

Machine Learning-Based Groundwater Level Simulation for Industrial Water Management: A Comparative Study of ARIMA, BP-ANN and Random Forest

Jia Cuixia¹, Lu Bin¹, Pengjun Bai^{2,*}

¹Dingxi Hydrology and Water Resources Survey Center of Gansu Province, Dingxi 743000, Gansu, China

²Dingxi Institute of Water Resources Science Dingxi 743000, Gansu, China

Abstract

Timely and accurate simulation of regional groundwater level dynamics is critical to guarantee stable water supply for agricultural production and manufacturing sectors in arid semi-arid loess regions. This study adopts Variable Projection Importance (VPI) to screen feature variables controlling groundwater fluctuation, and constructs three data-driven models: Autoregressive Integrated Moving Average (ARIMA), Backpropagation Artificial Neural Network (BP-ANN) and Random Forest (RF). The three models are applied to simulate monthly groundwater dynamics across seven county-level administrative divisions of water-deficient Dingxi City with scattered agricultural processing and light manufacturing enterprises. Training and independent validation results show the overall simulation accuracy ranking follows RF > BP-ANN > ARIMA. For one-year groundwater level prediction, RF provides higher fitting precision, reliability and better regional adaptability, which can be prioritized for groundwater dynamic simulation to support industrial and agricultural water allocation in Dingxi. In terms of feature contribution, lagged historical groundwater level (L1) contributes over 80% to model outputs for most districts, followed by air humidity and precipitation; human interference factors including total groundwater exploitation and manufacturing water intake occupy low weight. Taking county-level administrative units (basic industrial water management units) as independent modeling objects, this paper compares the performance of three mainstream data-driven models, quantifies the contribution of manufacturing groundwater extraction to aquifer fluctuation, and provides a complete, operable modeling process for groundwater simulation in similar arid areas with small-scale manufacturing clusters. The research results can offer quantitative support for balancing groundwater exploitation limits among manufacturing, farming and ecological protection in Dingxi.

Keywords: Data-driven model, Groundwater level dynamics, Random Forest (RF), Backpropagation Artificial Neural Network (BP-ANN), Autoregressive Integrated Moving Average (ARIMA), Variable Projection Importance (VPI), Industrial water consumption, Groundwater simulation, Dingxi City

Received on 13 August 2026, accepted on 11 September 2026, published on 07 October 2026

Copyright © 2026 Jia Cuixia *et al.*, licensed to EAI. This is an open access article distributed under the terms of the [CC BY-NC-SA 4.0](#), which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetsis.14473

*Corresponding author. Email: 876355045@qq.com

1. Introduction

Dingxi City in central Gansu suffers severe shortage of surface water resources; groundwater is the core water source for residents' daily life, traditional agriculture and

local agricultural product manufacturing. In recent years, the expansion of potato processing, feed and small textile factories has continuously increased groundwater extraction pressure. Long-term excessive groundwater use has intensified the supply-demand conflict between manufacturing and agricultural sectors. Severe over-

extraction in partial areas has triggered persistent groundwater drawdown, regional groundwater depression cones and water quality deterioration, restricting sustainable groundwater utilization and regional industrial-economic development.

China completed the national groundwater monitoring network in 2018, deploying 20,469 national professional monitoring stations nationwide. The system realizes automatic wireless collection, real-time transmission and centralized storage of groundwater level and temperature time-series data, laying a solid data foundation for refined groundwater management oriented to industrial water scheduling [1-2]. Under this background, accurate simulation of groundwater dynamic evolution becomes a core technical demand for regional water resource planning, especially for formulating reasonable groundwater intake quotas for manufacturing enterprises. Clarifying groundwater fluctuation rules and quantifying the impact of industrial water exploitation on aquifers is practically significant for mitigating water supply risks of manufacturing clusters and realizing coordinated water utilization of production, life and ecology.

Traditional physically-based groundwater numerical simulation tools including MODFLOW, FEFLOW and Groundwater Modeling System (GMS) can accurately characterize complex groundwater circulation systems [3]. However, such models require massive hydrogeological survey data, long-term spatiotemporal monitoring datasets and repeated parameter calibration, bringing high labor and time costs. These disadvantages limit their application to short-cycle groundwater prediction for emergency industrial water scheduling.

With the popularization of hydrological big data, data-driven machine learning models have been widely adopted in groundwater research. Such models focus on statistical correlations between groundwater levels and multi-source control factors (precipitation, industrial groundwater exploitation, irrigation recharge, etc.), with low data acquisition threshold and strong nonlinear fitting capacity without complex hydrogeological mechanism deduction [4-5]. Timely capturing groundwater variation laws via data-driven algorithms can significantly improve water department efficiency in supervising groundwater intake of manufacturing factories. Widely used data-driven approaches include Multiple Linear Regression (MLR) [6], ARIMA time series model, BP-ANN neural network, RF ensemble learning and Long Short-Term Memory (LSTM) deep learning, all of which have achieved stable simulation performance in existing hydrological research [7-10].

From the perspective of industrial water management, manufacturing enterprises generally rely on groundwater as stable production water; unregulated excessive pumping by processing factories accelerates aquifer depletion. Multiple studies have adopted machine learning to quantify manufacturing groundwater consumption and predict aquifer responses under different production expansion scenarios. Relevant research has verified that multi-factor data-driven models can effectively couple industrial water

extraction and groundwater fluctuation characteristics [11-13].

Nevertheless, existing groundwater simulation research still has obvious deficiencies. Few studies take county-level administrative regions (the basic unit for industrial water quota management) as independent modeling units to compare multiple machine learning models simultaneously [14]. Most existing literatures ignore differentiated groundwater extraction characteristics of local manufacturing industries during feature screening, lacking targeted model analysis matching industrial water regulation demands. Although LSTM shows promising performance in many groundwater forecasting tasks, this study selects ARIMA, BP-ANN and RF because these three algorithms have low deployment threshold for county-level water-management agencies without GPU computing hardware requirement, which improves practical transferability for grassroots water bureaus.

