting Regression), based on the 500 m circular buffer zone, respectively.



(caption on next page)

Fig. 8. Empirical cumulative distribution functions of observed nitrate values for test data (Observed_value), machine learning (ML_value) estimates, Bias correction using empirical distribution matching (EDM_value). Predicted and observe nitrate concentration of groundwater in training dataset and testing dataset after Empirical distribution matching (EDM) method bias correcting for GB (Gradient Boosting Regression), RF (Random Forest), XG (Extreme Gradient Boosting), AD (Adaptive Boosting) in 500 m buffer contribution area.

(GB: R2 = 0.627) suggests that certain statistical ensemble methods, such as GB and RF, as well as XG, may tend to overfit the data because of small size of samples dataset, leading to overly optimistic model performance. Predictive performance measured by cross-validation or testing datasets reflects more realistic predictive performance values.

To comprehensively assess the performance of the models, we incorporated uncertainty assessment by conducting quantile regression. In this study, we calculated the MPI and PICP values for the contributing area of the optimal radius (500 m) for each ML model (Table 2). The results indicated that the GB and RF models had the lowest PICP uncertainty, with values of 0.924 and 0.902, respectively. The XG and AD models exhibited lower PICP values at 0.864 and 0.892, respectively. Compared to the XG and AD models, the GB and RF models demonstrated PICP values closer to the 90% confidence level. Since the PICP values of the models are not equal, the MPI measure may not be considered when assessing the determinacy of the models. Based on the quantified indicators of model performance and uncertainty, the GB model outperformed the other four machine learning models, demonstrating the best model performance (higher value of R2 and lower value of RMSE and MAE) and the lowest uncertainty (quantile regression analysis).

3.3. Variable importance analysis

Statistical models used for predicting groundwater nitrate concentrations rely on 34 spatial predictive variables. Variable importance rank was shown in Fig. 9. Conductivity of aquifer, lithology, percentage of arable land, precipitation and nitrogen leakage are top 5 most important predictors in GB. For RF, percentage of arable land, nitrogen leakage, conductivity of aquifer, max value of NDVI and distance of point source pollution are the five most influential predictors. The AD method showed conductivity of aquifer, lithology, max value of NDVI, soil texture and nitrogen leakage. The ranking for XG indicates that conductivity of aquifer, max value of NDVI, distance of point source pollution, precipitation and percentage of arable land.

In conclusion, the variable importance analysis of the different models reveals that hydrogeological conditions, soil parameters, nitrogen input factors, and land use predictors, commonly examined in previous studies [20–22,26,42,70], are significant in predicting groundwater nitrate concentrations. It is worth noting that the availability and relevance of spatial predictors can vary across different regions, leading to variations in the ranking of predictor variables. Therefore, when determining the most important predictors, these regional differences should be taken into consideration. However, it is crucial to evaluate the overall performance of the models rather than relying solely on the rankings within each model. The models should be assessed based on their ability to accurately predict groundwater nitrate concentrations and their suitability for specific study areas. By considering the comprehensive performance and ranking results, a more robust understanding of the factors influencing nitrate pollution can be achieved, facilitating effective management and mitigation strategies.

3.4. Spatial distribution of groundwater nitrate concentrations

The spatial distribution of groundwater nitrate concentrations in Chongqing was analyzed based on the predictions derived from each of the four models (Fig. 10). Overall, the results indicated a general similarity in the spatial distribution patterns among the models, with only localized variations observed in specific areas. The high nitrate concentration areas were found primarily in the southeastern and northeastern regions of Chongqing, while the central urban and western areas exhibited lower nitrate concentrations. In addition, elevated nitrate concentrations were identified in the mountainous regions of the central urban area, displaying a north-to-south distribution pattern along karst valley which carries industrial, agricultural, and urban domestic wastewater discharge activities. Based on land use, hydrogeological conditions, and nitrogen input indicators, it can be inferred that groundwater with elevated nitrate levels is primarily concentrated in agricultural areas, urban peripheries, and carbonate rock high-permeability zones.

In accordance with the administrative divisions and geographical features of the study area, we partitioned the region into the main urban area (MUA_CQ), northeast region (NE_CQ), and southeast region (SE_CQ). Subsequently, we conducted a comprehensive statistical analysis of nitrate concentration intervals based on the predictions of the four machine learning models for groundwater nitrate

Table 2
Uncertainty quantization results using the QR method.

Model	Uncertainty quantization	Train	Test
GB	PICP	0.933	0.924
	MPI	2.333	2.292
RF	PICP	0.912	0.902
	MPI	2.613	2.534
XG	PICP	0.887	0.864
	MPI	2.172	2.3
AD	PICP	0.903	0.892
	MPI	2.64	2.727

