Water Resources Research

RESEARCH ARTICLE
10.1029/2024WR037753

Mapping Shallow Groundwater Solute Footprints in Arid Regions Using a Hydrologically Enhanced Species Distribution Model

Jianguo Li[1], Zunyi Xie[2,3], Dehua Mao[1], Ruijun Long[4], An Liu[5], Qiang Yu[6], Zongming Wang[1], Xin Yang[4], Chongliang Zhong[7], Madden Solomon[8], and Daniel Ramp[8]

[1]State Key Laboratory of Black Soils Conservation and Utilization, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China, [2]Faculty of Geographical Science and Engineering, Henan University, Zhengzhou, China, [3]Key Laboratory of Geospatial Technology for the Middle and Lower Yellow River Regions, Ministry of Education, Henan University, Kaifeng, China, [4]State Key Laboratory of Grassland and Agro-Ecosystems, International Centre for Tibetan Plateau Ecosystem Management, College of Ecology, Lanzhou University, Lanzhou, China, [5]College of Chemistry and Environmental Engineering, Shenzhen University, Shenzhen, China, [6]State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Soil and Water Conservation, Northwest A&F University, Yangling, China, [7]College of Animal Science and Technology, Northeast Agricultural University, Harbin, China, [8]Centre for Compassionate Conservation, TD School, University of Technology Sydney, Ultimo, NSW, Australia

Abstract Solute concentrations and accumulation in shallow groundwater can trigger a negative domino effect of environmental and economic problems. Accurately tracking the solute chemistry of groundwater is essential for assessing their adverse impacts and pinpointing hotspots of contaminants that require conservation measures. However, existing mapping methods have been greatly limited by the peculiarities of shallow groundwater, such as its challenging hydraulic connection, uneven distribution, and complex driving factors prevalent in arid regions. To address these challenges, we designed a novel framework that integrates species distribution models (SDMs) with traditional hydrological models and advanced machine learning algorithms to predict the spatial distribution of groundwater solutes in a multidisciplinary effort. We carried out a systematic collection of shallow groundwater, deep groundwater, and surface water samples from three adjacent hydrological units in the arid regions of northwest China. By employing the SDMs framework originally utilized in ecology to assess biological species suitability, we could simulate and predict solute concentrations in groundwater. The results emphasized that solutes in surface water were important variables in the final models for Na^+, K^+, SO_4^{2-}, and Cl^-, while deep groundwater influenced Ca^{2+}. In addition, integrating predictor variables into the SDMs enabled the discovery of additional valuable information. Our results highlighted the improved power for mapping groundwater by combining SDMs with multidimensional driving factors. This novel framework could not only clarify the solute movement mechanisms but also reveal their spatial patterns via a transdisciplinary approach, offering a versatile tool for groundwater management and policies.

Plain Language Summary This study examines solute contamination in shallow groundwater, which can lead to serious environmental and economic issues. Accurately tracking solutes is essential for understanding their impact and identifying areas that need protection. However, current methods are constrained by the sparse field samples of shallow groundwater, including its uneven distribution, data collection difficulties, and complex movement patterns, especially in arid regions. To address these challenges, we developed a new approach combining species distribution models (SDMs), hydrological models, and machine learning. This method improves predictions of solute distribution in groundwater with limited water samples from shallow and deep groundwater, as well as surface water, in northwest China. Using the SDM framework, we successfully predicted solute concentrations in shallow groundwater, and found that surface water influenced solutes like Na^+, K^+, SO_4^{2-}, and Cl^-, while deep groundwater affected Ca^{2+}. Integrating multiple factors into the model provided new insights into the complex interactions between groundwater and surface water. Our proposed innovative framework enhances the mapping of groundwater contamination and offers a valuable tool for groundwater management, especially in water-scarce regions.

Key Points:

- Shallow groundwater solute footprints captured using species distribution model
- Species distribution models combined with hydrological models improve solute mapping accuracy
- Transdisciplinary approach reveals solute movement patterns in groundwater systems

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Z. Xie,
zunyixie@henu.edu.cn

Citation:

Li, J., Xie, Z., Mao, D., Long, R., Liu, A., Yu, Q., et al. (2025). Mapping shallow groundwater solute footprints in arid regions using a hydrologically enhanced species distribution model. *Water Resources Research*, *61*, e2024WR037753. https://doi.org/10.1029/2024WR037753

Received 19 JUL 2024
Accepted 16 JUN 2025

Author Contributions:

Conceptualization: Jianguo Li, Zunyi Xie
Data curation: Jianguo Li, Chongliang Zhong
Formal analysis: Jianguo Li
Methodology: Jianguo Li, Zunyi Xie, Ruijun Long, An Liu, Daniel Ramp
Supervision: Zunyi Xie, Dehua Mao, Ruijun Long, Qiang Yu
Visualization: Jianguo Li, Xin Yang
Writing – original draft: Jianguo Li
Writing – review & editing: Jianguo Li, Zunyi Xie, Zongming Wang, Madden Solomon, Daniel Ramp

1. Introduction

Groundwater is an indispensable and widespread resource for sustaining life, ecosystems, and economic stability. However, the water quality of groundwater has been increasingly deteriorating due to increased consumption and climate change (Kaushal et al., 2018; Wu et al., 2021). Solute dynamics and accumulation in subsurface water can induce a series of water quality problems, such as salinization, acidification, mineralization, alkalization, softening, and hardening (Wu et al., 2021). In addition, shallow groundwater (SGW, unconfined aquifers), serving as the transitional zone between surface water and deep groundwater (confined aquifers) (Figure 1), is the most active and variable layer in the groundwater system (Waller, 1994). This makes SGW particularly vulnerable to ecological challenges and anthropogenic activities (Smith et al., 2018; Thaw et al., 2022). Thus, the solute issues of SGW are prone to triggering a domino effect, leading to a series of negative impacts including soil salinization, vegetation degeneration, desertification, and sandstorms. These impacts may ultimately break the ecosystem balance, threaten human lives (Lapworth et al., 2013), and endanger biodiversity (Giam et al., 2018).

Despite these significant impacts, the regular water quality monitoring of SGW over large-scale area has been challenging due to the lack of available data worldwide, especially in dryland regions (Setiawan et al., 2022). As shown in Figure 1, the dynamics and transport of SGW solutes often involve both natural processes and human activities, closely intertwining with the surrounding environments. SGW solutes may accumulate through multiple pathways, including hydraulic processes, topography, climate, plant interactions, soil properties, and anthropogenic activities such as irrigation, abstraction, pesticides and fertilizers, urbanization and sewage disposal (Nolan & Weber, 2015). High order interactions and complex relationships between solutes and associated drivers thus often exist. For example, irrigation can move groundwater level upwards, potentially increasing salt and sodium concentrations within the root zone (Shouse et al., 2010) (Figure 1). This in turn triggers the implementation of additional mitigation methods to reduce soil salinization in cropping systems. These relationships could be more complicated in arid regions (Ben Messaoud et al., 2021; Dandge & Patil, 2022), and understanding the dynamics of groundwater chemistry in such areas is crucial for sustainable development goals (e.g., SDG6-clean water) (Sadoff et al., 2020).

Previous studies have extensively investigated the sources and dynamics of groundwater chemistry to address three primary questions: What factors drive the movements of groundwater chemistry? How to quantify and evaluate the transport of these chemicals? And how can we predict their spatial distributions? Previous studies have made many efforts to address these general questions. For example, one promising approach is the introduction of models applicable in hydro-geochemistry, such as box models and physics-based models (Maavara et al., 2021; Rogers et al., 2021). These methods have proven to be more effective than traditional observational techniques in tracing the reactive chemistry of surface water and groundwater and quantifying their relative contributions. They not only reduce the need for time-consuming and labor-intensive tasks, but also eliminate the requirement for specialized training (Barthold et al., 2011). Additionally, many novel algorithms have been proposed to define important driving factors and their interactive impacts on geochemical cycles at terrestrial-aquatic interfaces (Cheng et al., 2021; Enguehard et al., 2022). Despite significant advancements in the field, three major limitations remain to be addressed. Gap 1 (hydraulic connection): Hydraulic connection substantially influences solute dynamics, but previous studies were mostly confined to specific regional or site-specific groundwater systems, such as geomorphic units, administrative units, hydrological units, or typical land use locations (key result 1 in Table S1 in Supporting Information S1). Thus, the simulation of solute dynamics may suffer from unrealistic representations if the hydraulic factors are neglected. This limitation calls for a more comprehensive approach that considers the broader hydrological context and incorporates hydraulic connectivity between aquifers; Gap 2 (incorporating variables): As solute dynamics in groundwater are driven by both intrinsic and extrinsic factors, it is necessary to incorporate these factors into mapping models to enhance accuracy. Yet, these predictor variables exhibit spatial-temporal heterogeneity, and the scale mismatch between them and groundwater pose challenges to gain deeper insights into the underlying mechanisms, especially at large-scale areas (key result 2 in Table S1 in Supporting Information S1). Addressing this limitation requires to effectively capture the complex interplay between driving factors and solute dynamics across various scales; Gap 3 (spatial prediction): Existing mapping methods heavily rely on spatial interpolation techniques such as kriging and inverse distance weighting for groundwater prediction (e.g., option 2 in Figure 2, key result 3 in Table S1 in Supporting Information S1). However, these traditional approaches encounter challenges in arid regions due to the patchy and sparse distributions of groundwater field observations (Figures 2a and 2c). The limited availability

10.1029/2024WR037753

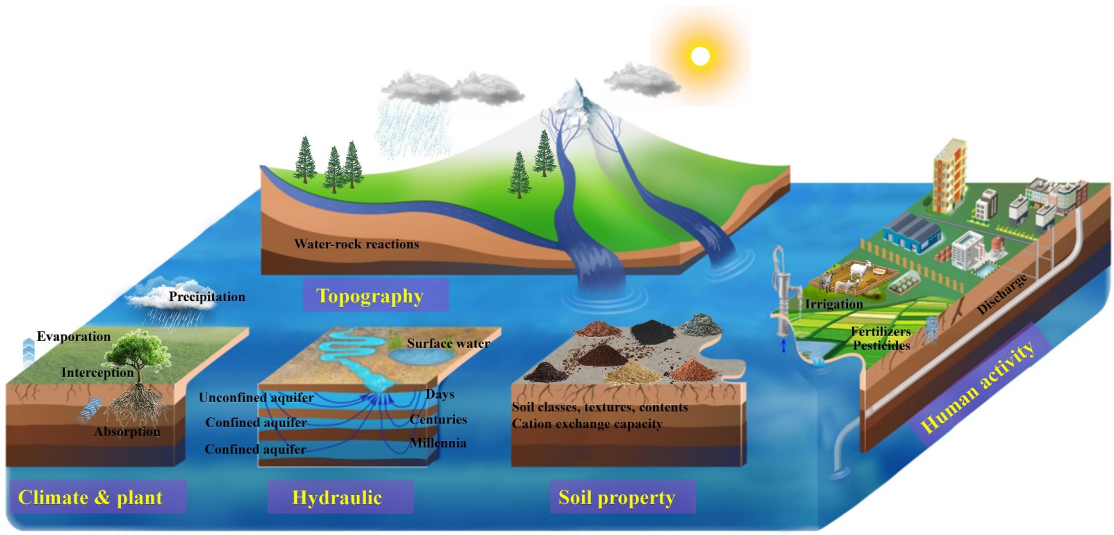

Figure 1. A schematic illustration of groundwater solute dynamics, showing the hydraulic connections and interplay between natural processes (e.g., climate, soil properties, plant uptake, and groundwater-surface water interactions) and human disturbances (e.g., urbanization, agriculture, waste disposal), as well as their combined impact on groundwater systems.