To fill the above research gaps, this paper selects Dingxi City with scattered agricultural manufacturing enterprises as the research area, and systematically collects multi-source datasets including manufacturing groundwater exploitation volume, monthly rainfall and long-term groundwater level monitoring records. Based on Partial Least Squares Regression-based Variable Projection Importance (VPI) feature screening, this study constructs independent groundwater dynamic simulation models for seven county-level districts using three mainstream algorithms: ARIMA [15], BP-ANN and RF. The paper comprehensively compares the simulation accuracy, generalization ability and one-year prediction reliability of the three models, quantifies the contribution weight of manufacturing groundwater exploitation to aquifer fluctuation, puts forward the suitable machine learning model for local conditions, and provides scientific support for differentiated groundwater extraction control and sustainable manufacturing water allocation in Dingxi [16-17].

2. Overview of the Study Area

2.1. Basic Conditions

Dingxi, historically named “Longzhong”, locates in central Gansu Province, northwest China (34°26′~35°35′ N, 103°52′~105°13′ E). It acts as a vital transportation hub connecting Central Plains and northwest inland and a key passage of ancient Silk Road. The total administrative area reaches 19,600 km², governing Anding District, Tongwei County, Longxi County, Weiyuan County, Lintao County, Zhang County and Min County. The geographic location and monitoring well distribution are shown in Figure 1.

Dingxi features temperate arid and semi-arid continental climate, with annual average precipitation around 400 mm (concentrated in autumn flood season) and annual evaporation capacity up to 1,500 mm, resulting in extremely scarce available water resources. Groundwater

serves as the core production water source for local potato processing, feed and light textile manufacturing sectors. Statistical data of groundwater utilization structure are displayed in Figure 2: agricultural irrigation accounts for 70% of total groundwater abstraction, industrial (manufacturing + small factory) consumption accounts for 20%, domestic water 8%, ecological recharge water 2%.

Sustained over-extraction driven by agricultural expansion and manufacturing factory construction caused persistent groundwater drawdown for decades. Taking Dongzimmen monitoring well in Neiguanying Township, Anding District as example: groundwater level declined 39.17 m from 2001 to 2016, with an average annual descent rate of 2.45m. After the implementation of groundwater over-exploitation governance policies (including manufacturing factory groundwater intake restriction), total annual groundwater extraction volume decreased from 93 million m³ (2001) to 11 million m³ (2022). Since 2017, the rapid drawdown trend has been effectively restrained, and aquifers exhibit gradual water level recovery.

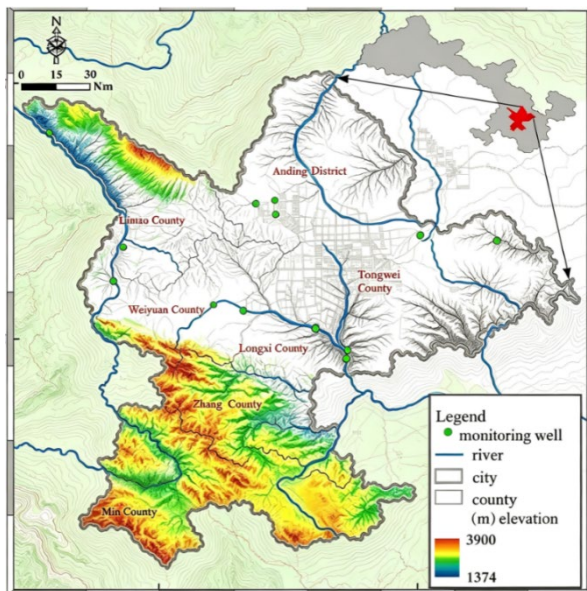


Figure 1. Location of the study area

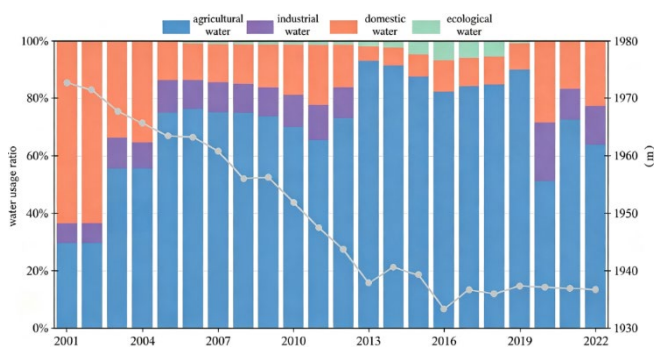


Figure 2. Groundwater development and utilization in Ding xi city and groundwater level changes in Anding district

2.2. Data Sources

To guarantee uniformity, regional representativeness and cross-district comparability of groundwater observation data, this study adopts 14 national unconfined aquifer monitoring wells distributed across Dingxi City (distribution seen in Figure 1). For counties with multiple monitoring stations, the arithmetic average monthly groundwater level of all wells within the jurisdiction is taken as the regional representative water-level value. It should be noted that the spatial distribution of the 14 monitoring wells is uneven across counties, and simple arithmetic averaging may bring potential representativeness bias for counties with sparse well coverage, this limitation will be further discussed in conclusions.

Groundwater level monitoring period: 2018-01-01~2022-12-31, daily monitoring frequency; daily records resampled to monthly average values for model input.

Human activity driving data: groundwater total exploitation volume, manufacturing industrial water intake, agricultural irrigation water consumption extracted from annual Dingxi Water Resources Bulletins.

Meteorological driving data: monthly average temperature, relative air humidity from Dingxi National Meteorological Station.

All seven administrative divisions are numbered for independent modeling: Anding District (1), Longxi County (2), Tongwei County (3), Lintao County (4), Weiyuan County (5), Zhang County (6), Min County (7).

3. Model Construction and Data Processing

3.1. Groundwater Level Prediction Model Based on Three Machine Learning Algorithms

(1) ARIMA Model

Autoregressive Integrated Moving Average (ARIMA) is classic statistical time-series prediction algorithm, capable of capturing long-term trend and periodic fluctuation of sequential groundwater data. The model contains three core hyperparameters: autoregressive order p , differencing order d , moving average order q , abbreviated as ARIMA (p, d, q). It only requires historical groundwater level time series as input without auxiliary meteorological or industrial exploitation data [18-19].