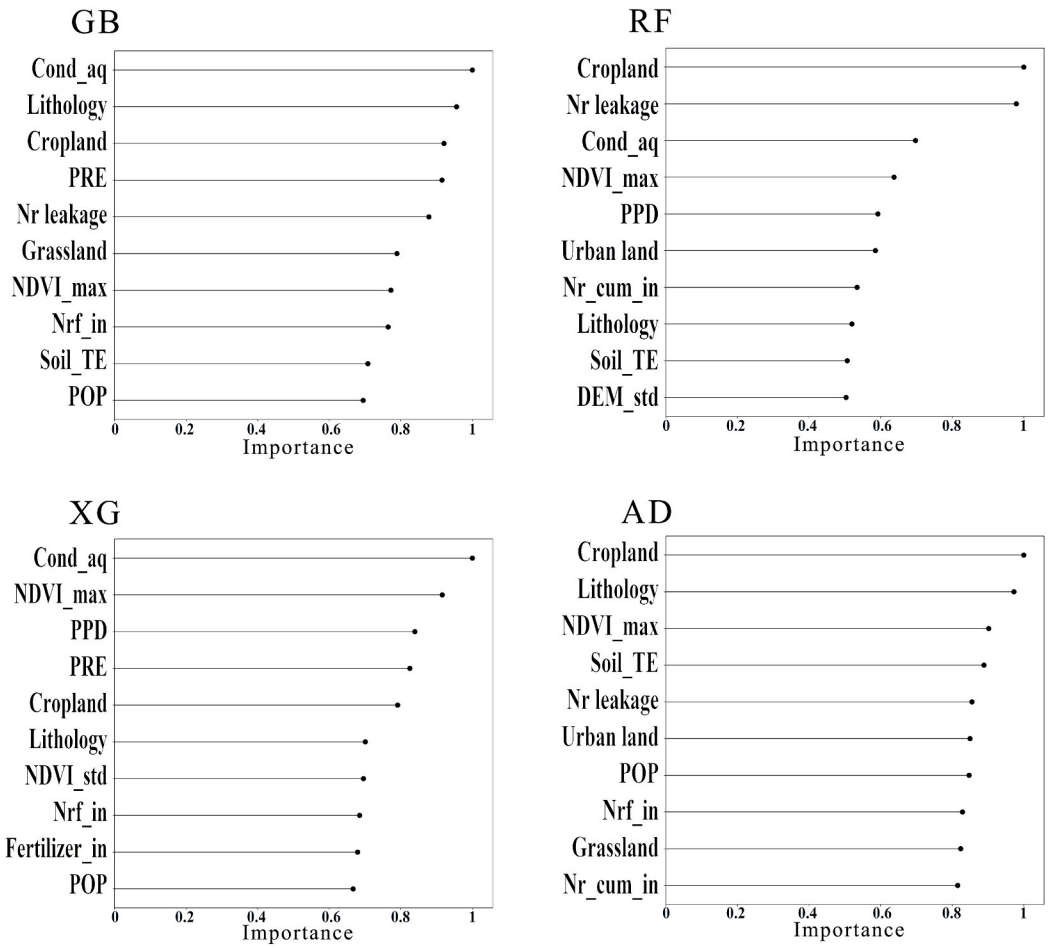


Fig. 9. Importance of Predictive Indicators for prediction of nitrate concentration in Chongqing city by Gradient Boosting Regression (GB), Random Forest Regression (RF), XGBoost Regression (XG), and AdaBoost Regression (AD) Models.

concentration across total areas (Fig. 11). In the assessment of groundwater nitrate concentration results across different models (GB, RF, XG, AD), a notable pattern emerges. Within both the main urban area and the northeastern region, extensive zones characterized by low nitrate concentrations cover more than 50% of the total area. As nitrate concentrations increase, there is a gradual reduction in the extent of these zones. In contrast, areas with high nitrate nitrogen concentration (>10 mg/L) represent a considerably smaller proportion, consistently accounting for less than 1% of the total area. In the southeastern region, the area with nitrate nitrogen concentrations ranging from 1 to 2.5 mg/L is the most extensive, covering approximately 46.28%–53.16% of the total area. Conversely, for other concentration ranges, the distribution areas of nitrate nitrogen decrease as the concentration increases.

Moreover, a positive correlation was observed between the predicted groundwater nitrate concentrations and the distribution of karst areas in Chongqing (Fig. 12). Through the overlay analysis of hydrogeological maps and the predicted results of groundwater nitrate nitrogen, it can be observed that groundwater with high nitrate concentrations (>10 mg/L) is predominantly distributed within carbonate rock formations, accounting for 57.2% of the total distribution. This proportion exceeds that of mixed sedimentary rock and siliceous sedimentary rock formations. The models‘ predicted results align with the groundwater vulnerability assessment in Chongqing [71]. Karst areas are characterized by soluble rock formations, such as limestone or dolomite, which facilitate rapid water infiltration and the development of underground channels and cavities. These geological features enhance the vulnerability of groundwater to nitrate contamination. The association between nitrate concentrations and karst areas suggests that the presence of karst formations in Chongqing contributes to the elevated groundwater vulnerability levels observed in certain regions. Compared to non-carbonate rock regions, the karst fractures and conduits within carbonate aquifers enhance the potential risk of groundwater contamination by facilitating focused groundwater recharge [72]. Simultaneously, the compounding effects of agricultural activities, urban development, and industrial operations have significantly intensified the deterioration of the groundwater environment [42]. These findings underscore the importance of incorporating spatial predictors and understanding the influence of local geological factors for effective assessment and management of groundwater nitrate contamination in Chongqing.