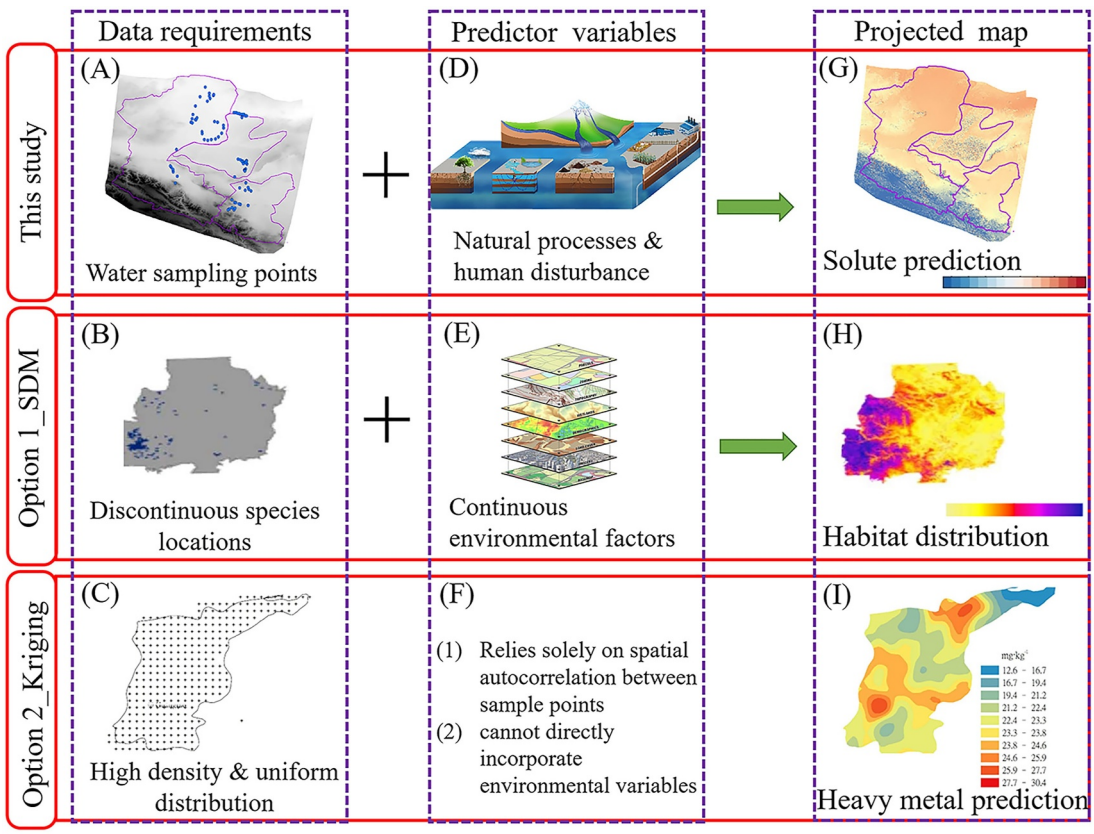

Figure 2. Comparison of this study's approach with two existing ones, species distribution model (SDM) and Kriging options. *Data requirements*: All require input distribution data but Kriging requires uniform high density data, (a) water sampling points in this study, (b) discontinuous biological species locations in SDM, (c) high-density, uniformly distributed sampling data in Kriging. *Predictor variables*: Both integrate continuous environmental variables but Kriging depends solely on spatial autocorrelation, (d) natural processes and human disturbances, (e) continuous environmental factors in SDM, (f) reliance solely on spatial autocorrelation in Kriging. *Projected maps*: Both output spatially continuous map but Kriging produces non-contextual predictions, (g) continuous solute prediction in this study, (h) habitat distribution, (i) heavy metal prediction without geographic context in Kriging.

of representative samples hampers precise interpolation, necessitating the development of novel methods that overcome the data scarcity issue and couple into accurate predictors.

Our proposed solution is the "ecologicalization" of solute to address these gaps, which is to borrow a classic ecological approach, species distribution models (SDMs) and apply it towards solute distributions in lieu of biological species (Option 1 in Figure 2). SDMs involve three steps: selecting biological observations, identifying predictor variables, and quantifying their relationships to create distribution maps by projecting this onto environmental layers (Figures 2b, 2e, and 2h). These steps align with the three scientific gaps as follows. First, SDMs are designed to handle sparse and discontinuous biological data, such as dispersed populations (Figure 2b), which contrasts with traditional Kriging methods that require dense, continuous sampling (Figure 2c). Second, SDMs predict species distribution by modeling relationships between species occurrences and environmental factors (Elith & Leathwick, 2009), integrating geographic layers (Figure 2e). In contrast, Kriging ignores environmental factors (Figure 2f). Third, SDMs project the model onto geographic layers, capturing environmental influences like terrain (Figure 2h), while Kriging results deviate (Figure 2i). Thus, the key steps in the SDM framework systematically address the three scientific gaps, positioning SDMs as a promising tool for filling these knowledge gaps. While SDMs have demonstrated significant predictive power in various research fields (Clemente et al., 2019; Harrigan et al., 2014; Li et al., 2020), their application to groundwater chemistry remains largely unexplored.

Here, we proposed a novel framework for predicting the spatial distribution of solutes in shallow groundwater (SGW) that builds on the classic SDM conceptual framework by modifying and combining it with hydrological models, statistical algorithms, and machine learning techniques. We "ecologicalized" solute concentrations into the incidence (presence/absence) of biological species. The solute predictor variables, which link five-dimensional drivers that govern groundwater dynamics (topography, climate, hydraulic, soil property and human activity, Figure 1), were then coupled into the SDMs to predict the spatial patterns of groundwater chemistry. In particular, we aimed to: (a) establish hydraulic connection between surface water and groundwater in the adjacent catchments for solute transport analysis; (b) better interpret the dynamic mechanism of solutes in SGW by employing a comprehensive system that integrates five-dimensional predictor variables; and (c) improve the accuracy of groundwater chemistry mapping. Our study highlights the potential of SDMs for better mapping SGW solutes on regional to large scales from an ecological perspective.

2. Study Area

Our study areas included the Badain Jaran Desert (BJD), Heihe River Basin (HRB), and Shiyanghe River Basin (SYH) in northwest China (Figure 3a). This area covers approximately 38×10^4 km^2 and presents extremely complex hydraulic connectivity at the junction areas of the above three hydrological units. Previous research has provided valuable information on the groundwater flow paths, geological structure and other background compositions (Chen et al., 2006; Ding & Wang, 2007; Wu et al., 2010) (Figure 3). The water properties in this area are notably unique, with nearly 100 hypersaline and freshwater lakes coexisting in the arid BJD. Furthermore, the salinity of groundwater in the Hexi Corridor ranks among the highest levels observed in China (Wang et al., 2017), rendering it an ideal location for our study due to the solute diversity within these water bodies.

There are exceptionally intricate conditions in this study area due to its complex natural background. The background topography exhibits significant spatial heterogeneity, with elevation ranging from 700 to 5,000 m asl. The area is surrounded by high mountains, deserts, and other landform units. Additionally, the landscape of BJD is characterized by continuous crescentic mega-dunes, which are the tallest sand dunes on earth, with a relative height of 200–300 m (Wang & Zhou, 2018). These unique traits make groundwater flow directions and hydraulic connections highly uncertain. Therefore, the traditional spatial interpolation methods applicable in flat areas are not suitable for this study area. In addition, the geological variability of this region is evident (Figure 3b). The BJD geological structure, located in the Alxa block, has a gentle landform with denuded low hills and inter-mountain depressions. The surface is covered by Quaternary sediments that form vast Gobi and desert areas (BB′ in Figure 3) (Ma & Edmunds, 2006). The HRB has a deep Quaternary sedimentary layer that acts as an efficient underground water storage area. The base is composed of Neo-Tertiary or Cretaceous layers overlain by hundreds of meters to over a thousand meters of Quaternary loose material with high groundwater content (AA′ in Figure 3) (Wang et al., 2013). In Figure 3 CC′, the Jinchang basin is characterized by interbedded sub-consolidated fluvio-lacustrine sandstones from the Pliocene and early Pleistocene periods, with an average thickness of 100–200 m in

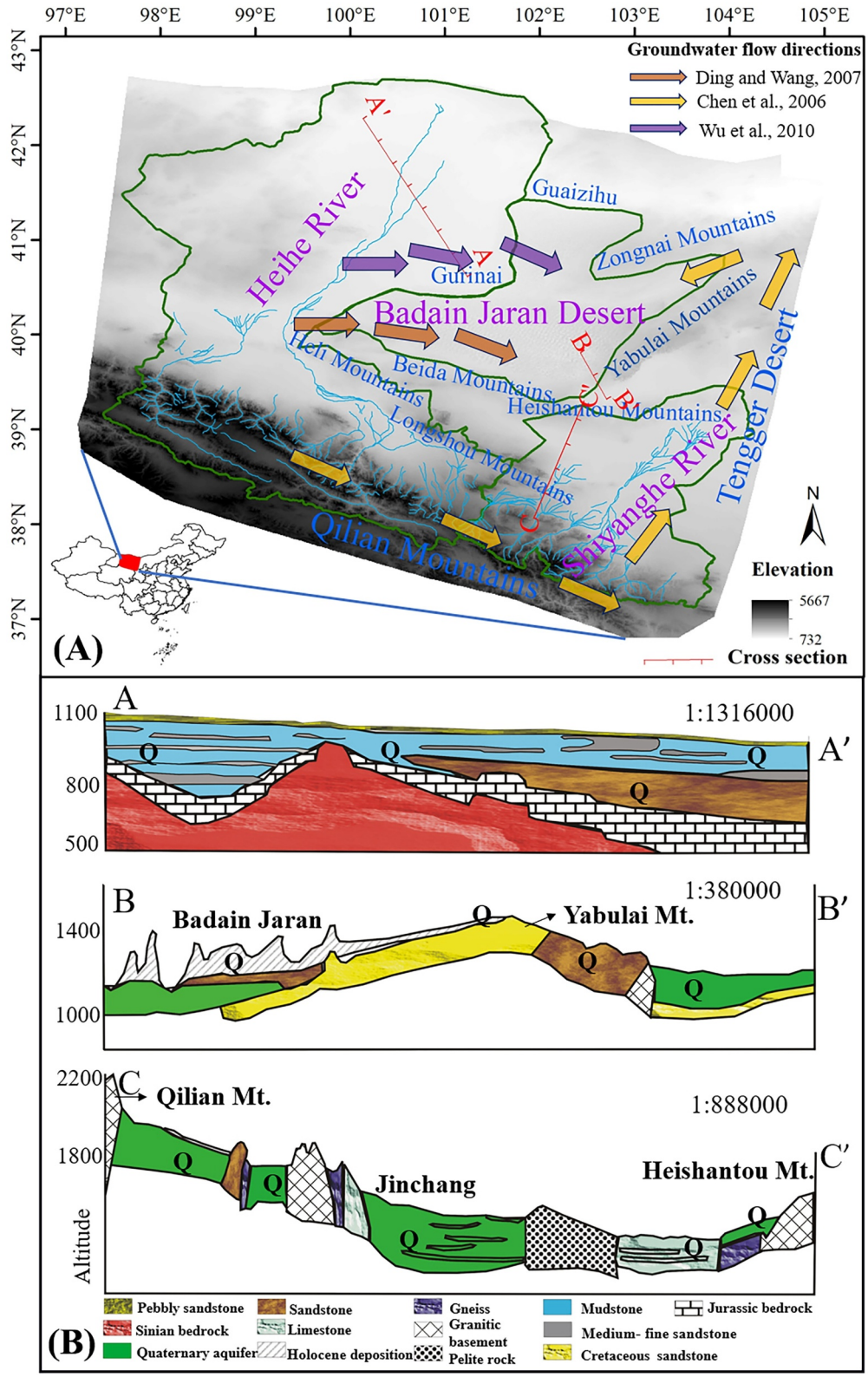

Figure 3. (a) Overview of cross-sections, landforms, and groundwater flow directions within three hydrological units in the Badain Jaran Desert region. (b) Cross-sectional profiles corresponding to the locations marked in (a), showing detailed geological structures, aquifers, and sedimentary layers.

the southern region and sandy-clay and clay interbeds, with a thickness of 70–100 m in the northern region (Ma et al., 2010). The high alpine and remote desert areas experience a diverse range of climates, characterized by diminishing levels of moisture with decreasing altitude. In the middle and downstream regions of these catchments, the annual precipitation averages around 200 and 50 mm, respectively, while evaporation rates soar beyond 2,000 mm. The annual average temperatures are approximately 6°C and 8°C in the middle and lower regions (Li et al., 2022). The BJD is located at the boundary of the East Asian Monsoon and receives an annual precipitation of 50–60 mm, primarily during June-August. It is prone to an annual evaporation of over 3,500 mm, with an average temperature of 7–8°C (Wang & Zhou, 2018).

The anthropogenic factors of the study area show apparent spatial heterogeneity as well. The corridor plains, located in the middle reaches of HRB and SYH catchments, serve as population centers and substantial agricultural production bases, popularly known as "ancient Silk Road," "Golden Zhangye" and "Silver Wuwei" (Zhou et al., 2015). Human activities in the middle and lower reaches of the catchments such as spring irrigation and summer floods have enhanced the interactions between surface water and groundwater. The discharge of pesticides, chemical fertilizers, and wastewater from human settlements has significantly impacted groundwater solutes. In contrast, the alpine regions, Gurinai and Guaizi Lake have relatively sparse population densities. Moreover, in the BJD, the population density is less than 1 person per 10 km^2. These distinct natural and anthropogenic 'factors on a large spatial scale have made the solute transport mechanism in SGW complex.

3. Methods

3.1. Adapting SDM for Hydrological Application

3.1.1. Groundwater and Surface Water Chemistry Data

Groundwater and surface water solute measurements throughout the study area were compiled from a comprehensive collection of published or publicly available sources. A summary of the collected solute measurements, including basic statistics and sources, can be found in the Supporting Information S1, Table S2 in Supporting Information S1. We strictly adhered to the classic methodology and process for collecting groundwater chemical indicator data (Amini et al., 2008; Podgorski & Berg, 2022; Podgorski et al., 2018). Based on our research objectives and unique features of the region, the following criterions for solute selection were applied to minimize possible bias: (a) Samples were collected from georeferenced locations along the groundwater flow paths, aiming to provide a representative and informative data set (Figure 3a); (b) Solute indicators were employed to assess water quality; (c) Well depths, along with any information regarding the type of surface water bodies if applicable, were meticulously documented in the literature; (d) Conventional observation indices, such as ion measurements in mg/L or mol/L, were utilized for quantifying water quality; (e) The sampling process and laboratory analysis procedures were thoroughly documented, and groundwater samples were primarily obtained from boreholes, monitoring wells, or other appropriate sampling techniques. In total, we collected 82 SGW samples, 114 deep groundwater samples, and 170 surface water samples, along the potential groundwater flow directions (Figure 4, Table S2 in Supporting Information S1). Major ions, such as Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, SO_4^{2-}, Cl^{-}, HCO_3^{-}, were selected as the solute tracers to investigate the dynamics of solutes in SGW.