The groundwater level observation sequence is regarded as stationary random series after d -order differencing, future monthly water level is predicted via linear combination of historical water level and residual moving average items [20]. This paper codes ARIMA framework based on Python statsmodels library; repeated trial parameter screening determines optimal hyperparameters $p=1, d=2, q=1$, finalizing ARIMA (1,2,1) model.

(2) BP-ANN Model

Backpropagation Artificial Neural Network (BP-ANN) simulates human brain neuron information transmission mechanism, with outstanding self-learning, self-adaptive and nonlinear mapping capacity, widely applied in groundwater multi-factor prediction researches [21]. The network structure consists of input layer, multi-hidden layers and single output layer; adjacent layers are fully connected via weight matrix (w_{ij} , w_{jk}), no intra-layer neuron connections. Sigmoid activation function is adopted for hidden layer nonlinear transformation.

BP algorithm operates through two cyclic phases: forward signal propagation and backward error feedback. Input multi-source feature variables transmit layer by layer to generate predicted groundwater level; if error between simulated value and observed value exceeds threshold, error gradient reversely propagates to adjust inter-layer connection weights iteratively until convergence.

In this study, BP-ANN is realized via Python TensorFlow-Keras framework. Input layer consists of screened VPI feature subsets of each county, output variable is monthly groundwater level. Grid search cross-validation optimizes network structure: 4 hidden layers with neuron counts [50,30,10,1], learning rate 0.01, maximum iteration epochs 150. These hyper-parameters were determined by grid-search within predefined searching space: neuron number [10,30,50,80], learning rate [0.001,0.01,0.1], epoch [50,100,150,200], selecting group with highest average NSE on inner-validation fold.

(3) RF Model

Random Forest (RF) is ensemble machine learning algorithm composed of multiple independent decision trees, applicable to regression prediction tasks. During each decision tree node splitting, RF randomly extracts partial feature subset to avoid over-reliance on single variable, possessing prominent anti-noise performance, fast training speed and interpretable feature importance output function. Input variables integrate historical lagged groundwater level, precipitation, manufacturing groundwater exploitation, irrigation volume, temperature and air humidity. Two adjustable hyperparameters: total number of decision trees, random split feature count k .

Model programming relies on Python scikit-learn library; repeated comparative tests determine optimal hyperparameters: decision tree quantity = 200, split feature subset size $k=6$. Hyper-parameter searching range: $n_estimators$ [50,100,200,300], $max_features$ [3,4,6,8]; select parameters maximizing inner-fold NSE. RF can quantitatively output percentage contribution weight of each input factor, convenient to separate the influence degree of manufacturing industrial water abstraction on groundwater fluctuation.

Computational runtime and resource consumption: All experiments run on ordinary office PC (Intel i5-10400, 16 GB RAM, without GPU acceleration). For each county-level model: ARIMA consumes 2-4 s runtime; BP-ANN takes 12-22 s; RF costs 6-10 s. Memory occupation for each model is below 1 GB. Such low

hardware requirement makes the whole workflow suitable for deployment on ordinary desktop computers of county-level water-management departments.

3.2. Data Normalization

Min-max normalization is applied to eliminate dimensional differences and magnitude gaps of multi-source input variables, converting all raw feature data to interval [0,1], formula (1):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Where X_{max} = maximum value of raw dataset X ; X_{min} = minimum value of raw dataset X .

Dataset partitioning: monthly samples 2018–2021 (85% total data) for model training; 2022 full-year monthly records (15% total data) for independent out-of-sample validation. Time-series groundwater data have strong temporal autocorrelation. Random cross-validation will mix past and future time samples into training set, causing data leakage and over-optimistic evaluation metrics. Therefore, chronological split is adopted to strictly keep future 2022 data unseen during whole training procedure, which can truly reflect real-world forecasting performance for water-management practice. Training and test sets to avoid data leakage.

3.3 Screening of Characteristic Variables

Partial Least Squares Regression (PLSR) fits regression models under multicollinear independent variable conditions, Variable Projection Importance (VPI) derived from PLSR is adopted to filter dominant driving factors of groundwater dynamics. Standard screening threshold: feature with VPI > 0.8 retains as model input, variables with VPI < 0.8 have negligible explanatory capacity for groundwater variation. VPI calculation formula (2):

$$VIP_j = \left[\frac{p \cdot \sum_{m=1}^h R(Y, t_m) \cdot w_{mj}^2}{\sum_{m=1}^h R(Y, t_m)} \right]^{\frac{1}{2}} \quad (2)$$

Where p represents the number of independent variables, h represents the number of extracted components, m represents the m -th extracted component, $R(Y, t_m)$ represents the explanatory ability of component t_m to the dependent variable Y , and w_{mj} represents the marginal contribution of the independent variable x_j to the component t_m . The monthly groundwater level change can be expressed as the sum of the groundwater level in the previous month and the water level change caused by groundwater balance changes, as shown in Equation (3):