This model can be utilized by groundwater regulatory authorities to formulate groundwater management plans in accordance with

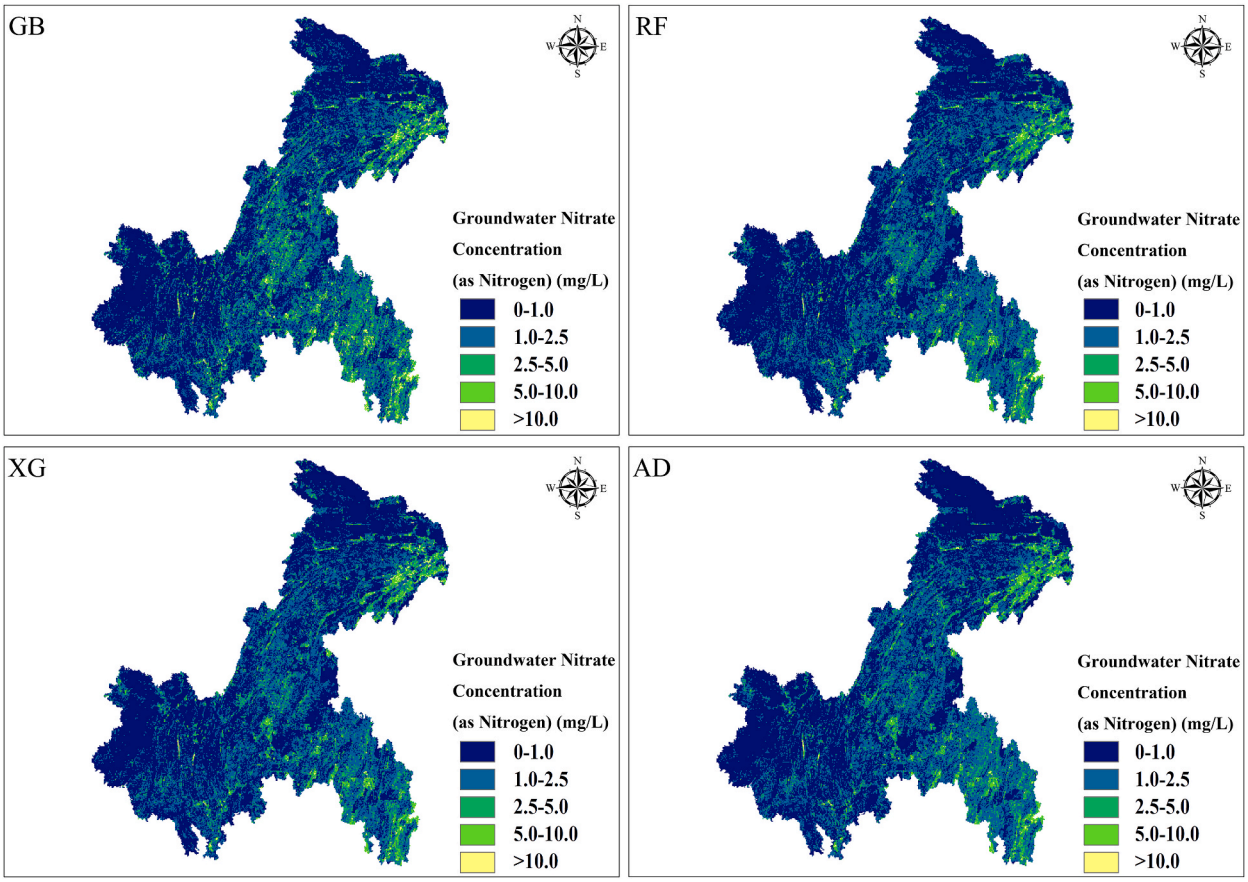


Fig. 10. The spatial distribution of groundwater nitrate concentrations in the Chongqing region predicted by the Random Forest (RF), XGBoost (XG), AdaBoost (AD), and Gradient Boosting (GB) respectively with a resolution of 1 km × 1 km grid cell.

the Chongqing Groundwater Management Regulations. These regulations mandate the management and utilization of groundwater without significantly compromising its quality. The model can extend spatial monitoring results by strategically sampling the most vulnerable areas, providing cost-effective information for monitoring schemes. This method involves optimizing groundwater monitoring stations and sample collection areas, focusing monitoring efforts on areas with high nitrate concentrations as key groundwater environmental monitoring targets. This is particularly relevant in regions where groundwater serves as the primary water supply, enabling cost-effective groundwater quality monitoring. The approach establishes a theoretical foundation for groundwater management decisions in areas heavily reliant on groundwater for water supply.

4. Conclusion

Groundwater resources is a significant position as a vital strategic asset and serve as a crucial link within the ecological environment. Consequently, accurately assessing the concentration of pollutants in groundwater assumes paramount importance, encompassing both effective measures against groundwater pollution and the formulation of robust strategies for its prevention and control. Four machine learning method (GB, RF, XG and AD) were used to assess the nitrate concentration of groundwater with multi-source spatial data across the whole area of Chongqing. Four machine learning models accurately predict the groundwater nitrate concentration. However, GB model achieved best predictive performance (highest training data $R^2 = 0.913$, MAE = 0.306, RMSE = 0.38, PICP = 0.933, testing data $R^2 = 0.627$, MAE = 0.529, RMSE = 0.705, PICP = 0.924) a circular buffer with 500 m radius.