Based on both common standards in Asia and field observations in the study area (Fendorf et al., 2010; Wang et al., 2013), we used a buried depth of 20 m as the threshold for distinguishing between shallow and deep groundwater. Furthermore, to investigate the hydraulic connections between different hydrological units in the study area, we divided the groundwater into seven mapped zones by comprehensively considering the spatial distributions of SGW observation wells, watershed sections, and potential groundwater flow directions. These zones were designated as Zone 1 (Z1), Z2, and Z3, representing the upper, middle, and lower reaches of the HRB, respectively. Additionally, Z4 was situated in the Gurinai area, Z5 in the BJD, Z6 in the Guaizihu Lake area, and Z7 in the SYH basin (Figure 4). To aid clarity, we further categorized surface water as Layer 1 (L1), which includes precipitation, lake, and river water, while shallow and deep groundwater were classified as Layer 2 (L2) and Layer 3 (L3), respectively (Figure 4).

3.1.2. Selection of Predictor Variables for SDM

The predictor variables were selected as proxies to model the subsurface conditions if: (a) They had been proven by previous studies to have profound impacts on the solute dynamics; (b) They needed to have broad applicability

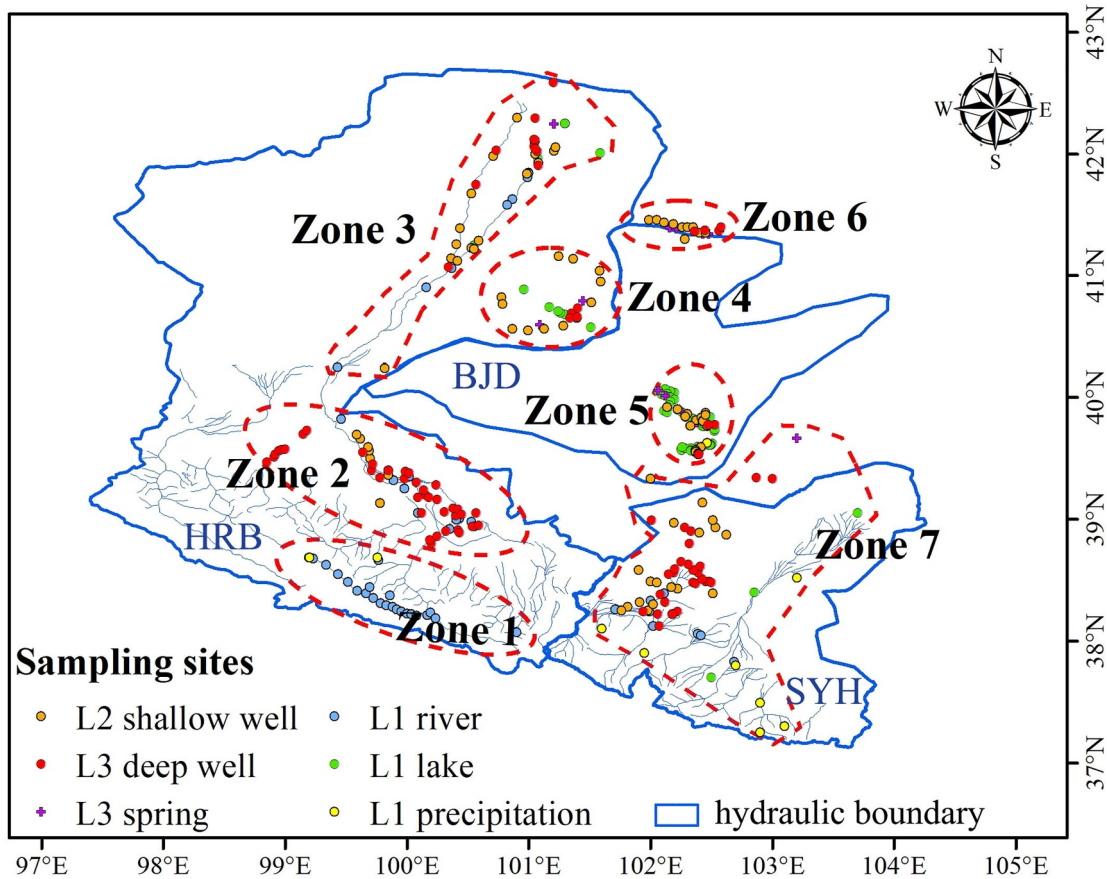

Figure 4. Sampling sites for various water bodies across zones. Sites include shallow wells (L2), deep wells (L3), springs (L3), rivers, lakes, and precipitation (L1). Red dashed lines indicate zones, and blue outlines mark hydraulic boundaries.

beyond this study area; and (c) these factors had to be spatially quantifiable and exhibit minimal collinearity with each other. In total, we collected five types of explanatory variables (Table 1). These variables include topographical patterns; climate and vegetation variables, including normalized difference vegetation index (NDVI), precipitation, evaporation, and ground surface temperature (GST); human activity variables, such as gross domestic product (GDP), population density, land use categories; and soil properties. The land use data contains six categories: agriculture, forest, grassland, water, intensive-use areas, and unexploited areas (Table 1). Given that SDMs require predictor variables in a continuous map format, we established innovative hydraulic variables and generated continuous maps to address Gap 1 (hydraulic connection). A key improvement in our study is the inclusion of spatially continuous maps of groundwater solutes as predictor variables. These hydraulic variables were derived from preliminary output maps using SDMs that considered only topography conditions, with detailed generation steps described in Section 3.2.2. Other variables were retrieved from various sources, including social, statistical and geographic data. For example, terrain, climate and vegetation, and human activity data were extracted from the Resources and Environment Data Center of the Chinese Academy of Sciences (http://www.resdc.cn). Additionally, we aquired 250 m-resolution soil grid layers and 10 indicators, including cation exchange capacity, soil coarseness, silt, clay contents, and pH, from the open database licensed by SoilGrids organization under the Zenodo organization (https://zenodo.org/). Overall, we selected 30 predictor variables from five categories (Table 1, Figure 1) to more accurately capture the underlying mechanism governing water solutes, thereby improving the model's predictive performance to reflect real-world scenarios.

3.1.3. SDM Algorithm

To enhance the applicability and versatility of our framework, we moved beyond relying solely on individual SDMs such as MAXENT, BIOCLIM, and DOMAIN. Instead, we transitioned to an ensemble framework that

Table 1
Potential Predictor Variables for Solutes in Shallow Groundwater

Variables	Comments	Resolution	Unit
Topography patterns			
DEM(6x)	elevation, flow direction, slope, rough, tpi, tri	30 m	m asl
Hydraulic connections			
$SGW_{chemistry}$	Shallow groundwater chemistry	30 m	mg/L
$DGW_{chemistry}$	Deep groundwater chemistry	30 m	mg/L
Climate & plant variables			
NDVI	Normalized difference vegetation index	1 km	-
evaporation	Annual evaporation amount	500 m	mm
GST	Ground surface temperature	500 m	°C
precipitation	Annual precipitation amount	500 m	mm
Human variables			
GDP	gross domestic product	1 km	10^4 RMB km^{-2}
P_ density	population density	1 km	inhabitants km^{-2}
Land use(6x)	Six different land use types	30 m	m^2
Soil properties			
soil. cation	cation exchange capacity	250 m	_
soil. coarse	soil coarse fragment	250 m	volumetric %
soil. texture	soil texture classes	250 m	1,2,3,…12
soil. silt	soil silt content	250 m	% (kg/kg)
soil. clay	soil clay content	250 m	% (kg/kg)
soil. sand	soil sand content	250 m	% (kg/kg)
soil. carbon	soil organic carbon content	250 m	5 g kg^{-1}
soil. pH	soil pH	250 m	_
soil. water	soil water content	250 m	volumetric %
soil. bulk	soil bulk density	250 m	10 kg m^{-3}

Note. The six topographic patterns were extracted from the digital elevation model (DEM) products. The land use types conclude agricultural land, forest land, grassland, water bodies, built-up land and unexploited land. Soil property indicators were sampled at a depth of 10 cm below the surface, and the toil texture was classified 12 types: sand, loamy sand, sandy loam, sandy clay loam, loam, silt loam, silt, silty clay loam, clay, clay loam, sandy clay, and silty clay.

enables the incorporation of a wider range of techniques and methodologies. Drawing from foundational SDMs tutorials (Guisan & Thuiller, 2005; Hijmans & Elith, 2013), we further designed and modified a SDMs framework. The formula for SDMs is generally expressed as:

$$\text{logit}(Y) = \alpha + \sum_{i=1}^{n}(\beta_i X_i) \quad (1)$$

Logit (Y) represents the logarithm of the log odds for the probability or potential distribution of species, while X represents the related variables, such as competitors, predators, climate, and terrain. β is the regression coefficient that measures the contribution of each variable on the species distribution.

3.2. Integrated Hydrologic-SDM-ML Framework

To address the three knowledge gaps, the framework includes four steps (Figure 5). Step 1 develops preliminary maps for solutes in shallow and deep groundwater. Step 2 extracts two raster layers using surface sampling points, establishing solute dynamics between hydrological layers. Step 3 optimizes relationships between solutes and

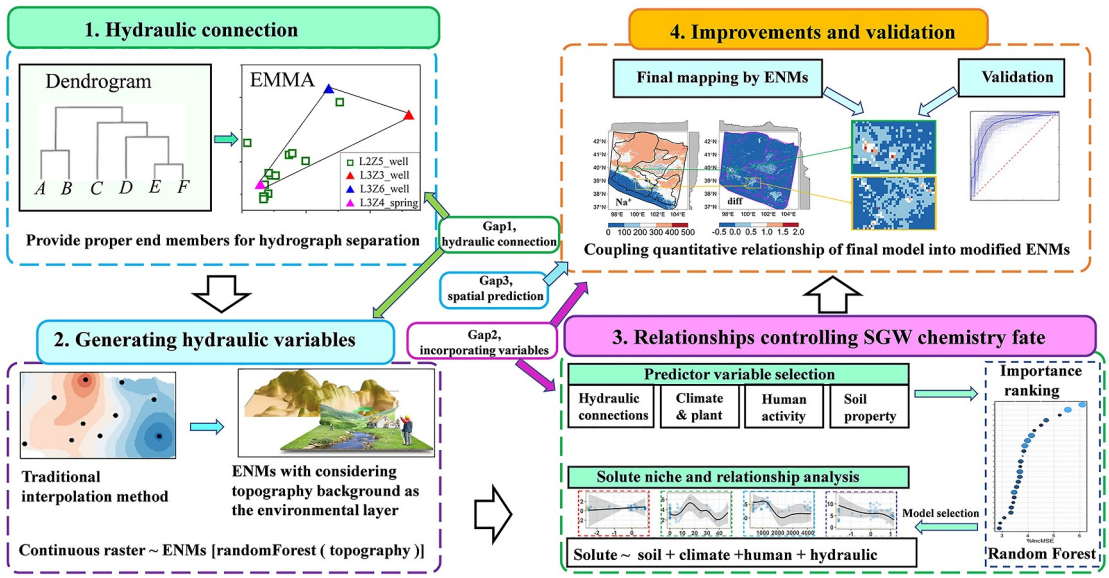

Figure 5. Flowchart outlining the main steps to address specific knowledge gaps. Step 1: Hydraulic connection- identify appropriate end members for hydrograph separation using dendrograms and EMMA (end-member mixing analysis). Step 2: Generating hydraulic variables- apply SDMs (species distribution models) with continuous rasters, incorporating topography as an environmental layer to address Gap 1 (hydraulic connection). Step 3: Relationships controlling SGW solute fate- select key predictor variables and analyze solute distribution and relationships. Step 4: Improvements and validation- conduct final mapping with modified SDMs by integrating quantitative relationships and validate the results to improve model accuracy.

predictor variables, and Step 4 projects these relationships into SDMs to resolve Gap 2 (variable incorporation) and Gap 3 (spatial prediction). By integrating hydrological models, statistical analyses, and machine learning, the framework enhances both efficacy and applicability.

3.2.1. Hydraulic Connection Analysis (Step 1)

We used hierarchical cluster analysis and the end member mixing analysis (EMMA) tool to establish hydraulic connections between adjacent hydrological units. By clustering water samples based on chemical dissimilarity, we identified potential end members and quantified hydraulic connections across seven zones.

3.2.1.1. Hierarchical Cluster Analysis

This section establishes a quantitative understanding of the connections between neighboring hydrological units (Step 1 in Figure 5), which provides a basis for generating spatially continuous raster data for groundwater and subsequent analysis of solute dynamics. We used hierarchical clustering to categorize water samples by their chemical compositions. This method groups objects by dissimilarity, initially placing each in its own cluster. It then merges the most similar clusters iteratively until one remains, representing a distinct hydrogeochemical state. To accurately capture chemical features, we combined Euclidean distance and Ward's method. The Euclidean distance between water types α and β for N quantitative chemical concentrations was defined as:

$$\mathrm{Ed}_{\alpha\beta} = \sqrt{\sum_{i=1}^{N} (X_{i\alpha} - X_{i\beta})^2} \tag{2}$$

where Ed is the Euclidean distance, with $X_{i\alpha}$ and $X_{i\beta}$ being the concentrations of specific soluble ion "i" for water types α and β, respectively. Next, we employed the Ward's minimum variance method to construct a dendrogram for classifying the water samples based on their chemical compositions. Based on the data characteristics in this study, we divided the data set into six distinct groups.