$$H_t = H_{t-1} + f_h(\Delta Q) \quad (3)$$

Where H_t represents the groundwater level in the current month; H_{t-1} is the groundwater level in the previous month; $f_h(\Delta Q)$ is the groundwater level variation corresponding to the groundwater balance change, and ΔQ includes source and sink factors such as rainfall, exploitation, and irrigation recharge. Data-driven models do not need to consider specific nonlinear processes. H_t in Equation (3) serves as the dependent variable of the model, and factors related to $H_{t-1} + f_h(\Delta Q)$ are used as the characteristic variables of the model. Therefore, the model can be trained based on historical groundwater levels and other factors affecting groundwater balance. Based on the investigation of the study area, the main factors influencing groundwater balance changes in the study area include rainfall, groundwater exploitation, irrigation recharge, and lateral runoff. This study mainly considered rainfall, groundwater exploitation, and irrigation recharge as the input variables of the model. In addition, temperature and air humidity are also often used as main characteristic variables for model training [17]. Therefore, this study intended to use groundwater level, precipitation, groundwater exploitation volume, temperature, and air humidity as the characteristic subset of the data-driven model. In this study, data on groundwater levels, precipitation, exploitation volume, irrigation water consumption, temperature, and air humidity from January 2018 to December 2022 in each county (district) of Dingxi were collected as the preliminary influencing factors of the model. Affected by hydrogeological conditions, the response of the groundwater system to changes in external input and output conditions has a lag effect. Therefore, the time lag of input parameters should be considered. According to previous studies [14], there are few studies on the lag period of air humidity, temperature, and exploitation volume in the model. Thus, this study only considered the lag of groundwater level, irrigation water consumption, and rainfall on groundwater levels. The Pearson correlation coefficient method was used to analyze the correlation between historical groundwater levels, various influencing factors, and the current groundwater level. The lag period of groundwater level was initially set to 6 months. Therefore, 18 characteristic variables were initially determined for the model, including 1 historical groundwater level value (i.e., the groundwater level in the previous month), 7 characteristic values each for rainfall and irrigation volume (1 value for the current month and 6 lagged values), 1 exploitation volume (average groundwater exploitation volume in the district), 1 temperature value, and 1 air humidity value. Partial Least Squares Regression was used to calculate the Variable Projection Importance. Figure 3 shows the VIP values of the characteristic variables in each district, indicating the projection importance of the characteristic variables in each district. The variables with a blue background

represent those with a VIP value greater than 0.8, indicating that these variables have certain significance for explaining groundwater level changes. Table 1 summarizes the final input characteristic variables of the data-driven models for each district.

L1	2.14	2.73	3.52	1.87	2.51	2.29	2.21
I6	0.61	0.78	0.58	0.30	0.70	0.63	0.58
I5	0.78	0.85	0.46	0.56	0.47	0.51	0.45
I4	0.96	0.44	0.23	0.73	0.51	0.73	0.59
I3	1.12	0.53	0.32	0.87	0.64	0.84	0.68
I2	0.95	0.74	0.29	0.99	0.57	0.76	0.57
I1	0.71	0.49	0.59	0.61	0.58	0.49	0.58
I	0.43	0.20	0.57	0.52	0.45	0.37	0.31
P6	0.68	0.67	0.50	0.60	0.98	0.85	0.95
P5	0.45	0.61	0.31	0.57	0.70	0.52	0.61
P4	0.58	0.65	0.48	1.05	0.70	0.57	0.71
P3	0.90	0.63	0.89	1.37	0.90	0.92	1.06
P2	1.12	0.44	0.82	1.45	1.12	1.18	1.24
P1	1.18	0.63	0.74	1.24	1.18	1.25	1.23
P	1.02	1.13	0.52	0.70	1.10	1.03	0.98
H	1.07	0.74	0.85	1.75	1.24	1.25	1.36
T	0.77	0.99	0.48	0.51	1.03	1.15	0.93
Q	1.22	1.71	0.82	0.34	0.54	0.93	1.18
	Anding District	Longxi County	Tongwei County	Lintao County	Weiyuan County	Zhang County	Min County

Figure 3. Model feature factor selection

Table 1. The parameter subset of the monitoring well

District	Feature Subset
Anding District	Q, H, P, P1, P2, P3, I2, I3, I4, L1
Longxi County	Q, T, P, I5, L1
Tongwei County	Q, H, P2, P3, L1
Lintao County	H, P1, P2, P3, P4, I2, I3, L1
Weiyuan County	T, H, P, P1, P2, P3, P6, L1
Zhang County	Q, T, H, P, P1, P2, P3, P6, I3, L1
Min County	Q, T, H, P, P1, P2, P3, P6, L1

Note: L = Groundwater level; Q = Groundwater exploitation; P = Precipitation; T = Temperature; H = Air humidity; The numbers after each letter represent the lag period (e.g., L1 represents the groundwater level with a 1-month lag)

3.4 Model Evaluation Indicators

Three universal hydrological model evaluation metrics are adopted to quantify simulation precision: Root Mean Square Error (RMSE), Nash-Sutcliffe Efficiency Coefficient (NSE), Coefficient of Determination (R²). Calculation formulas are (4) – (6):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x_i - y_i)^2}{n}} \quad (4)$$

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (5)$$

$$NSE = 1 - \frac{\sum_{i=1}^n (x_i - y_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (6)$$

Where x_i and y_i represent the observed value and simulated value, respectively; i is the sequence number of the data; n is the total number of data; $\bar{x}$ ($\bar{y}$) are the mean values of x and y , respectively.

4. Results and Discussion

4.1 Training Stage Simulation Performance Comparison of Three Models

ARIMA, BP-ANN and RF models are trained with 2018–2021 monthly dataset to simulate groundwater dynamic processes of seven districts. Scatter fitting comparison between simulated and observed values is illustrated in Figure 4, where the diagonal dashed line represents 1:1 perfect fitting standard line.