This study reveals the distribution patterns of groundwater nitrate concentrations in the regional scope of Chongqing city. Additionally, we identified elevated nitrate concentrations in the karst areas of Chongqing, including the northeast, southeast, and karst river valley regions, a finding not previously discovered in regional groundwater surveys and assessments. These areas have scarce surface water resources, making them crucial for rural populations heavily reliant on groundwater wells for domestic water supply. This dependence may contribute to regional health concerns, such as an increased risk of cancer and reproductive. Although the proportion of areas with groundwater nitrate concentrations exceeding 10 mg/L is relatively limited, future work can leverage high-resolution (1 km grid) maps of groundwater nitrate concentrations overlaid with population density grids to unveil regions at risk for groundwater drinking water safety. This study underscores the utility of machine learning as a effective tool for mapping groundwater

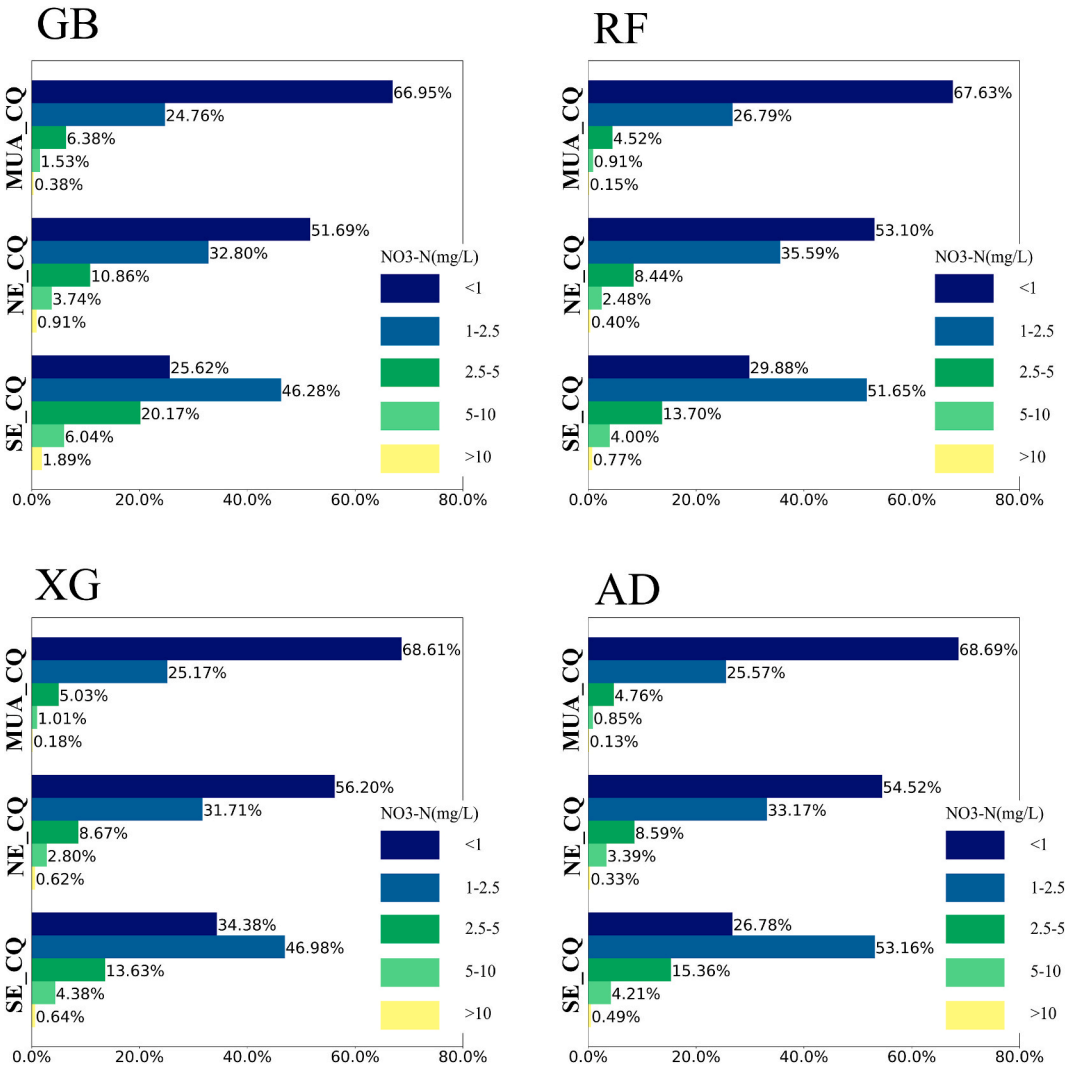


Fig. 11. Depicts the spatial statistics of predicted groundwater nitrate concentrations in the main urban area (MUA_CQ), northeast region (NE_CQ), and southeast region (SE_CQ). The predictions were generated using the Random Forest (RF), XGBoost (XG), AdaBoost (AD), and Gradient Boosting (GB) models, and are presented at a resolution of 1 km × 1 km grid cells.

quality on a large regional scale.

However chemical environmental variables related to groundwater were not considered in this study. During the groundwater transport process, different redox conditions have a crucial impact on the nitrite and nitrate ions in groundwater. However, in large-scale studies, due to data limitations and high computational costs, simulating the chemical processes of groundwater becomes challenging. Therefore, in future work involving the simulation of groundwater pollutant concentrations at distributed small-scale, consideration can be given to incorporating information on groundwater chemical processes.

In our study, using only spatial data as predictive indicators, we efficiently predicted groundwater nitrate at the municipal level through the use of GIS and machine learning methods. This represents a significant advancement in the evaluation of groundwater quality in Chongqing. The outcomes of this research can serve as valuable data references for areas lacking groundwater monitoring.

Funding source disclosure

The research presented in this paper was funded by the investment theme of Chongqing Science And Technology Development Foundation and Chongqing Institute of Geology and Mineral Resources (NO.: cstc2020jcyj-msxmX1074)

Data availability statement

According to the regulations of the groundwater quality data authority, the data that has been used is confidential.