3.2.1.2. End Member Mixing Analysis

The EMMA, a typical hydrograph separation method, can explain the chemical ingredients of a water body as a mixture of potential end members using conservative tracers. The method is described as the following equations (Christophersen et al., 1990):

$$\begin{cases} 1 = a_1 + a_2 + a_3 \\ C_t^1 = C_1^1 a_1 + C_2^1 a_2 + C_3^1 a_3 \\ C_t^2 = C_1^2 a_1 + C_2^2 a_2 + C_3^2 a_3 \end{cases} \tag{3}$$

where a_1, a_2, and a_3 are the discharge fractions of end members. $C_t{}^1$ and $C_t{}^2$ are the tracer concentrations of mixed groundwater. $C_n{}^1$ and $C_n{}^2$ are the tracer concentrations of the nth end member. In this study, the $C_t{}^1$ and $C_t{}^2$ are the tracer concentrations of shallow groundwater in the BJD, which was assumed to be recharged from different water types in the three hydrological units.

3.2.2. Generating Hydraulic Variable Layers (Step 2)

We incorporated topographic variables into the SDMs to generate spatially continuous raster layers for both SGW and deep groundwater. We then extracted these raster layers from the sampling locations to establish hydraulic connections between water samples at different depths. The topographic features, including flow direction, elevation, slope, TPI, roughness, and TRI, were extracted from the Digital Elevation Model (DEM) product. We then utilized the Random Forest algorithm to establish the relationships between solute concentrations and these topographic features (step 2 in Figure 5). Random Forest is an ensemble algorithm known for its robust performance in regression and classification modeling (Wang et al., 2016). The regression tree splitting criterion involves selecting the input variable with the lowest Gini index,

$$I_G(t_{X(xi)}) = 1 - \sum_{j=1}^{m} f(t_{X(xi)}, j)^2 \tag{4}$$

$f(t_{x(xi)}, j)^2$ represents the proportion of samples with the value x_i belonging to leaf j as node t. The predicted value of an observation is calculated by averaging the outputs of multiple regression trees. Then, the prediction value of the random forest regression is obtained by averaging the regression prediction values of the decision trees, as shown in the following equation,

$$M(X) = \frac{1}{m} \sum_{i=1}^{n} h_i(x) \tag{5}$$

The groundwater chemistry and geology relationship established by Random Forest in this study can be further mapped onto geographic layers of environmental information to improve the prediction accuracy. Although this approach may introduce uncertainties due to the spatially limited data from the water sampling points, it captures the influence of surface water and deep groundwater on the solute dynamics of shallow groundwater. This improves prediction accuracy and makes the simulation more realistic and reliable compared to traditional interpolation methods.

3.2.3. Analyzing Relationships Between Solutes and Predictor Variables (Step 3)

To comprehensively analyze the relationships between solutes and predictor variables, we employed model selection and incorporated several algorithms (step 3 in Figure 5). Generalized linear model (GLM) allows for direct observation of the quantitative associations between solutes and variables using linear functions. Random Forest was used for variable importance analysis on the initial set of 30 predictor variables. This analysis assisted in selecting the top two drivers in each dimension for subsequent model selection to avoid the over-fit analysis. Generalized additive model (GAM) was then utilized to visualize nonlinear effects in the relationships

graphically. Consequently, we quantitatively assessed the contributions of individual variable, determining their magnitude and direction of influence on the solute dynamics.

During model selection, we employed the Akaike Information Criterion corrected for small sample sizes (AICc) to choose the final model from the best models with the highest area under the curve, as determined through the cross-validation process. The formula for calculating AICc is as follows:

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2K(K+1)}{n-K-1} \tag{6}$$

K represents the number of free parameters in the model, including regression coefficients, intercept, and other model parameters, while n represents the number of observed samples.

3.2.4. Final Mapping and Model Assessment (Step 4)

To generate final prediction maps, we projected the quantitative relationships of final models that were identified in the previous step. The top predictor variables corresponding to each solute in the final model were then integrated into the modified SDMs, with topography serving as the background layer. Although we employed statistical analysis in Step 3 for visual representation of the relationships, we utilized the more robust Random Forest algorithm in this stage. The Random Forest can help fit the regression relationships between solutes and the environmental layers representing various variables in the final model to generate improved output maps by Equation 1.

To evaluate the predictive performance of our model for groundwater, we utilized K-fold cross-validation, in which the data set was randomly divided into 80% training and 20% testing sets (Podgorski & Berg, 2020). We also examined different measures, such as the Area Under the receiver operating characteristic curve (AUC), sensitivity (the true positive rate), and specificity (the true negative rate). Considering the characteristics of our data and practical requirements, we also comprehensively evaluated the reliability and applicability of the model by comparing the preliminary maps from Step 2, incorporating our expertise, and considering relevant literature records (Amini et al., 2008).

4. Results

4.1. Hydraulic Connection Variables

Based on the expertise of the study area (Figure 3a) and the results of dendrogram analysis (Figure 6a), we selected L2Z5 (Layer2Zone5, shallow groundwater in BJD, Figure 4) as the receiving waterbody, while other water bodies at different depths were considered as potential end members.

Most waters in L1 (Layer1, surface water) exhibited similar characteristics, except for the saline lakes in the BJD region (Figure 6a). The cluster analysis revealed that river water from Zones 1, 2, 3, and Z7 (HRB and SYH in Figure 4, indicated by yellow rectangles in Figure 6a), along with precipitation in Zones 1 and 7 (also represented by yellow rectangles in Figure 6a), shared similar geochemical features and were grouped together on the same branches (purple trees in Figure 6a). We determined that precipitation in Z5 (L1Z5_precipitation, blue tree in Figure 6a) and Z7 (L1Z7_precipitation, purple tree in Figure 6a) were the dominant end members, contributing 78.07% and 21.93%, respectively (Figure 6b, Table 2).

We observed distinct clustering patterns in the groundwater distribution (Figure 6a). Groundwater around the BJD region formed clusters on the purple branches, as indicated by the red rectangles in Figure 6a. Conversely, groundwater in the middle and lower regions of the HRB and SYH areas clustered on the blue branches (represented by the green rectangles in Figure 6a). Consequently, for the EMMA process in L2 (Layer2, shallow groundwater), we identified the optimal combination as L2Z4 and L2Z7 (shallow groundwater in Zone 4 and Zone 7, respectively, Figure 4). The discharge fractions for them were calculated as 7.17% and 92.83%, respectively (Figure 6c, Table 2). Moving on to L3 (Layer3, deep groundwater), we considered well water from L3Z3, L3Z6, and spring water from L3Z4 as the end members (Figure 6d). Their respective discharge fractions were calculated as 6.39%, 11.88%, and 81.73% (Table 2). The specific sources of lake water remain controversial, however, the description of the hydrologic connections in this study is consistent with previous studies in the region (Ma et al., 2010; Wang et al., 2013).

10.1029/2024WR037753

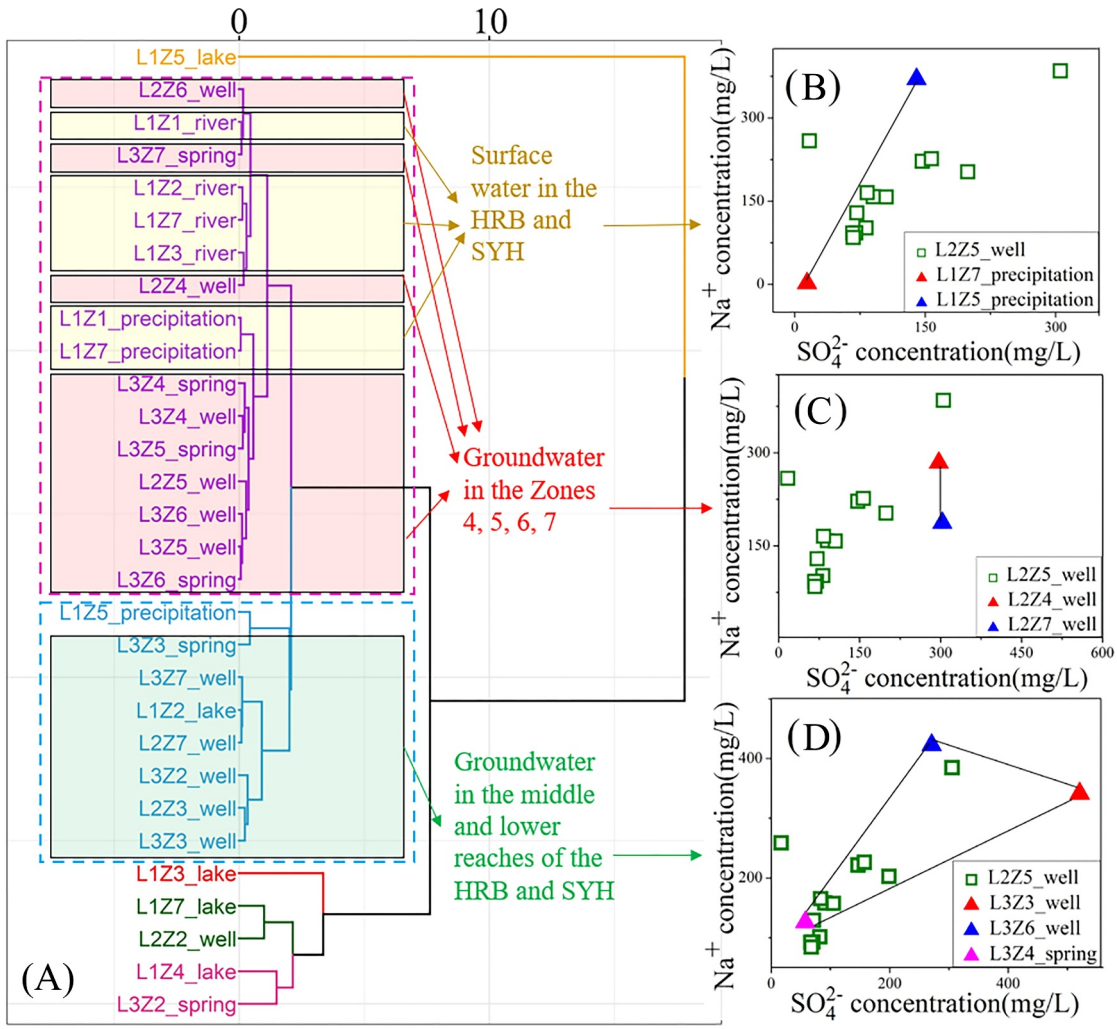

Figure 6. (a) Hierarchical cluster dendrogram illustrating the grouping of water samples, including surface water, shallow groundwater, and deep groundwater from different zones in the study area. (b) End-member mixing analysis (EMMA) for L2Z5 (shallow groundwater in Zone 5) using surface water, showing Na^+ and SO_4^{2-} concentrations. (c) EMMA for L2Z5 using shallow groundwater. (d) EMMA for L2Z5 using deep groundwater.

The saline lakes in BJD region displayed distinctive geochemical properties, preventing their grouping with other water types. The lake water in Z5 (L1Z5_lake) exhibited the longest orange branch length in the dendrogram compared to other waters (Figure 6a). These lakes were typically characterized by high levels of anions such as anions Cl^- (60,160.26 mg/L) and SO_4^{2-} (21,804.70 mg/L), as well as cations including Na^+ (60,990.44 mg/L)

Table 2
Final Tracers for End Member Mixing Analysis (EMMA), With Corresponding Discharge Fractions in the Study Area

Recharge components	EM quantity	EM	SO_4^{2-}	Na^+	Fraction (%)
layer 1 (Surface water)	2	L1Z5_precipitation	140	370	78.07
		L1Z7_precipitation	14.17	2.29	21.93
layer 2 (Shallow groundwater)	2	L2Z4_well	331.25	357.33	7.17
		L2Z7_well	302.97	187.21	92.83
layer 3 (Deep groundwater)	3	L3Z3_well	520.62	341.51	6.39
		L3Z6_well	270.76	422.7	11.88
		L3Z4_spring	57.47	126	81.73

Note. EM, end member.