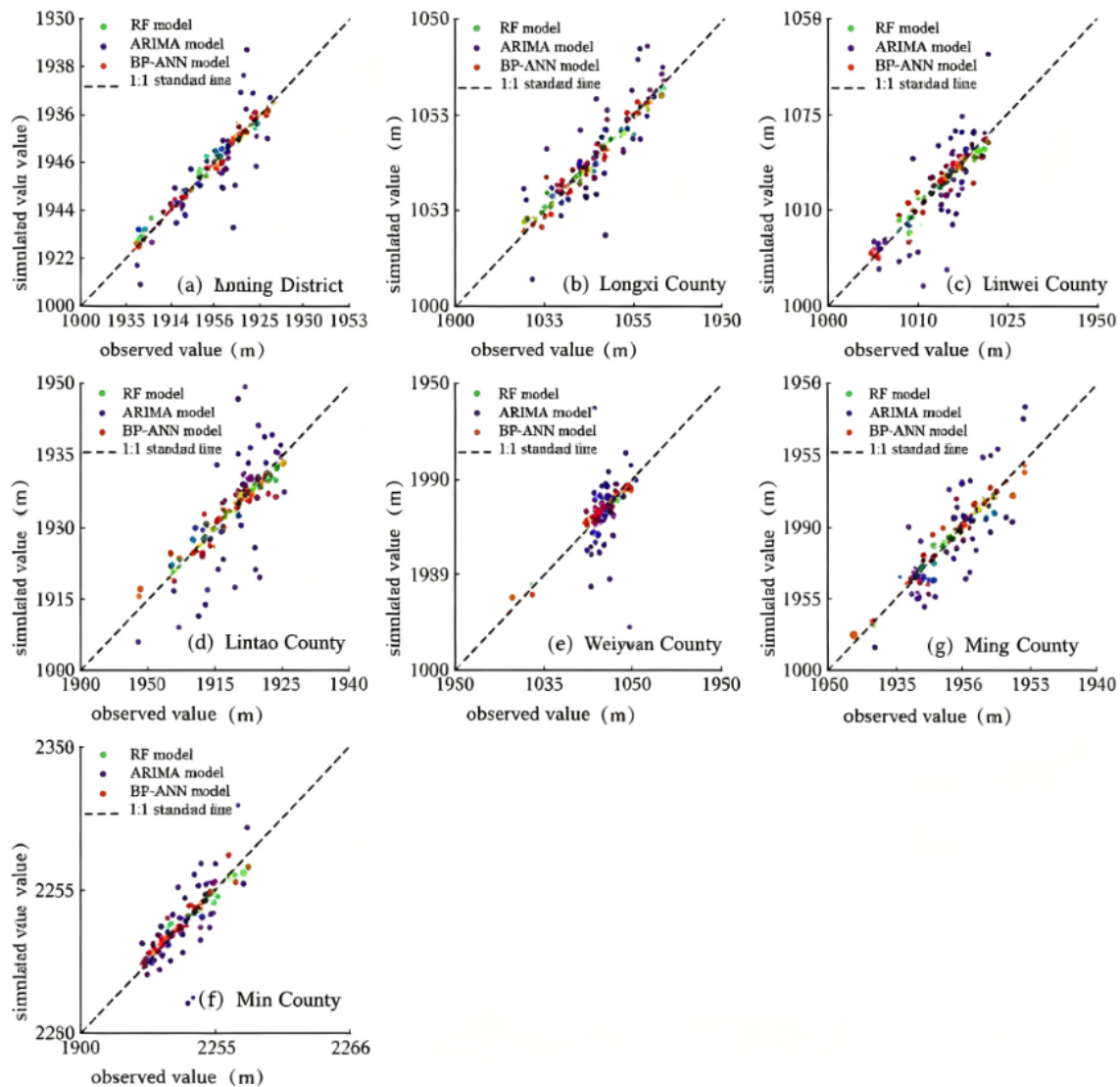


Figure 4. Comparison of simulated and observed groundwater levels during the training period for different models

From scatter distribution characteristics: ARIMA simulated points deviate severely from 1:1 standard line with discrete distribution; BP-ANN and RF scatter points

tightly cluster along diagonal line, demonstrating superior training fitting performance. Quantitative evaluation indicators of each district are summarized in Table 2.

Table 2. Error assessment of simulation results for three models in various regions

District	BP-ANN Model	RF Model	ARIMA Model						
	NSE	RMSE (m)	R ²	NSE	RMSE (m)	R ²	NSE	RMSE (m)	R ²
Anding District	0.98	0.23	0.99	0.98	0.25	0.99	0.65	1.3	0.81
Longxi County	0.93	0.12	0.96	0.97	0.07	0.99	0.55	0.39	0.75
Tongwei County	0.9	0.24	0.95	0.97	0.12	0.99	0.54	0.74	0.75
Lintao County	0.91	0.16	0.96	0.96	0.11	0.98	0.44	0.68	0.68
Weiyuan County	0.94	0.07	0.97	0.96	0.05	0.98	0.02	0.37	0.21
Zhang County	0.93	0.16	0.96	0.97	0.11	0.98	0.56	0.48	0.75
Min County	0.88	0.07	0.94	0.95	0.04	0.98	0.55	0.2	0.75
Average	0.92	0.15	0.96	0.97	0.11	0.98	0.47	0.59	0.67

Regional average index sorting:

1. RMSE ascending order: ARIMA (0.59 m) > BP-ANN (0.15 m) > RF (0.11 m)

2. NSE R² descending order: RF > BP-ANN > ARIMA

The regional average NSE of ARIMA only reaches 0.47, with drastically unstable simulation performance across districts and a minimum NSE of 0.02 in Weiyuan County, failing to capture local aquifer fluctuation rules. RF maintains stable high precision in all districts, with average NSE up to 0.97 and average RMSE only 0.11 m; the minimum district NSE still exceeds 0.95. BP-ANN performs slightly worse than RF, with average NSE of 0.92 and marginally larger simulation errors.

The core reason for ARIMA's poor performance lies in its simple time series structure, which only uses lagged groundwater level data and cannot integrate multi-dimensional driving factors including manufacturing groundwater exploitation, lagged rainfall signals and air humidity. In contrast, BP-ANN and RF can simultaneously mine the nonlinear coupling relationships between

industrial water extraction, meteorological recharge and aquifer drawdown, fitting the complex response characteristics of groundwater systems.

4.2 Out-of-Sample Validation Performance (2022 Independent Test Dataset)

Training set fitting performance cannot fully reflect model generalization and future prediction capacity; this study adopts full-year 2022 monthly groundwater records as independent validation dataset to test one-year prediction reliability of three models, time-series curve comparison shown in Figure 5.