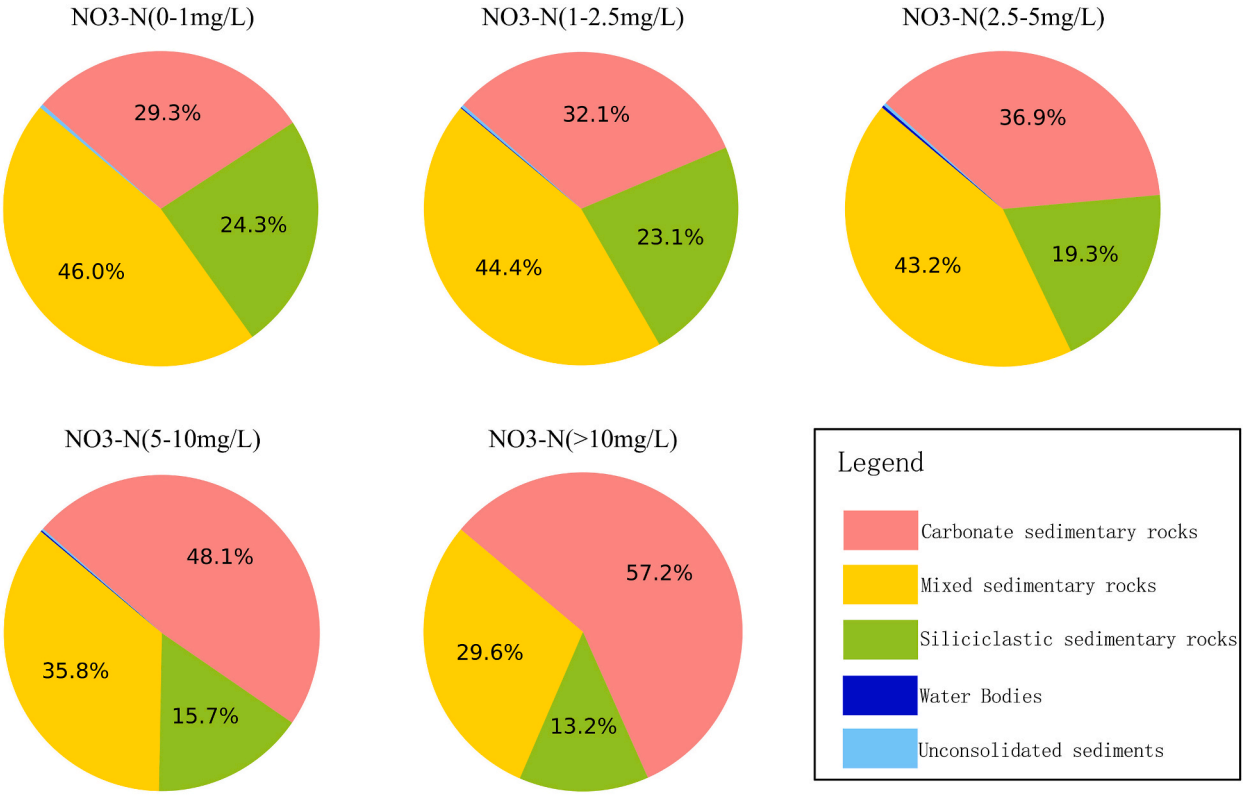


Fig. 12. Statistical graphs depicting the distribution of aquifer lithology across various nitrate nitrogen concentration intervals.

Additional information

No additional information is available for this paper.

CRediT authorship contribution statement

Yuanyi Liang: Writing – original draft, Methodology. **Xingjun Zhang:** Project administration. **Lin Gan:** Funding acquisition, Data curation. **Si Chen:** Funding acquisition, Data curation. **Shandao Zhao:** Supervision, Methodology. **Jihui Ding:** Formal analysis, Data curation. **Wulue Kang:** Validation, Software. **Han Yang:** Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgement

First and foremost, we extend our gratitude to the Chongqing Institute of Geological Environment Monitoring and Chongqing Ecology and Environment Bureau for providing water quality analysis data from groundwater monitoring stations. We also acknowledge the Chongqing Institute of Geology and Mineral Resources for supplying hydrogeological survey data and spring water quality analysis data. Additionally, we express our appreciation for the financial support received from the Chongqing Science And Technology Development Foundation (cstc2020jcyj-msxmX1074) and the self-funded resources of the Chongqing Institute of Geology and Mineral Resources.

References

[1] B. Gu, Y. Ge, S.X. Chang, W. Luo, J. Chang, Nitrate in groundwater of China: sources and driving forces, Global Environ. Change 23 (5) (2013) 1112–1121, https://doi.org/10.1016/j.gloenvcha.2013.05.004.

[2] M.H. Ward, R.R. Jones, J.D. Brender, T.M. De Kok, P.J. Weyer, B.T. Nolan, et al., Drinking water nitrate and human health: an updated review, Int. J. Environ. Res. Publ. Health 15 (7) (2018) 1557, https://doi.org/10.3390/ijerph15071557.