Table 3
Final Models of Shallow Groundwater Solutes Identified Through Model Selection Based on AICc

Solutes	Predict variables in final model	df	logLik	AICc	Weight
Ca^{2+}	8.66* (Ca.L3) + 13* GST-1.96(L6_unexploited) + 51.48	5	−592.48	1,195.4	0.45
Mg^{2+}	−27.21* (evaporation)-17.33* (soil. coarse) + 69.68	4	−696.5	1,401.29	0.31
Na^{+}	3.42* (L2_forest) + 5.21* (L3_grassland)-20.64* (Na.L1) + 18.08* (pdensity)-110.46* (precipitation) + 230.23	7	−839.42	1,693.65	0.07
K^{+}	5.82* (evaporation) + 0.44* (K.L1) +15.83	4	−499.81	1,007.91	0.15
SO_4^{2-}	−267.7* (GST) + 72.45* (pdensity)-303.62* (precipitation)-26.36* (SO4.L1)-64.67* (soil. coarse) + 321.39	7	−911.51	1,837.82	0.24
Cl^{-}	−18.75* (Cl. L1)-27.35* (L6_unexploited)-143.56* (precipitation) + 28.34* (soil. carbon) + 219.08	6	−830.73	1,674.07	0.39
HCO_3^{-}	5.09* (pdensity)-24.81* (soil. coarse) + 264.77	4	−837.95	1,684.18	0.09

Note. df = Degrees of freedom. Weight = Akaike weight.

and K^{+} (4,641.87 mg/L) (Table S2 in Supporting Information S1). In comparison, these solute concentrations were lower in L1Z4 (Gurinai areas in HRB, Figure 4), measuring at 4,987.00, 3,912.57, 5,074.86, and 184.20 mg/L, respectively (Table S2 in Supporting Information S1).

4.2. Tracking Footprints of SGW Solutes

The top models for mapping solutes in the SGW were selected through model selection using all possible combinations of explanatory variables. Top models with the lowest AICc values are presented in Table 3, which indicates their superior ability to fit the data compared to other models with high AICc values. The hydraulic interactions between the surface and deep groundwater layers are driving the SGW solute dynamics. We found Na^{+}, SO_4^{2-}, and Cl^{-} ions in the SGW were negatively correlated with the corresponding ions in surface water. Their correlation coefficients were −20.64, −26.36, and −18.5, respectively (Table 3). A non-linear relationship exhibited between Na^{+}, SO_4^{2-} in SGW and the corresponding ions in surface water, whereas Cl^{-} featured a weak linear negative correlation with Cl^{-} in surface water ($P < 0.05$*, Figure 7, Table S3 in Supporting Information S1). Climate variables also had significant impacts on solute, with ground surface temperature (GST) being a significant factor in the final models for Ca^{2+} and SO_4^{2-}, evaporation for Mg^{2+} and K^{+}, and precipitation for Na^{+}, SO_4^{2-}, and Cl^{-} (Table 3). GST was significantly correlated with HCO_3^{-} in SGW ($p < 0.05$*) and showed a first peak when the annual average GST was 3°C, reaching a steady state after the GST exceeded 7°C (Table S3 in Supporting Information S1, Figure 7).

The human influence was also observed to be pronounced, as evidenced by the positive correlations between "population density" and Na^{+}, SO_4^{2-}, and HCO_3^{-} in the final models (Table 3). This suggests that human activities could be the key source of these solutes. Nonlinear analysis of SO_4^{2-} showed a significant positive correlation with "population density" ($p < 0.05$*, Table S3 in Supporting Information S1, Figure 7). "Unexploited area" was identified as an explanatory factor for the final models of Ca^{2+} and Cl^{-}, with coefficients of −1.96 and −27.35, respectively (Table 3). In terms of soil properties, soil coarse fragments were negatively correlated with Mg^{2+}, SO_4^{2-}, and HCO_3^{-} (Table 3, Figure 7), while soil carbon showed a positive correlation with Cl^{-}.

4.3. Improvements of Mapping Solute With Predictor Variables

4.3.1. Maps With Only Topography Variables

Both SDMs obtained (topography-only, and all predictor variables) presented high mapping accuracy, with high AUC (area under the curve, 0.88–0.99), PCC (proportion of correctly predicted occurrences, 0.83–0.98), and high sensitivity (0.99–1) (Table S4 in Supporting Information S1). The output maps of solutes that considered solely topographical information by the SDMs were presented in the first columns of Figures 8 and 9. This means that only topography data was used to predict the spatial patterns of shallow groundwater solutes. Our findings indicated that all other ions showed a negative correlation with increasing altitude, except for Mg^{2+} and SO_4^{2-}

10.1029/2024WR037753

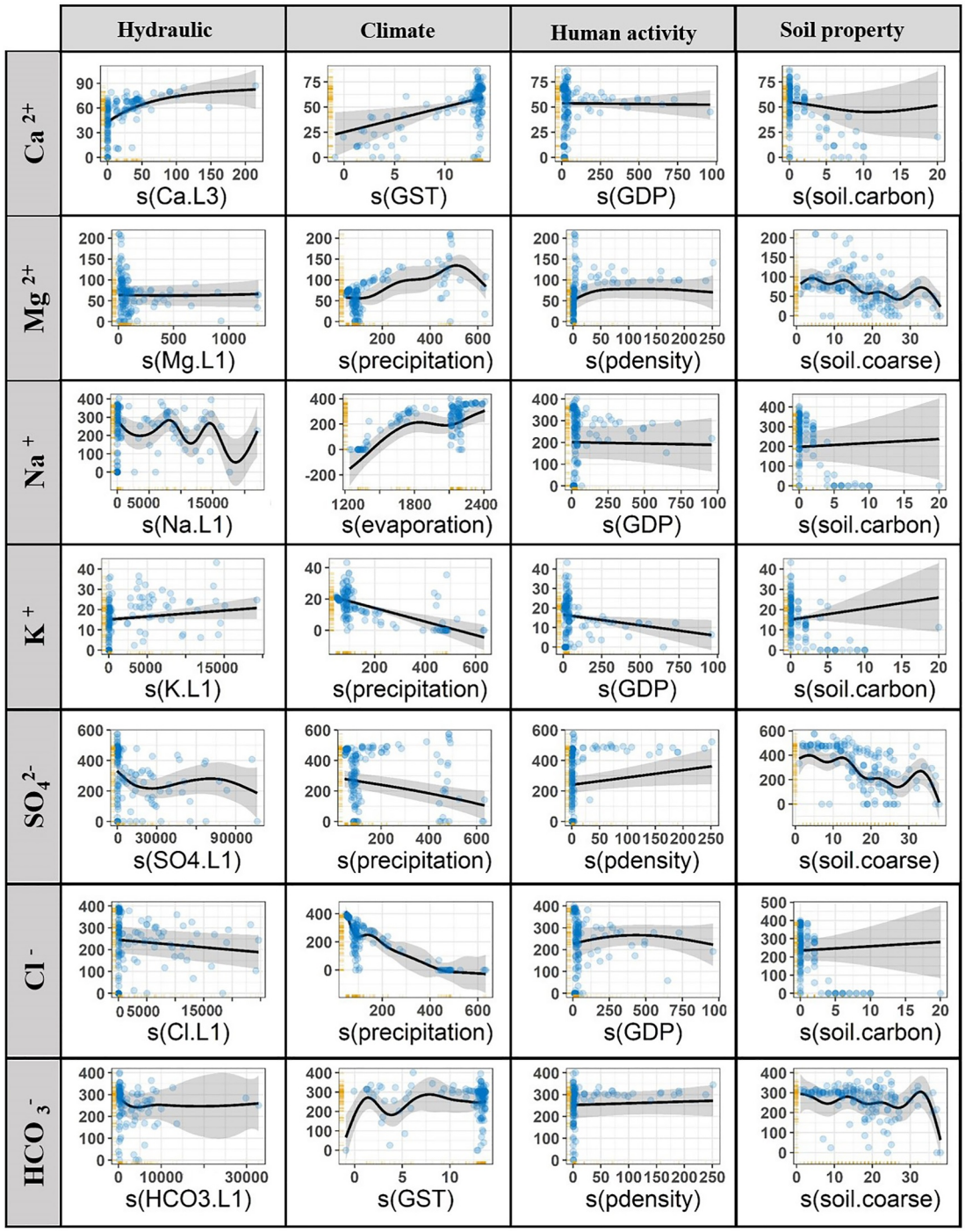

Figure 7. Smoothed fits of relationships between shallow groundwater solutes and four categories of explanatory variables: hydraulic (first column), climate (second column), human activity (third column), and soil properties (fourth column). Each panel shows the influence of a specific variable. Shaded areas represent confidence intervals around the fitted models.

which displayed high values in some high mountain areas. This pattern was particularly noticeable in the Yabulai and Longshou Mountains located on the southern edge of BJD. In the shallow groundwater of the BJD's lake area, Mg^{2+}, K^{+}, and SO_4^{2-} were predicted to be hotspots, while other ions corresponded to low-value areas.

10.1029/2024WR037753

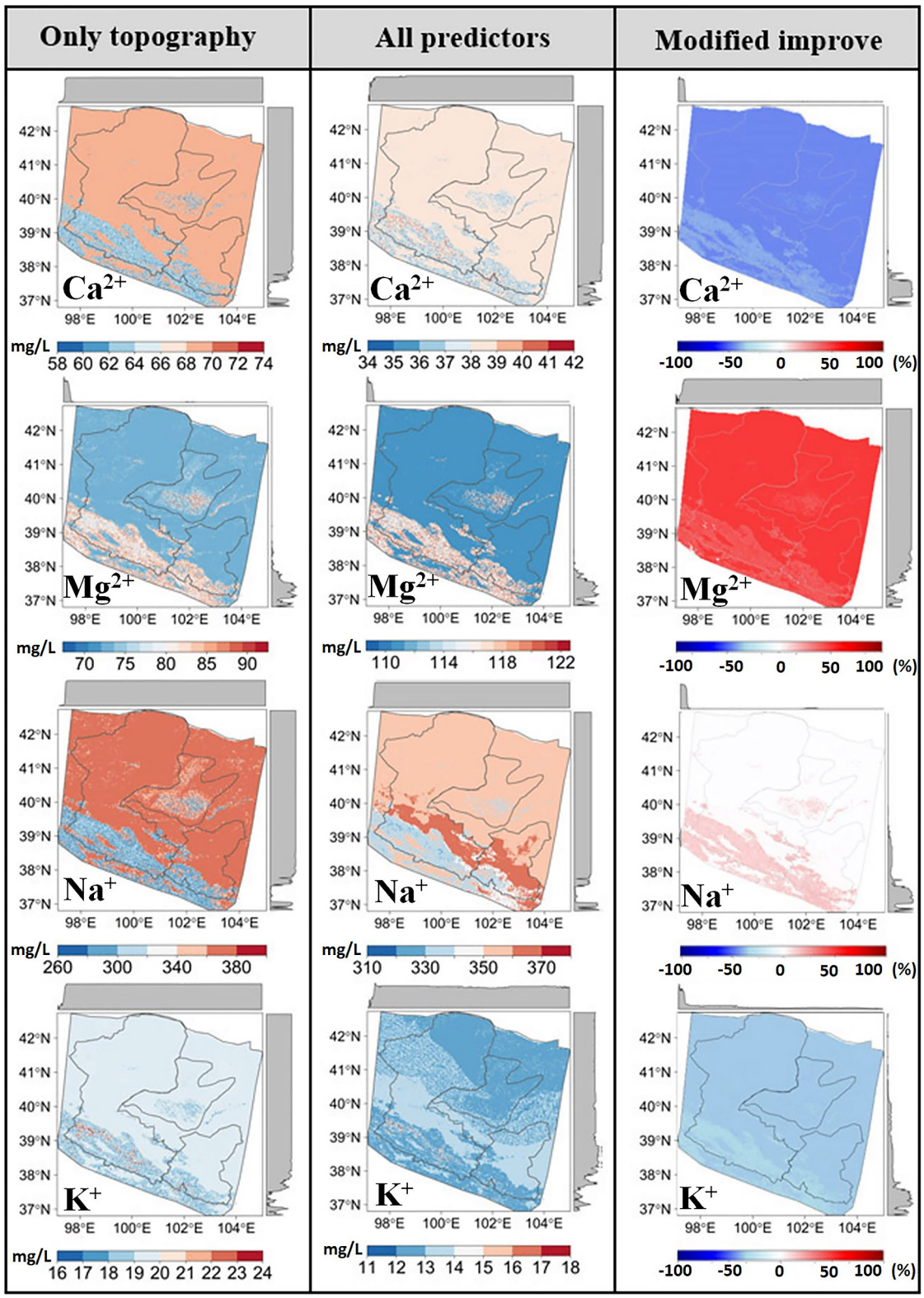

Figure 8. Comparative analysis of predicted spatial distributions of shallow groundwater cations using species distribution models (SDMs). First column: Predictions based on topography alone using SDMs (unit: mg/L). Second column: Predictions from the final SDM model incorporating all predictors (unit: mg/L). Third column: Difference maps illustrating the improvements, calculated by subtracting the first column from the second, then dividing by the first column to express the change as a percentage. Positive values represent overestimation, while negative values indicate underestimation.