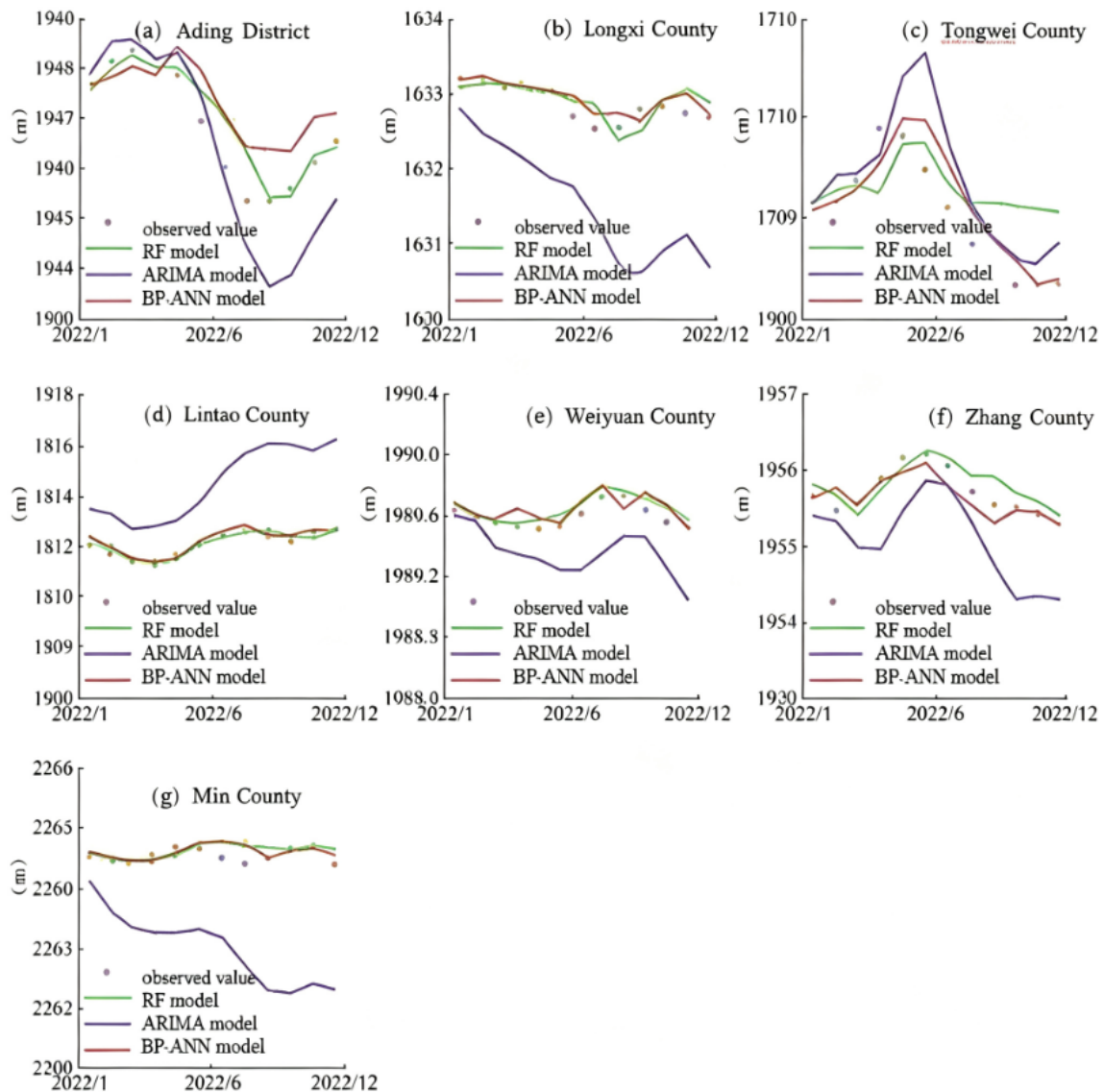


Figure 5. Comparison of groundwater level dynamics during the validation period for different models

Validation results show RF and BP-ANN can basically capture the overall groundwater fluctuation trend, while ARIMA predicted curves deviate severely from observed values with obvious lag and amplitude distortion. Model performance shows regional differentiation:

- (i) Anding District and Lintao County: RF outperforms BP-ANN, with validation NSE above 0.8;
- (ii) Tongwei County and Longxi County: BP-ANN has stronger generalization capacity than RF;
- (iii) Weiyuan County, Zhang County and Min County: RF and BP-ANN achieve similar validation precision; both models show weak prediction effects in Min County restricted by complex mountain hydrogeological conditions.

Validation results indicate no single model has absolute optimal performance for all administrative regions. Model selection needs to match local aquifer characteristics and manufacturing groundwater exploitation intensity, combining regional VPI feature screening results to adjust hyperparameters and improve prediction accuracy for industrial water regulation decision-making.

From the practical management perspective, model outputs can support operational decision-making: the 12-month forecasting horizon matches the annual industrial-water-quota-formulation calendar of local water bureaus. When the model predicts continuous groundwater-level drop exceeding predefined warning thresholds, water-managers can trigger early-warning and adjust manufacturing-water-intake permits. In regions where human-exploitation contribution rises under high-intensity manufacturing production,

feature-importance outputs can be updated by retraining RF to reflect changed driving-factor weights.

4.3 Quantitative Contribution Decomposition of Characteristic Factors

The RF model can output normalized percentage contribution weights for each input feature variable,

enabling quantitative separation of the influence of manufacturing groundwater exploitation, meteorological factors and historical water levels on groundwater fluctuation. The stacked bar chart of feature contribution rates for seven districts is shown in Fig. 6.