[3] Y.H. Loh, P. Jakszyn, R.N. Luben, A.A. Mulligan, P.N. Mitrou, K.-T. Khaw, N-Nitroso compounds and cancer incidence: the European prospective investigation into cancer and nutrition (EPIC)–Norfolk study, Am. J. Clin. Nutr. 93 (5) (2011) 1053–1061, https://doi.org/10.3945/ajcn.111.012377.
[4] H. Topaldemir, B. Taş, B. Yüksel, F. Ustaoğlu, Potentially hazardous elements in sediments and Ceratophyllum demersum: an ecotoxicological risk assessment in Miliç Wetland, Samsun, Türkiye, Environ. Sci. Pollut. Control Ser. 30 (10) (2023) 26397–26416, https://doi.org/10.1007/s11356-022-23937-2.
[5] B. Yüksel, F. Ustaoğlu, E. Arica, Impacts of a garbage disposal facility on the water quality of çavuşlu stream in Giresun, Turkey: a health risk assessment study by a validated ICP-MS assay, Aquatic sciences engineering 36 (4) (2021) 181–192, https://doi.org/10.26650/ASE2020845246.
[6] L.W. Canter, Nitrates in Groundwater, Routledge, 2019.
[7] J. Li, G. Yang, D. Zhu, H. Xie, Y. Zhao, L. Fan, et al., Hydrogeochemistry of karst groundwater for the environmental and health risk assessment: the case of the suburban area of Chongqing (Southwest China), Geochemistry 82 (2) (2022) 125866, https://doi.org/10.1016/j.chemer.2022.125866.
[8] J. Pu, D. Yuan, C. Zhang, H. Zhao, Hydrogeochemistry and possible sulfate sources in karst groundwater in Chongqing, China, Environ. Earth Sci. 68 (1) (2013) 159–168, https://doi.org/10.1007/s12665-012-1726-8.
[9] H. Zhang, Y. Xu, S. Cheng, Q. Li, H. Yu, Application of the dual-isotope approach and Bayesian isotope mixing model to identify nitrate in groundwater of a multiple land-use area in Chengdu Plain, China, Sci. Total Environ. 717 (2020) 137134, https://doi.org/10.1016/j.scitotenv.2020.137134.
[10] L. Aller, T. Bennett, J. Lehr, R. Petty, G. Hackett, DRASTIC: A Standardized System for Evaluating Ground Water Pollution Potential Using Hydrogeologic Settings, US Environmental Protection Agency, Washington, D.C: U.S. Environmental Protection Agency, 1987. EPA/600/2-85/018.
[11] I.S. Babiker, M. Mohamed, T. Hiyama, K. Kato, A GIS-based DRASTIC model for assessing aquifer vulnerability in Kakamigahara Heights, Gifu Prefecture, central Japan, Sci. Total Environ. 345 (1–3) (2005) 127–140, https://doi.org/10.1016/j.scitotenv.2004.11.005.
[12] D.C. Wheeler, B.T. Nolan, A.R. Flory, C.T. DellaValle, M.H. Ward, Modeling groundwater nitrate concentrations in private wells in Iowa, Sci. Total Environ. 536 (2015) 481–488, https://doi.org/10.1016/j.scitotenv.2015.07.080. Epub 2015/08/02.
[13] M.R. Khan, M. Koneshloo, P.S.K. Knappett, K.M. Ahmed, B.C. Bostick, B.J. Mailloux, et al., Megacity pumping and preferential flow threaten groundwater quality, Nat. Commun. 7 (1) (2016) 12833, https://doi.org/10.1038/ncomms12833.
[14] N. Dalla Libera, P. Fabbri, L. Mason, L. Piccinini, M. Pola, Geostatistics as a tool to improve the natural background level definition: an application in groundwater, Sci. Total Environ. 598 (2017) 330–340, https://doi.org/10.1016/j.scitotenv.2017.04.018. Epub 2017/04/28.
[15] K. Kalhor, R. Ghasemizadeh, L. Rajic, Alshawabkeh Ajgfsd, Assessment of groundwater quality and remediation in karst aquifers: A review 8 (2019) 104–121, https://doi.org/10.1016/j.gsd.2018.10.004.
[16] J. Podgorski, M. Berg, Global threat of arsenic in groundwater, Science 368 (6493) (2020) 845–850, https://doi.org/10.1126/science.aba1510.
[17] L.A. DeSimone, J.P. Pope, K.M. Ransom, Machine-learning models to map pH and redox conditions in groundwater in a layered aquifer system, Northern Atlantic Coastal Plain, eastern USA, J. Hydrol.: Reg. Stud. 30 (2020), https://doi.org/10.1016/j.ejrh.2020.100697.
[18] O. Rahmati, B. Choubin, A. Fathabadi, F. Coulon, E. Soltani, H. Shahabi, et al., Predicting uncertainty of machine learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods, Sci. Total Environ. 688 (2019) 855–866, https://doi.org/10.1016/j.scitotenv.2019.06.320.
[19] V. Rodriguez-Galiano, M.P. Mendes, M.J. Garcia-Soldado, M. Chica-Olmo, L. Ribeiro, Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain), Sci. Total Environ. 476–477 (2014) 189–206, https://doi.org/10.1016/j.scitotenv.2014.01.001.
[20] K.M. Ransom, B.T. Nolan, J. At, C.C. Faunt, A.M. Bell, J.A.M. Gronberg, et al., A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA, Sci. Total Environ. 601–602 (2017) 1160–1172, https://doi.org/10.1016/j.scitotenv.2017.05.192.
[21] L. Knoll, L. Breuer, M. Bach, Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning, Sci. Total Environ. 668 (2019) 1317–1327, https://doi.org/10.1016/j.scitotenv.2019.03.045.