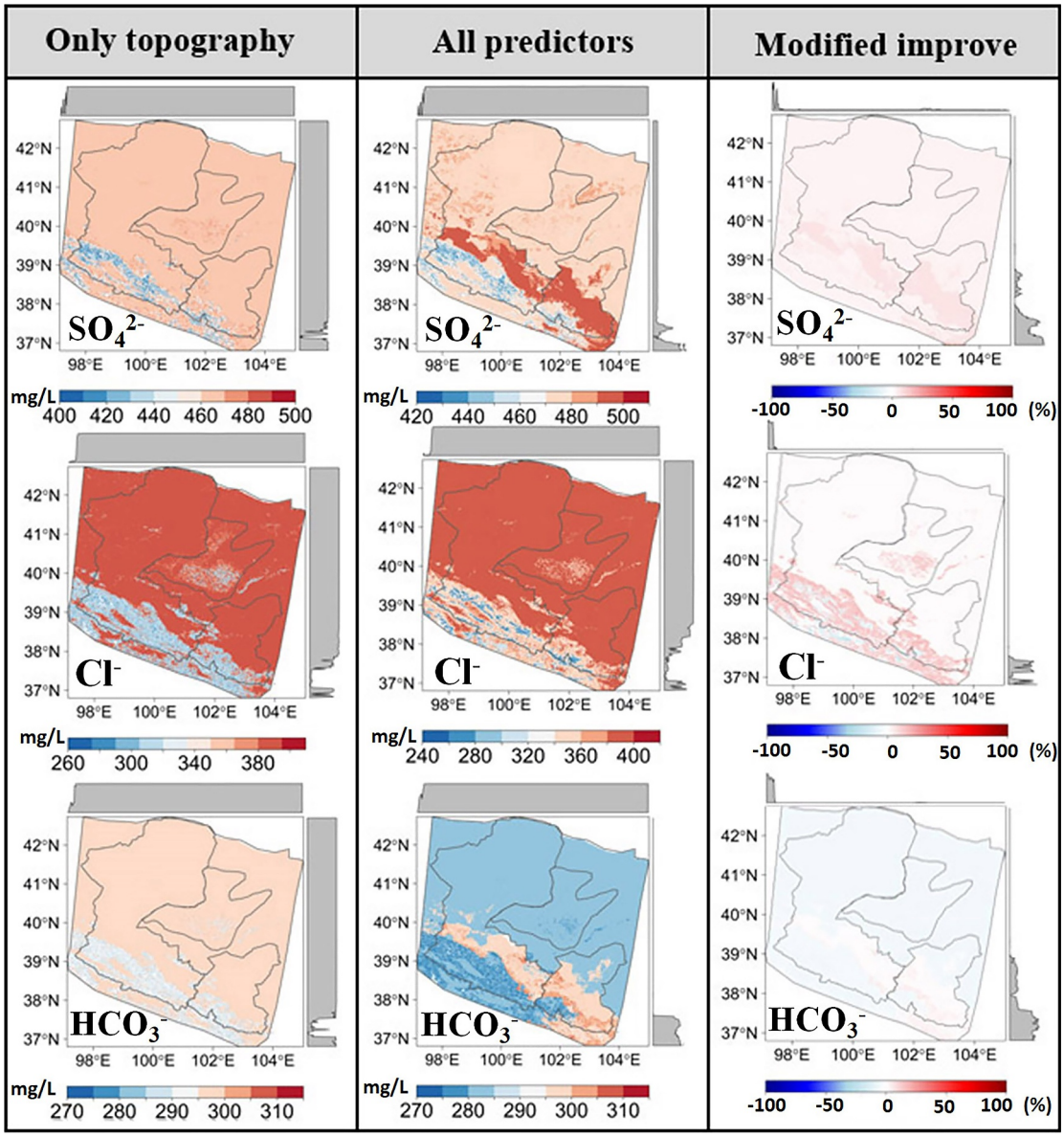

Figure 9. Comparative analysis of predicted spatial distributions of shallow groundwater anions using species distribution models (SDMs). First column: Predictions based on topography alone using SDMs (unit: mg/L). Second column: Predictions from the final SDM model incorporating all predictors (unit: mg/L). Third column: Difference maps illustrating the improvements, calculated by subtracting the first column from the second, then dividing by the first column to express the change as a percentage. Positive values represent overestimation, while negative values indicate underestimation.

4.3.2. Final Maps With All Predictor Variables

We integrated the SGW solute variables represented by the top model in Table 3 into SDMs for mapping the solute footprints. The resultant maps and improvements are displayed in the second and third columns of Figures 8 and 9, respectively. The output maps indicated a narrower predicted range for Ca^{2+}, K^{+}, and Cl^{-}, but a broader predicted range for Mg^{2+} and SO_4^{2-}. However, there were no significant changes in the variations predicted for Na^{+} and HCO_3^{-} (Figures 8 and 9). The incorporation of more intricate variable relationships facilitated the identification of additional hotspots in the predictions. The differences observed between the maps highlighted the importance of the additional variables in predicting the spatial patterns of shallow groundwater solutes. For instance, the predictor variables of Ca^{2+} can be represented by "$Ca^{2+} \sim 8.66*$ (Ca^{2+} . L3) +13* GST-1.96 (L6_ unexploited) + 51.48," and the narrower predicted range in the final map demonstrated the crucial role of "unexploited area" ($p < 2e{-}16$). The SO_4^{2-} is an important indicator of groundwater pollution resulting from

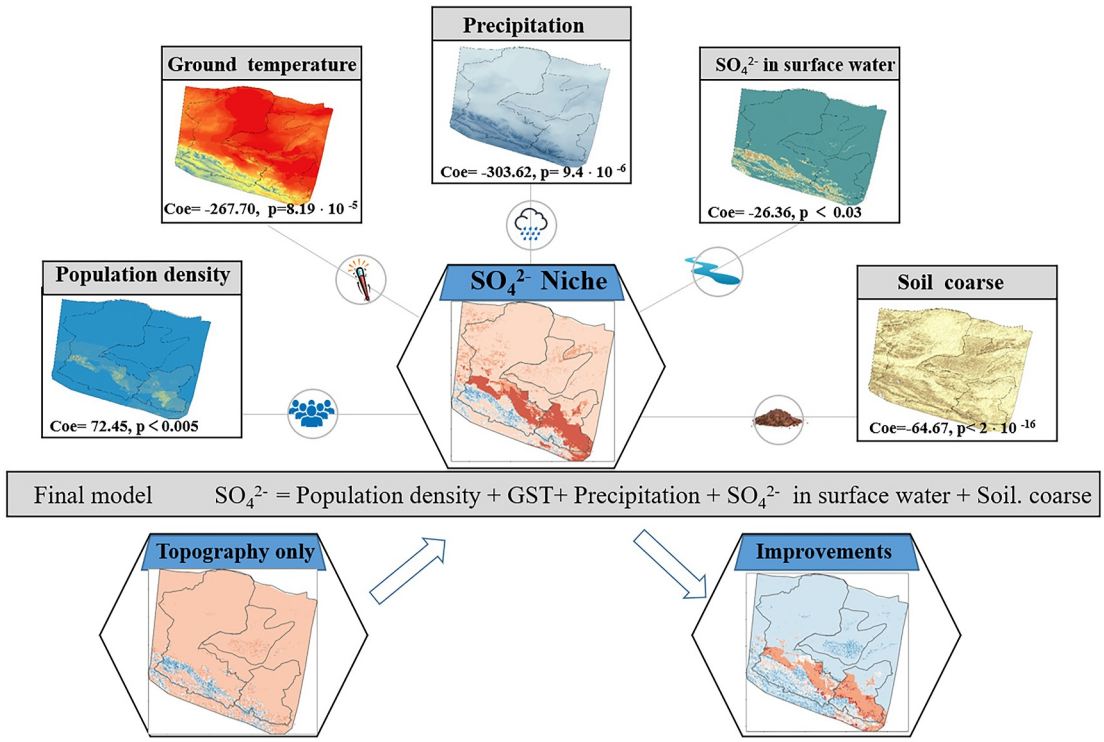

Figure 10. Improvements in spatial prediction by incorporating variables into the final models.

human activities. In this study, we identified its variables as "$SO_4^{2-} \sim 72.45$* (p density)-267.7* (GST) −303.62* (precipitation)-26.36* (SO_4^{2-} . L1)-64.67* (soil. coarse)" (Table 3, Figure 10). Our analysis revealed that, except for a positive correlation with "population density," all other factors showed a negative correlation with SO_4^{2-}. Therefore, the increased trend of SO_4^{2-} in the final maps demonstrated the significant impact of human activities ($p < 0.005$). Additionally, our investigation uncovered new hotspot points of SO_4^{2-} in big cities such as Jiuquan, Jiayuguan, Zhangye, and other areas along the Hexi Corridor (Figure 10).

5. Discussion

5.1. A Hydrologically Enhanced SDM With Multi-Dimensional Predictors

A key strength of our hydrologically enhanced SDM (Species Distribution Model) is its minimal dependence on field data, making it particularly suitable for environmental samples (e.g., groundwater, soil) that are challenging to collect in natural settings. For the sparse distribution of water sampling points under extreme drought conditions in this study, Option 1 (SDM) significantly outperforms Option 2 (traditional interpolation methods, such as Kriging) (Figure 2). Compared to traditional Kriging, SDMs offer more flexibility in the requirements for input biological observation data. The applicability of Kriging is constrained by the density and uniformity of the sample points, making it difficult to accommodate the sparse data characteristic present in this study. For instance, Figure 2c shows a typical example of Kriging application (Lv, 2019), which is based on high-density sampling with a 2 km × 2 km grid over a 1,138 km^2 area. Such data distribution is unsuitable for the needs of research like this study. In contrast, SDMs not only possess high compatibility and strong generalization ability, but also require much less spatial samples and greater flexibility in the uniformity of input data (Figure 2b). However, previous studies suggested that the minimum number of sampling points typically required to build a SDM is 10 (Pearson et al., 2007; van Proosdij et al., 2016). When the sample size is very small (e.g., only one observation point), although the model can predict similar areas based on environmental variables, the results tend to have high uncertainties, resulting in a limited practical value. In actual applications, to better capture the relationship between species distribution and environmental variables, and to increase prediction reliability, 10 to 30 observation points are generally needed (Hernandez et al., 2006). It is important to note that this minimum requirement is influenced by various factors, including the size and environment of study area, the distribution characteristics of species, and the complexity of

10.1029/2024WR037753

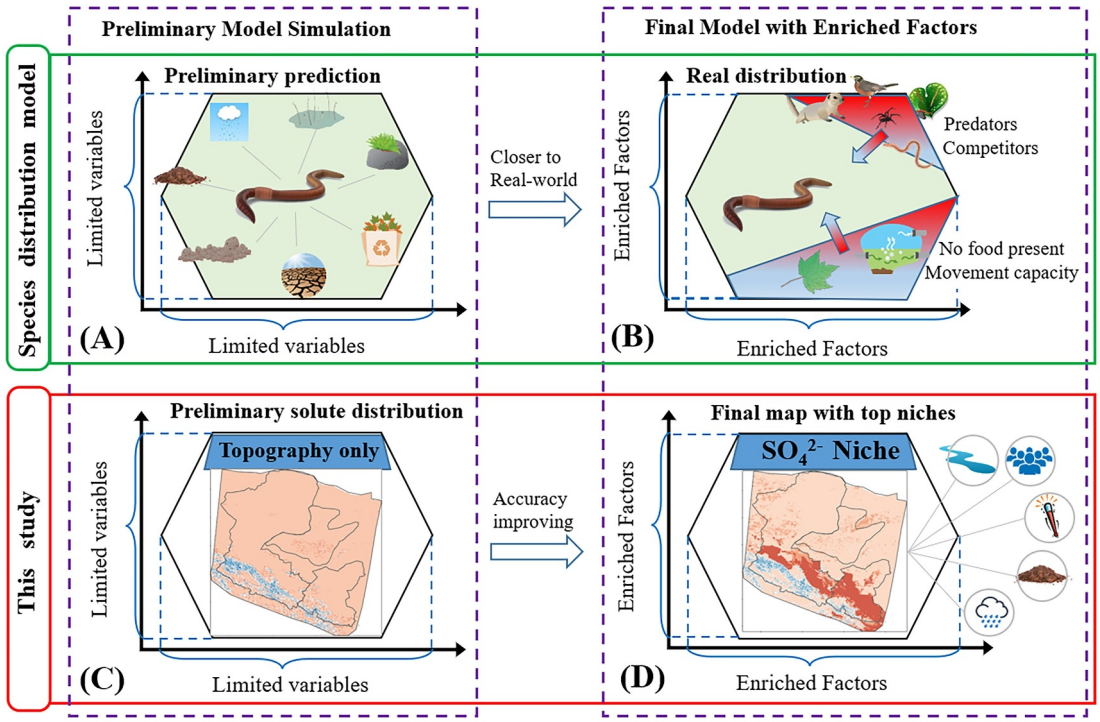

Figure 11. Conceptual framework showing the integration of enriched predictor variables in the SDM (Species Distribution Model) to enhance spatial accuracy. (a) and (b) Demonstrate that incorporating additional predictor variables in biological models can bring predictions closer to real-world distributions. (c) and (d) Show that this study enhances prediction accuracy by considering factors influencing groundwater solute dynamics.

the model. For species with a wide distribution or in larger study areas, more sample points are needed, while simpler distribution patterns may require fewer data. Additionally, when data are insufficient, model performance can be enhanced by generating background points or incorporating expert knowledge.