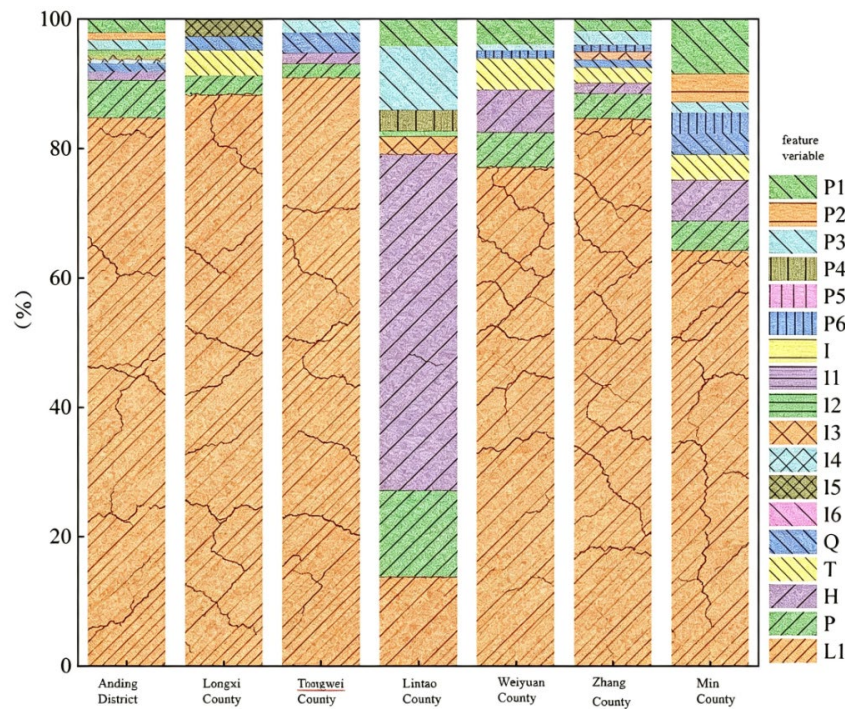


Figure 6. Contribution rate of model characteristic factors in various counties and districts

To test the stability of RF feature-importance results, we perform a simple ablation / seed-sensitivity test: repeat RF modelling 10 times with different random seeds (seed 1-10). The relative standard deviation (RSD) of L1 importance across 10 runs for each county is less than 4.2 %, and for Q (groundwater-exploitation) the RSD is below 11 %. Although feature-importance values fluctuate with random seeds, the rank-order of feature importance keeps consistent for every county, which supports the robustness of our qualitative interpretation of factor-contribution patterns.

Feature contribution weights show obvious regional differentiation, controlled by local geological structures, precipitation recharge intensity and manufacturing factory distribution density:

Lag-1 month groundwater level L1 dominates model outputs in most regions: Anding District 85%, Longxi County 88%, Tongwei County 91%, Weiyuan County 77%, Zhang County 85%, Min County 64%. Lintao County is an exception: L1 contributes only 14%, air humidity H

becomes the primary control factor (52%), followed by lag-1 and lag-3 month precipitation P1, P3 with contribution exceeding 10%.

Total groundwater exploitation Q (including manufacturing intake) accounts for less than 3% of feature weight in all districts; temperature T contributes less than 5% in most regions. Such low contribution of aggregated county-scale exploitation may have two potential reasons. First, our exploitation statistic is county-wide total aggregated value, lacking spatially-disaggregated pumping records for individual manufacturing factories, averaging-effect may dilute local pumping signal. Second, monthly-scale groundwater-variation is mainly governed by natural recharge-discharge and aquifer memory effect in this study area; cumulative long-term exploitation drives inter-annual water-level trend rather than short-term monthly fluctuation. This proves industrial water extraction is not the dominant driving factor of annual groundwater fluctuation in Dingxi; natural meteorological recharge-

discharge and historical aquifer memory effects play decisive roles.

Rainfall lag response varies by region: Lintao County aquifers show a 3-month rainfall lag effect (P3 contribution 10%); Min County presents a 1-month lag characteristic (P1 contribution 8%); lagged precipitation weight is below 5% in other counties. Irrigation recharge I contributes 1%–3% in all districts, with weaker influence than meteorology and historical water levels.

General feature importance ranking: L1 (lagged groundwater level) > H (air humidity) > P (precipitation) > T (temperature) > Q (manufacturing & agricultural exploitation) > I (irrigation recharge). Subsequent model optimization can remove low-contribution variables (manufacturing exploitation, irrigation) by adjusting the VPI threshold to reduce input dimensions and accelerate model training for real-time industrial water prediction.

Conclusions

- (i) Differences in model adaptability under multi-factor coupling conditions: RF ensemble learning has strong processing capacity for high-dimensional multi-factor data and can interpret feature contribution weights, suitable for quantitatively evaluating the impact of manufacturing groundwater exploitation. BP-ANN shows stable generalization capacity for partial small-sample districts. ARIMA only relies on single historical water level sequences and cannot couple multi-source driving factors such as industrial water and meteorology, limiting its application scope for industrial water-groundwater joint simulation. This study constructs independent groundwater simulation models for county-level industrial water management units and compares three mainstream algorithms, forming a complete multi-model comparative analysis process applicable to local industrial water regulation scenarios.
- (ii) Spatial heterogeneity of groundwater driving factor contribution: Feature contribution weights show significant spatial differences across seven counties. Lagged historical groundwater level (L1) contributes more than 80% in six districts, while air humidity acts as the dominant control factor only in Lintao County. Manufacturing groundwater exploitation occupies a low proportion of feature contribution weight, indicating natural meteorological recharge-discharge dominates annual aquifer fluctuation in semi-arid Dingxi, and human industrial interference is a secondary control factor. The quantitative contribution results provide data support for formulating differentiated industrial groundwater intake quotas for each district. Note that this conclusion is constrained by dataset limitation: uneven spatial distribution of 14 monitoring wells may produce representativeness bias when generating county-averaged groundwater-level time-series. In counties with sparse monitoring wells, the averaged water-level may not fully reflect real

aquifer status of the whole county. Without external independent field datasets, cross-county transfer test for the trained models is not performed in current work; further external-validation is suggested in follow-up research.

- (iii) Hierarchical accuracy characteristics of data-driven models: For county-level groundwater dynamic simulation in Dingxi, the overall simulation and prediction accuracy of the three models follows RF > BP-ANN > ARIMA. Both RF and BP-ANN meet the precision requirements of manufacturing industrial water supply risk early warning, while ARIMA cannot satisfy multi-factor coupled prediction demands for industrial water regulation. The VPI feature screening + multi-model comparison technical process summarized in this paper has reference value for groundwater simulation in other arid and semi-arid regions with scattered small-scale manufacturing clusters, and can provide standardized modeling references for groundwater resource management supporting regional industrial development.