[22] K.M. Ransom, B.T. Nolan, P.E. Stackelberg, K. Belitz, M.S. Fram, Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States, Sci. Total Environ. 807 (2022) 151065, https://doi.org/10.1016/j.scitotenv.2021.151065.
[23] D. Cain, D.R. Helsel, S.E. Ragone, Preliminary evaluations of regional ground-water quality in relation to land use, Ground Water 27 (2) (1989) 230–244, https://doi.org/10.1111/j.1745-6584.1989.tb00444.x.
[24] Y. Freund, R.E. Schapire (Eds.), A Desicion-Theoretic Generalization of On-Line Learning and an Application to Boosting. European Conference on Computational Learning Theory, Springer, Berlin, Heidelberg, 1995.
[25] T.D. Johnson, K. Belitz, Assigning land use to supply wells for the statistical characterization of regional groundwater quality: correlating urban land use and VOC occurrence, J. Hydrol. 370 (1–4) (2009) 100–108, https://doi.org/10.1016/j.jhydrol.2009.02.056.
[26] F. Sajedi-Hosseini, A. Malekian, B. Choubin, O. Rahmati, S. Cipullo, F. Coulon, et al., A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination, Sci. Total Environ. 644 (2018) 954–962, https://doi.org/10.1016/j.scitotenv.2018.07.054.
[27] V.F. Rodriguez-Galiano, J.A. Luque-Espinar, M. Chica-Olmo, M.P. Mendes, Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods, Sci. Total Environ. 624 (2018) 661–672, https://doi.org/10.1016/j.scitotenv.2017.12.152.
[28] A. El Bilali, A. Taleb, Y. Brouziyne, Groundwater quality forecasting using machine learning algorithms for irrigation purposes, Agric. Water Manag. 245 (2021) 106625, https://doi.org/10.1016/j.agwat.2020.106625.
[29] J. Yan, S. Jia, A. Lv, W. Zhu, Water resources assessment of China's transboundary river basins using a machine learning approach, Water Resour. Res. 55 (1) (2019) 632–655, https://doi.org/10.1029/2018wr023044.
[30] J. Spijker, D. Fraters, A. Vrijhoef, A machine learning based modelling framework to predict nitrate leaching from agricultural soils across The Netherlands, Environmental Research Communications 3 (4) (2021), https://doi.org/10.1088/2515-7620/abf15f.
[31] A. Lahjouj, A. El Hmaidi, K. Bouhafa, Boufala Mh, Mapping specific groundwater vulnerability to nitrate using random forest: case of Sais basin, Morocco, Modeling Earth Systems and Environment 6 (3) (2020) 1451–1466, https://doi.org/10.1007/s40808-020-00761-6.
[32] J.A.E. Agency, ALOS World 3D 30 Meter DEM. V3.2, OpenTopography, 2021.
[33] P. Gong, H. Liu, M. Zhang, C. Li, J. Wang, H. Huang, et al., Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017, Sci Bull (Beijing) 64 (6) (2019) 370–373, https://doi.org/10.1016/j.scib.2019.03.002.
[34] K. Didan. MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V061 [Data set], NASA EOSDIS Land Processes Distributed Active Archive Center, 2021. https://doi.org/10.5067/MODIS/MOD13Q1.061.
[35] W.R. Wieder, J. Boehnert, G.B. Bonan, M. Langseth, Regridded Harmonized World Soil Database v1.2, ORNL Distributed Active Archive Center, 2014.
[36] J. Hartmann, N. Moosdorf, The new global lithological map database GLiM: a representation of rock properties at the Earth surface, G-cubed 13 (12) (2012), https://doi.org/10.1029/2012gc004370.
[37] Y. Fan, H. Li, G. Miguez-Macho, Global patterns of groundwater table depth, Science 339 (6122) (2013) 940–943, https://doi.org/10.1126/science.1229881.
[38] S.E. Fick, Hijmans Rjjijoc, WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, Int. J. Climatol. 37 (12) (2017) 4302–4315, https://doi.org/10.1002/joc.5086.
[39] R.J. Zomer, A. Trabucco, D.A. Bossio, L.V. Verchot, Climate change mitigation: a spatial analysis of global land suitability for clean development mechanism afforestation and reforestation, Agriculture, Ecosystems & Environment 126 (1–2) (2008) 67–80, https://doi.org/10.1016/j.agee.2008.01.014.
[40] C.S. Bureau, Chongqing statistical yearbook, Available from, http://tjj.cq.gov.cn/zwgk_233/tjnj/, 2022.
[41] Z. Yu, J. Liu, G. Kattel, Historical nitrogen fertilizer use in China from 1952 to 2018, Earth System Science Data 14 (11) (2022) 5179–5194, https://doi.org/10.5194/essd-14-5179-2022.
[42] S. Wang, X. Zhang, C. Wang, X. Zhang, S. Reis, J. Xu, et al., A high-resolution map of reactive nitrogen inputs to China, Sci. Data 7 (1) (2020), https://doi.org/10.1038/s41597-020-00718-5.
[43] D.S.-M. Mark Friedl, in: N.L. DAAC (Ed.), MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid, 2015.
[44] I. Gaode, Gaode Map point of interesting search 2.0, Available from: https://lbs.amap.com/, 2016.
[45] A.J. Tatem, WorldPop, open data for spatial demography, Sci. Data 4 (1) (2017) 170004, https://doi.org/10.1038/sdata.2017.4.