Another key advantage of SDM is its ability to integrate multiple factors. Specifically, SDM can effectively utilize existing remote sensing data and spatiotemporal continuous map products, such as GIS (Figure 2e), to establish a comprehensive framework for more accurate predictions. In contrast, Kriging relies exclusively on the spatial autocorrelation between sample points and cannot incorporate environmental factors (Figure 2f). By integrating multiple factors, SDM produce predictions that more accurately reflect real-world conditions, making them highly applicable to a variety of environmental studies. Building on this strength, our study adopted a multi-dimensional factor system to examine the transport and spatial distribution of solutes in shallow groundwater within arid regions. Taking the example of SDM used for predicting biological distribution, Figure 11a illustrates the potential range of species distribution under optimal conditions, driven by abiotic factors such as climate and soil characteristics. However, in reality, their actual distribution is shaped by factors like predators, competitors, food availability, and movement capacity (Figure 11b). This highlights the importance of incorporating a broader range of influencing factors into models to improve prediction accuracy. Similarly, just as SDM improves predictions in ecological studies by considering a wide array of environmental factors, it can also be applied to groundwater research. For example, the dynamics of SGW solutes are influenced by various natural processes and human activities (Figure 1), emphasizing the need for a multi-factor approach. In our study, we analyzed the relationships among hydrology, environmental conditions, and human systems (Figures 11c and 11d) to predict the spatial patterns of SGW solutes. This multidimensional approach enabled us to better understand the mechanisms driving solute dynamics and their interactions with the surrounding environment.

While considering multiple factors is essential, the integration of hydrological connectivity is particularly crucial in groundwater models. Ignoring interactions between adjacent aquifer units or between surface water and groundwater in model algorithms can lead to significant deviations from actual solute dynamics (Woodward et al., 2016). Our approach addressed this gap by incorporating hydrological connectivity as a mechanistic constraint, improving upon traditional data-driven interpolation and correlation-based methods. This ensures that

predictions better reflect the physical reality of groundwater systems. Moreover, by integrating spatial data to inform estimates, our method avoids the rigid assumptions and computational demands of fully prescribed physics-based models. This approach represents a significant improvement over previous research by specifically targeting Gap 1 (hydraulic connection) within existing constraints. It is also crucial to accurately define aquifer boundaries when modeling the physical processes of groundwater solutes (Reilly, 1987; Yao et al., 2015). While the origin of many lakes in the BJD remains a subject of debate (Wang & Zhou, 2018), our study benefited from previous research to guide geological surveys and determine potential groundwater flow directions. Additionally, hydrograph separation was used to quantify water supply between different hydrological units, and the use of chemical signals in EMMA complemented complex geological and hydrogeological cross-sectional data.

The dynamics of SGW solutes in arid regions are being influenced by various extrinsic factors, including human activities, vegetation, climate, and soil conditions (Figure 1). These factors exhibit great spatial heterogeneity, posing challenges in understanding their impacts on SGW. However, with the advancements in the fields of remote sensing and GIS (Xie et al., 2024), monitoring the dynamics of these drivers at a large-scale is now possible through direct or indirect means. For instance, indicators such as GDP and population density can provide useful insights into human-induced overexploitation and pollution of groundwater, particularly in arid and semi-arid regions (Mukherjee et al., 2020). In this study, we incorporated as many variables as possible in the species distribution models from the existing available GIS/RS data products to simulate the environmental layers (Li et al., 2020). Notably, this study also incorporated soil physical and chemical data, which have been largely overlooked in previous studies when examining groundwater dynamics.

5.2. Enhanced Mapping Accuracy of Shallow Groundwater Solutes Through SDMs

Spatial monitoring of groundwater chemistry is essential for evaluating and sustainably managing water resources, as well as for understanding hydrogeochemical cycles. Mapping the spatial distributions of groundwater chemistries helps link these patterns to potential health risks, such as diseases caused by water contamination (Ma et al., 2022). Traditionally, spatial interpolation methods such as kriging, inverse distance weighting, and Thiessen polygons have been widely used in groundwater studies. These methods are based on the first law of geography, which asserts that geographically closer entities tend to exhibit stronger relationships (Tobler, 1970). However, traditional techniques often overlook key characteristics of groundwater systems, such as hydrological connectivity, heterogeneous flow paths, and dynamic solute transport. To address these challenges, this study integrates hydrological connectivity as a mechanistic constraint, offering a novel approach to improving spatial predictions of groundwater chemistry, especially when data are sparse. While methods like Universal Kriging, Moving Window Kriging, and geographically weighted regression have been developed to address spatial heterogeneity, they still struggle with complex geographical conditions. These methods work well in uniform terrains but tend to yield inaccurate predictions in regions with significant spatial heterogeneity, such as arid areas. In such regions, the varied topography strongly influences the migration of SGW solutes, rendering traditional interpolation methods fail to capture the complexities of solute dynamics. To address these challenges, we employed SDMs, originally developed for ecological applications, to enhance the prediction of groundwater chemistry distribution. SDMs have demonstrated success across various disciplines, including ecological and environmental studies (Li et al., 2022).

Unlike traditional interpolation methods, SDMs allow for the incorporation of both natural and anthropogenic variables, providing more accurate spatial predictions of solute distributions and unveiling new insights into hydrogeochemical cycles in arid environments. Our results emphasize several key quantitative relationships that significantly influence the spatial patterns of shallow groundwater solutes. For instance, the predicted spatial distribution of Ca^{2+} revealed that "unexploited area" played a critical role in its distribution. This variable was instrumental in narrowing the predicted range of Ca^{2+} in the final maps (Figures 8 and 9), demonstrating the significant impact of land use on geochemical cycling. Similarly, the results for SO_4^{2-}, a key indicator of groundwater pollution from human activities, revealed that its distribution was primarily driven by population density (positive correlation), while factors such as precipitation, GST, and soil type (coarse) exhibited negative correlations. The inclusion of these variables in the final model enabled the identification of new SO_4^{2-} hotspots in urban areas such as Jiuquan, Jiayuguan, Zhangye, and other regions along the Hexi Corridor (Figure 10), highlighting the critical role of human activities in driving solute dynamics. Additionally, the predicted ranges for other solutes, such as Mg^{2+}, K^+, Cl^-, and HCO_3^-, revealed varying spatial patterns influenced by their respective

predictor variables (Figures 8 and 9). For example, the broader predicted range for Mg^{2+} and SO_4^{2-} underscores the impact of geological and anthropogenic factors, while the narrower predicted range for Ca^{2+}, K^+, and Cl^- reflects the localized nature of their driving factors. These findings demonstrate the ability of SDMs to capture intricate variable relationships and highlight the spatial heterogeneity of groundwater chemistry in arid regions. By integrating these key relationships, our study advances the understanding of groundwater solute dynamics and highlights the potential of SDMs in identifying the complex interactions among driving factors.

SDMs have demonstrated strong predictive power across diverse research fields, including human emotions (Li et al., 2021), infectious diseases (Harrigan et al., 2014), municipal planning (Clemente et al., 2019), and heavy metal pollution (Li et al., 2020). Building on these successes, the modified SDMs employed in this study provide a novel approach to investigating the spatial distribution and driving mechanisms of solutes in SGW within arid regions from an ecological perspective. However, the framework presented here has some limitations, particularly regarding the spatial resolution of available data sets. For instance, while the demonstration region is relatively large, the coarse resolution of input data—such as soil and land-use information—may constrain the accuracy of fine-scale predictions. Future work should focus on integrating higher-resolution data sets and exploring the scalability of this approach to smaller or more heterogeneous regions, which would enhance our understanding of both spatial and temporal variations in solute dynamics. A key challenge for further development of SDMs in groundwater prediction lies in the need for predictor variables to be represented in continuous raster format. Currently, obtaining large-scale raster data for variables such as hydrogeological conditions, bedrock type, fracture density, and depth to groundwater is difficult. The inclusion of such hydrological data could significantly improve the accuracy and predictive capabilities of SDMs.

6. Conclusion

Mapping shallow groundwater solute footprints in arid regions has presented challenges due to its uneven spatial distributions, complex transport mechanisms, and difficulties in data acquisition. In this study, we addressed these challenges by proposing a novel approach that integrates SDMs from ecology with traditional hydrological models, machine learning techniques, remote sensing, and GIS methods. Using this interdisciplinary approach, we collected comprehensive predictor variables that encompass both natural processes and human activities. This novel framework enabled us to predict solute concentrations in groundwater by adapting the SDMs approach traditionally used in ecology for modeling species suitability. Our findings demonstrated that incorporating hydraulic connections between the surface layer and deep groundwater across adjacent catchments significantly improved the accuracy of solute dynamics predictions. Additionally, we observed negative correlations between Na^+, SO_4^{2-}, and Cl^- in the SGW and their counterparts in surface water. Moreover, integrating solute variables into the SDMs facilitated our identification of additional hotspots. For instance, incorporating parameters such as "unexploited area" narrowed the prediction range for Ca^{2+}, while "population density" revealed hotspots of SO_4^{2-} in Hexi Corridor. This study highlights the value of combining conventional hydrological methods with SDMs from an ecological perspective to improve the accuracy of mapping SGW solutes in large-scale arid regions. This approach demonstrates transferability to watersheds with diverse land-use patterns and groundwater regimes, offering a robust framework for cross-system water quality management.

Data Availability Statement

The topography-only and final maps of shallow groundwater and deep groundwater, their improvements, the original data set of water chemistry, as well as the related R script in the study are available at Zenodo at https://zenodo.org/records/12735837.

Acknowledgments
This work was supported by the Natural Science Foundation of China (42371311), the Science and Technology Department of Jilin Province (YDZJ202501ZYTS495) and Gansu Province (23JRRA1122), the Strategic Priority Research Program of the Chinese Academy of Sciences, China (XDA28020501), the Black Soil Granary Technology and Talent Special Fund (E455S30401), Henan Province International Science & Technology Cooperation Program (252102521024), and the Natural Science Foundation of Henan Province (232300420164).

References

Amini, M., Abbaspour, K. C., Berg, M., Winkel, L., Hug, S. J., Hoehn, E., et al. (2008). Statistical modeling of global geogenic arsenic contamination in groundwater. *Environmental Science & Technology*, *42*(10), 3669–3675. https://doi.org/10.1021/es702859e

Barthold, F. K., Tyralla, C., Schneider, K., Vaché, K. B., Frede, H. G., & Breuer, L. (2011). How many tracers do we need for end member mixing analysis (EMMA)? A sensitivity analysis. *Water Resources Research*, *47*(8). https://doi.org/10.1029/2011WR010604

Ben Messaoud, R., Lachaal, F., Leduc, C., & Mlayah, A. (2021). Discharge of treated wastewater: Hydrodynamic and hydrogeochemical impacts on the Kairouan plain aquifer (Central Tunisia). *Environmental Earth Sciences*, *80*(10), 1–15. https://doi.org/10.1007/s12665-021-09667-7

Chen, J., Zhao, X., Sheng, X., Dong, H., Rao, W., & Su, Z. (2006). Formation mechanisms of megadunes and lakes in the Badain Jaran Desert, Inner Mongolia. *Chinese Science Bulletin*, *51*(24), 3026–3034. https://doi.org/10.1007/s11434-006-2196-8

Cheng, Y., Bhoot, V. N., Kumbier, K., Sison-Mangus, M. P., Brown, J. B., Kudela, R., & Newcomer, M. E. (2021). A novel random forest approach to revealing interactions and controls on chlorophyll concentration and bacterial communities during coastal phytoplankton blooms. *Scientific Reports*, *11*(1), 19944. https://doi.org/10.1038/s41598-021-98110-9

Christophersen, N., Neal, C., Hooper, R. P., Vogt, R. D., & Andersen, S. (1990). Modelling streamwater chemistry as a mixture of soilwater end-members—A step towards second-generation acidification models. *Journal of Hydrology*, *116*(1–4), 307–320. https://doi.org/10.1016/0022-1694(90)90130-P

Clemente, P., Calvache, M., Antunes, P., Santos, R., Cerdeira, J. O., & Martins, M. J. (2019). Combining social media photographs and species distribution models to map cultural ecosystem services: The case of a Natural Park in Portugal. *Ecological Indicators*, *96*, 59–68. https://doi.org/10.1016/j.ecolind.2018.08.043

Dandge, K., & Patil, S. (2022). Spatial distribution of ground water quality index using remote sensing and GIS techniques. *Applied Water Science*, *12*(1), 1–18. https://doi.org/10.1007/s13201-021-01546-7

Ding, H., & Wang, G. (2007). Study on the formation mechanism of the lakes in the Badain Fijaran Desert. *Arid Zone Research*, *24*(1), 1–7.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology Evolution and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Enguehard, L., Falco, N., Schmutz, M., Newcomer, M. E., Ladau, J., Brown, J. B., et al. (2022). Machine-learning functional zonation approach for characterizing terrestrial–aquatic interfaces: Application to Lake Erie. *Remote Sensing*, *14*(14), 3285. https://doi.org/10.3390/rs14143285

Fendorf, S., Michael, H. A., & van Geen, A. (2010). Spatial and temporal variations of groundwater arsenic in South and Southeast Asia. *Science*, *328*(5982), 1123–1127. https://doi.org/10.1126/science.1172974

Giam, X., Olden, J. D., & Simberloff, D. (2018). Impact of coal mining on stream biodiversity in the US and its regulatory implications. *Nature Sustainability*, *1*(4), 176–183. https://doi.org/10.1038/s41893-018-0048-6

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Harrigan, R. J., Thomassen, H. A., Buermann, W., & Smith, T. B. (2014). A continental risk assessment of West Nile virus under climate change. *Global Change Biology*, *20*(8), 2417–2425. https://doi.org/10.1111/gcb.12534

Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, *29*(5), 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

Hijmans, R. J., & Elith, J. (2013). *Species distribution modeling with R*. R Cran Project.