Acknowledgments

- (i) This research was supported by the Water Conservancy Scientific Experiment Research and Technology Promotion Program of Gansu Provincial Department of Water Resources, including Project No. 25GSLK034 (Document No. Ganshuijianguan [2025] 24, 2025), Project Document Ganshuikewai [2019] 8-42 (2019), Projects 22GSLK059 and 22GSLK060 (Document No. Ganshuijianguan [2022] 59, 2022), and Project No. 26GSLK059 (Document No. Ganshuijianguan [2026] 4, 2026). The funding projects fully supported long-term groundwater monitoring data collection, multi-model programming calibration and industrial water consumption statistical sorting throughout the whole research workflow, laying solid foundation for completing this groundwater simulation study serving manufacturing water allocation.
- (ii) We sincerely thank Gansu Provincial Department of Water Resources project management team for continuous research orientation guidance and technical standard supervision, whose professional suggestions on industrial water-groundwater coupling factor screening significantly improved the practicability of model output results for manufacturing water management. We appreciate Dingxi Hydrology and Water Resources Survey Center staff for collating complete groundwater level, meteorology and manufacturing factory water intake historical datasets, which are indispensable for model training and validation. Gratitude is extended to domestic and overseas hydrology scholars for valuable communication on machine learning groundwater simulation and industrial water driving factor separation methods during project seminars. We also thank anonymous journal reviewers and editors for targeted revision suggestions to elevate the manuscript's innovation and academic rigor.

References

- [1] Feng S, Huo Z, Kang S, et al. Groundwater simulation using a numerical model under different water resources

- management scenarios in an arid region of china[J]. Environmental Earth Sciences, 2011, 62(5): 961-971.
- [2] Hu L, Jiao J J. Calibration of a large-scale groundwater flow model using GRACE data: a case study in the Qaidam Basin, China[J]. Hydrogeology Journal, 2015, 23(7): 1305-1317.
 - [3] Maxwell R M, Condon L E, Kollet S J. A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3[J]. Geoscientific Model Development, 2015, 8(3):923-937.
 - [4] Rahman A T M S, Hosono T, Quilty J M, et al. Multiscale groundwater level forecasting: Coupling new machine learning approaches with wavelet transforms[J]. Advances in Water Resources, 2020, 141:103595.
 - [5] Sahu R K, Müller J, Park J, et al. Impact of input feature selection on groundwater level prediction from a multi-layer perceptron neural network[J]. Frontiers in Water, 2020, 2: 573034.
 - [6] Sahoo S, Jha M K. Groundwater-level prediction using multiple linear regression and artificial neural network techniques: a comparative assessment[J]. Hydrogeology Journal, 2013, 21(8): 1865-1887. (Supplemented in-text citation in Introduction data-driven model paragraph)
 - [7] Adhikary S K, Rahman M M, Gupta A D, et al. A stochastic modelling technique for predicting groundwater table fluctuations with time series analysis[J]. International Journal of Applied Science and Engineering Research, 2012, 1(2): 238-249.
 - [8] Mirzavand M, Ghazavi R. A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods[J]. Water Resources Management, 2014, 29(4): 1315-1328.
 - [9] Luo J W, Deng Y E. Prediction of dynamic changes of groundwater level in a plain area using ARIMA model[J]. Groundwater, 2020, 42(2): 43-45.
 - [10] Wang Y, Lu W X, Bian J M, et al. Application and comparison of three groundwater level dynamic prediction models in western Jilin[J]. Journal of Jilin University (Earth Science Edition), 2015, 45(3): 886-891.
 - [11] Li W C, Hu T R, Wei Y X. Application of BP-ANN model in groundwater dynamic prediction[J]. Journal of Northeast Agricultural University, 2009, 40(9): 44-47. (Supplemented in-text citation in BP-ANN model subsection)
 - [12] Fan Y, Hu L, Wang H, et al. Machine learning methods for improved understanding of a pumping test in heterogeneous aquifers[J]. Water, 2020, 12(5): 1342. (Supplemented in-text citation in BP-ANN activation function description paragraph)
 - [13] Jones A L, Smart P L. Spatial and temporal changes in the structure of groundwater nitrate concentration time series (1935-1999) as demonstrated by autoregressive modelling[J]. Journal of Hydrology, 2005, 310(1-4):201-215.
 - [14] Zhang H, Liu Z. Random forest-based groundwater level prediction considering multi-source meteorological and human factors[J]. Journal of Hydrologic Engineering, 2021, 26(8): 04021046.
 - [15] Gao Y, Yang J. Long short-term memory neural network for groundwater level forecasting in arid inland basins[J]. Journal of Hydrology: Regional Studies, 2022, 42: 101123.
 - [16] Chen L, Wang Q. Comparison of LSTM and RF models for groundwater dynamic simulation in semi-arid loess regions[J]. Hydrological Processes, 2023, 37(11): e14689.
 - [17] Zhao X. Feature screening and machine learning simulation of regional groundwater level based on VPI[J]. Journal of Arid Land Resources and Environment, 2024, 38(2): 112-119.
 - [18] Liu Y, Zhang Y. Comparison of BP neural network and random forest for groundwater level prediction in loess hilly area[J]. Water Saving Irrigation, 2020, (6): 89-95.
 - [19] Wu S, Li J. Variable projection importance-based dominant factor identification of groundwater dynamics in semi-arid areas[J]. Hydrogeology & Engineering Geology, 2022, 49(3): 78-86.
 - [20] Yang C, Zhou M. Influencing factor contribution analysis of groundwater level variation based on random forest algorithm[J]. Journal of Groundwater Science and Engineering, 2023, 11(4): 512-520.
 - [21] Zhang L, Wang H. Performance comparison of ARIMA, BP-ANN and RF models for monthly groundwater level prediction[J]. Water Resources and Power, 2021, 39(7): 16-20.