[46] J. Chen, M. Gao, S. Cheng, W. Hou, M. Song, X. Liu, et al., Global 1 km × 1 km gridded revised real gross domestic product and electricity consumption during 1992–2019 based on calibrated nighttime light data, Sci. Data 9 (1) (2022) 202, https://doi.org/10.1038/s41597-022-01322-5.
[47] L.E. Condon, R.M. Maxwell, Evaluating the relationship between topography and groundwater using outputs from a continental-scale integrated hydrology model, Water Resour. Res. 51 (8) (2015) 6602–6621, https://doi.org/10.1002/2014wr016774.
[48] E. Nixdorf, Y. Sun, M. Lin, O. Kolditz, Development and application of a novel method for regional assessment of groundwater contamination risk in the Songhua River Basin, Sci. Total Environ. 605–606 (2017) 598–609, https://doi.org/10.1016/j.scitotenv.2017.06.126.
[49] P. Li, D. Karunanidhi, T. Subramani, K. Srinivasamoorthy, Sources and consequences of groundwater contamination, Archives of Environmental Contamination and Toxicology 80 (1) (2021) 1–10, https://doi.org/10.1007/s00244-020-00805-z.
[50] M. Liaw Aw, Classification and regression by randomForest, R. News 23 (23) (2002).
[51] C.G. Tianqi Chen, 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, San Francisco, California, USA, 2016, pp. 785–794.
[52] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232, https://doi.org/10.1214/aos/1013203451.
[53] P. Deville, C. Linard, S. Martin, M. Gilbert, F.R. Stevens, A.E. Gaughan, et al., Dynamic population mapping using mobile phone data, Proc Natl Acad Sci U S A 111 (45) (2014) 15888–15893, https://doi.org/10.1073/pnas.1408439111.
[54] F.R. Stevens, A.E. Gaughan, C. Linard, A.J. Tatem, Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data, PLoS One 10 (2) (2015) e0107042, https://doi.org/10.1371/journal.pone.0107042.
[55] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[56] P. Baudron, F. Alonso-Sarría, J.L. García-Aróstegui, F. Cánovas-García, D. Martínez-Vicente, J. Moreno-Brotóns, Identifying the origin of groundwater samples in a multi-layer aquifer system with Random Forest classification, J. Hydrol. 499 (2013) 303–315, https://doi.org/10.1016/j.jhydrol.2013.07.009.
[57] W. Shangguan, T. Hengl, J. Mendes de Jesus, H. Yuan, Y. Dai, Mapping the global depth to bedrock for land surface modeling, J. Adv. Model. Earth Syst. 9 (1) (2017) 65–88, https://doi.org/10.1002/2016ms000686.
[58] D. Xue, Reconstruction of all-weather daytime and nighttime MODIS aqua-terra land surface temperature products using an XGBoost approach, Rem. Sens. 13 (22) (2021), https://doi.org/10.3390/rs13224723.
[59] J. Zhang, K. Liu, M. Wang, Downscaling groundwater storage data in China to a 1-km resolution using machine learning methods, Rem. Sens. 13 (3) (2021) 523, https://doi.org/10.3390/rs13030523.
[60] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees 77 (4) (2008) 802–813, https://doi.org/10.1111/j.1365-2656.2008.01390.x.
[61] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, 2013.
[62] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of machine learning research 13 (2) (2012) 281–305. https://dl.acm.org/doi/10.5555/2188385.2188395.
[63] P.M. Lerman, Fitting segmented regression models by grid search, J. Roy. Stat. Soc. C Appl. Stat. 29 (1) (1980) 77–84, https://doi.org/10.2307/2346413.
[64] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, N. De Freitas, Taking the human out of the loop: a review of Bayesian optimization, Proc. IEEE 104 (1) (2016) 148–175, https://doi.org/10.1109/JPROC.2015.2494218.
[65] A. Qgis, Free and Open Source Geographic Information System, Open source geospatial foundation project, 2015.
[66] K. Belitz, P.E. Stackelberg, Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models, Environ. Model. Software 139 (2021), https://doi.org/10.1016/j.envsoft.2021.105006.
[67] G. Bassett Jr., R. Koenker, Asymptotic theory of least absolute error regression, J. Am. Stat. Assoc. 73 (363) (1978) 618–622, https://doi.org/10.1080/01621459.1978.10480065.
[68] R. Koenker, K.F. Hallock, Quantile regression, J. Econ. Perspect. 15 (4) (2001) 143–156. http://www.jstor.org/stable/2696522.
[69] D.P. Solomatine, D.L. Shrestha, A novel method to estimate model uncertainty using machine learning techniques, Water Resour. Res. 45 (12) (2009), https://doi.org/10.1029/2008WR006839.
[70] J. Huang, J. Xu, X. Liu, J. Liu, L. Wang, Spatial distribution pattern analysis of groundwater nitrate nitrogen pollution in Shandong intensive farming regions of China using neural network method, Math. Comput. Model. 54 (3–4) (2011) 995–1004, https://doi.org/10.1016/j.mcm.2010.11.027.
[71] X.P. Wei, Pjb, C.Y. Zhao, Assessment of karst groundwater vulnerability in Chongqing based on revised RISKE model, Acta Ecol. Sin. 34 (3) (2014) 589–596, https://doi.org/10.5846/stxb201210301504.
[72] A. Hartmann, S. Jasechko, T. Gleeson, Y. Wada, B. Andreo, J.A. Barberá, et al., Risk of groundwater contamination widely underestimated because of fast flow into aquifers, 10.1073/pnas.2024492118 118 (20) (2021) e2024492118.