Kaushal, S. S., Likens, G. E., Pace, M. L., Utz, R. M., Haq, S., Gorman, J., & Grese, M. (2018). Freshwater salinization syndrome on a continental scale. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(4), E574–E583. https://doi.org/10.1073/pnas.1711234115

Lapworth, D., MacDonald, A., Tijani, M., Darling, W., Gooddy, D., Bonsor, H., & Araguás-Araguás, L. (2013). Residence times of shallow groundwater in West Africa: Implications for hydrogeology and resilience to future changes in climate. *Hydrogeology Journal*, *21*(3), 673–686. https://doi.org/10.1007/s10040-012-0925-4

Li, J., Li, Z., Brandis, K. J., Bu, J., Sun, Z., Yu, Q., & Ramp, D. (2020). Tracing geochemical pollutants in stream water and soil from mining activity in an alpine catchment. *Chemosphere*, *242*, 125167. https://doi.org/10.1016/j.chemosphere.2019.125167

Li, J., Xie, Z., Qiu, X., Yu, Q., Bu, J., Sun, Z., et al. (2022). Heavy metal habitat: A novel framework for mapping heavy metal contamination over large-scale catchment with a species distribution model. *Water Research*, *226*, 119310. https://doi.org/10.1016/j.watres.2022.119310

Li, Y., Fei, T., Huang, Y., Li, J., Li, X., Zhang, F., et al. (2021). Emotional habitat: Mapping the global geographic distribution of human emotion with physical environmental factors using a species distribution model. *International Journal of Geographical Information Science*, *35*(2), 227–249. https://doi.org/10.1080/13658816.2020.1755040

Lv, J. (2019). Multivariate receptor models and robust geostatistics to estimate source apportionment of heavy metals in soils. *Environmental Pollution*, *244*, 72–83. https://doi.org/10.1016/j.envpol.2018.09.147

Ma, J., & Edmunds, W. M. (2006). Groundwater and lake evolution in the Badain Jaran Desert ecosystem, Inner Mongolia. *Hydrogeology Journal*, *14*(7), 1231–1243. https://doi.org/10.1007/s10040-006-0045-0

Ma, J., Pan, F., Chen, L., Edmunds, W. M., Ding, Z., He, J., et al. (2010). Isotopic and geochemical evidence of recharge sources and water quality in the Quaternary aquifer beneath Jinchang city, NW China. *Applied Geochemistry*, *25*(7), 996–1007. https://doi.org/10.1016/j.apgeochem.2010.04.006

Ma, R., Yan, M., Han, P., Wang, T., Li, B., Zhou, S., et al. (2022). Deficiency and excess of groundwater iodine and their health associations. *Nature Communications*, *13*(1), 7354. https://doi.org/10.1038/s41467-022-35042-6

Maavara, T., Siirila-Woodburn, E. R., Maina, F., Maxwell, R. M., Sample, J. E., Chadwick, K. D., et al. (2021). Modeling geogenic and atmospheric nitrogen through the East River watershed, Colorado Rocky Mountains. *PLoS One*, *16*(3), e0247907. https://doi.org/10.1371/journal.pone.0247907

Mukherjee, I., Singh, U. K., Singh, R. P., Kumari, D., Jha, P. K., & Mehta, P. (2020). Characterization of heavy metal pollution in an anthropogenically and geologically influenced semi-arid region of east India and assessment of ecological and human health risks. *Science of the Total Environment*, *705*, 135801. https://doi.org/10.1016/j.scitotenv.2019.135801

Nolan, J., & Weber, K. A. (2015). Natural uranium contamination in major US aquifers linked to nitrate. *Environmental Science and Technology Letters*, *2*(8), 215–220. https://doi.org/10.1021/acs.estlett.5b00174

Pearson, R. G., Raxworthy, C. J., Nakamura, M., & Townsend Peterson, A. J. (2007). Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *Journal of Biogeography*, *34*(1), 102–117. https://doi.org/10.1111/j.1365-2699.2006.01594.x

Podgorski, J., & Berg, M. (2020). Global threat of arsenic in groundwater. *Science*, *368*(6493), 845–850. https://doi.org/10.1126/science.aba1510

Podgorski, J., & Berg, M. (2022). Global analysis and prediction of fluoride in groundwater. *Nature Communications*, *13*(1), 4232. https://doi.org/10.1038/s41467-022-31940-x

Podgorski, J. E., Labhasetwar, P., Saha, D., & Berg, M. (2018). Prediction modeling and mapping of groundwater fluoride contamination throughout India. *Environmental Science & Technology*, *52*(17), 9889–9898. https://doi.org/10.1021/acs.est.8b01679

Reilly, T. E. (1987). *A conceptual framework for ground-water solute-transport studies with emphasis on physical mechanisms of solute movement*. Department of the Interior, US Geological Survey.

Rogers, D. B., Newcomer, M. E., Raberg, J. H., Dwivedi, D., Steefel, C., Bouskill, N., et al. (2021). Modeling the impact of riparian hollows on river corridor nitrogen exports. *Frontiers in Water*, *3*, 590314. https://doi.org/10.3389/frwa.2021.590314

Sadoff, C. W., Borgomeo, E., & Uhlenbrook, S. (2020). Rethinking water for SDG 6. *Nature Sustainability*, *3*(5), 346–347. https://doi.org/10.1038/s41893-020-0530-9

Setiawan, I., Morgan, L., Doscher, C., Ng, K., & Bosserelle, A. (2022). Mapping shallow groundwater salinity in a coastal urban setting to assess exposure of municipal assets. *Journal of Hydrology: Regional Studies*, *40*, 100999. https://doi.org/10.1016/j.ejrh.2022.100999

Shouse, P. J., Goldberg, S., Skaggs, T. H., Soppe, R., & Ayars, J. E. (2010). Changes in spatial and temporal variability of SAR affected by shallow groundwater management of an irrigated field, California. *Agricultural Water Management*, *97*(5), 673–680. https://doi.org/10.1016/j.agwat.2009.12.008

Smith, R., Knight, R., & Fendorf, S. (2018). Overpumping leads to California groundwater arsenic threat. *Nature Communications*, *9*(1), 2089. https://doi.org/10.1038/s41467-018-04475-3

Thaw, M., GebreEgziabher, M., Villafañe-Pagán, J. Y., & Jasechko, S. (2022). Modern groundwater reaches deeper depths in heavily pumped aquifer systems. *Nature Communications*, *13*(1), 5263. https://doi.org/10.1038/s41467-022-32954-1

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(sup1), 234–240. https://doi.org/10.2307/143141

van Proosdij, A. S., Sosef, M. S., Wieringa, J. J., & Raes, N. J. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, *39*(6), 542–552. https://doi.org/10.1111/ecog.01509

Waller, R. M. (1994). *Ground water and the rural homeowner*. US Department of the Interior, US Geological Survey.

Wang, L., Dong, Y., & Xu, Z. (2017). A synthesis of hydrochemistry with an integrated conceptual model for groundwater in the Hexi Corridor, northwestern China. *Journal of Asian Earth Sciences*, *146*, 20–29. https://doi.org/10.1016/j.jseaes.2017.04.023

Wang, L. a., Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal*, *4*(3), 212–219. https://doi.org/10.1016/j.cj.2016.01.008

Wang, P., Yu, J., Zhang, Y., & Liu, C. (2013). Groundwater recharge and hydrogeochemical evolution in the Ejina Basin, northwest China. *Journal of Hydrology*, *476*, 72–86. https://doi.org/10.1016/j.jhydrol.2012.10.049

Wang, X. S., & Zhou, Y. (2018). Investigating the mysteries of groundwater in the Badain Jaran Desert, China. *Hydrogeology Journal*, *26*(5), 1639–1655. https://doi.org/10.1007/s10040-018-1750-1

Woodward, S. J., Wöhling, T., & Stenger, R. (2016). Uncertainty in the modelling of spatial and temporal patterns of shallow groundwater flow paths: The role of geological and hydrological site information. *Journal of Hydrology*, *534*, 680–694. https://doi.org/10.1016/j.jhydrol.2016.01.045

Wu, J., Xu, N., Wang, Y., Zhang, W., Borthwick, A. G., & Ni, J. (2021). Global syndromes induced by changes in solutes of the world's large rivers. *Nature Communications*, *12*(1), 5940. https://doi.org/10.1038/s41467-021-26231-w

Wu, Y., Zhang, Y., Wen, X., & Su, J. (2010). Hydrologic cycle and water resource modeling for the Heihe River Basin in northwestern China. *Science*.

Xie, Z., Zhao, Y., Jiang, R., Zhang, M., Hammer, G., Chapman, S., et al. (2024). Seasonal dynamics of fallow and cropping lands in the broadacre cropping region of Australia. *Remote Sensing of Environment*, *305*, 114070. https://doi.org/10.1016/j.rse.2024.114070

Yao, Y., Zheng, C., Tian, Y., Liu, J., & Zheng, Y. (2015). Numerical modeling of regional groundwater flow in the Heihe River Basin, China: Advances and new insights. *Science China Earth Sciences*, *58*(1), 3–15. https://doi.org/10.1007/s11430-014-5033-y

Zhou, S., Huang, Y., Yu, B., & Wang, G. (2015). Effects of human activities on the eco-environment in the middle Heihe River Basin based on an extended environmental Kuznets curve model. *Ecological Engineering*, *76*, 14–26. https://doi.org/10.1016/j.ecoleng.2014.04.020

References From the Supporting Information

Barroso, M. F., Ramalhosa, M. J., Olhero, A., Antão, M., Pina, M., Guimarães, L., et al. (2015). Assessment of groundwater contamination in an agricultural peri-urban area (NW Portugal): An integrated approach. *Environmental Earth Sciences*, *73*(6), 2881–2894. https://doi.org/10.1007/s12665-014-3297-3

Chen, J., Sun, X., Gu, W., Tan, H., Rao, W., Dong, H., et al. (2012). Isotopic and hydrochemical data to restrict the origin of the groundwater in the Badain Jaran Desert, Northern China. *Geochemistry International*, *50*(5), 455–465. https://doi.org/10.1134/S0016702912030044

Feng, Y., Sun, Z., Bu, J., & Cai, H. (2017). The hydrogeochemical characteristics of the river water in the section from Bayi Glacier to Huargzangsi of the Heihe River, Qilian Mountains. *Journal of Glaciology and Geocryology*, *39*(3), 680–687.

Gao, Y., Chen, J., Zhang, C., & Yan, Y. (2011). Hydrochemical characteristics of the irrigation area in the middle reaches of the Heihe River Basin. *Arid Land Geography*.

Jung, Y.-Y., Shin, W.-J., Seo, K.-H., Koh, D.-C., Ko, K.-S., & Lee, K.-S. (2022). Spatial distributions of oxygen and hydrogen isotopes in multi-level groundwater across South Korea: A case study of mountainous regions. *Science of the Total Environment*, *812*, 151428. https://doi.org/10.1016/j.scitotenv.2021.151428

Li, Z., Song, L., Tian, Q., Luo, Z., & Li, Y. (2016). Spatial and temporal variation of chemical characteristics and source analysis of precipitation of Shiyang River Basin. *Earth and Environment*, *44*(6), 637–646. https://doi.org/10.1007/s11356-017-0504-2

Long, X., Liu, F., Zhou, X., Pi, J., Yin, W., Li, F., et al. (2021). Estimation of spatial distribution and health risk by arsenic and heavy metals in shallow groundwater around Dongting Lake plain using GIS mapping. *Chemosphere*, *269*, 128698. https://doi.org/10.1016/j.chemosphere.2020.128698

Lu, Y., Wang, N., Li, G., Li, Z., Dong, C., & Lu, J. (2010). Spatial distribution of lakes hydro-chemical types in Badain Jaran Desert. *Journal of Lake Sciences*, *22*(5), 774–782.

Ma, N., & Yang, X. (2008). Environmental isotopes and water chemistry in the Badain Jaran desert and in its southeastern adjacent areas, Inner Mongolia and their hydrological implications. *Quaternary Sciences*, *28*(4), 702–711.

Trásy, B., Kovács, J., Hatvani, I. G., Havril, T., Németh, T., Scharek, P., & Szabó, C. (2018). Assessment of the interaction between surface-and groundwater after the diversion of the inner delta of the River Danube (Hungary) using multivariate statistics. *Anthropocene*, *22*, 51–65. https://doi.org/10.1016/j.ancene.2018.05.002

Wang, W., Li, W., Cai, Y., An, Y., Shao, X., Wu, X., & Yin, D. (2021). The hydrogeochemical evolution of groundwater in the middle reaches of the Heihe River Basin. *Earth Science Frontiers*, *28*(4), 184.

Wu, X. (2018). *An investigation on groundwater origination in deserts indicated by Hydrogen and Oxygen Isotopes, taking the Badain Jaran Desert as an example*. China University of Geosciences.

Zhang, G. (2017). *The groundwater source discussion of Gurinai Lake area of Alxa in Inner Mongolia*. China University of Geosciences.

Zhang, Y. (2020). *Hydrochemical characteristics and influencing factors of reservoirs under different environmental background in Shiyang River Basin*. Northwest Normal University.