AQUA
Water Infrastructure, Ecosystems and Society

IWA PUBLISHING

 AQUA — Water Infrastructure, Ecosystems and Society Vol 70 No 8, 1272 doi: 10.2166/aqua.2021.205

Prediction and analysis of water resources demand in Taiyuan City based on principal component analysis and BP neural network

Junhao Wu[a], Zhaocai Wang[b,*] and Leyiping Dong[c]

[a] College of Economics and Management, Shanghai Ocean University, Shanghai 201306, China
[b] College of Information, Shanghai Ocean University, Shanghai 201306, China
[c] College of AIEN, Shanghai Ocean University, Shanghai 201306, China
*Corresponding author. E-mail: zcwang1028@163.com

ABSTRACT

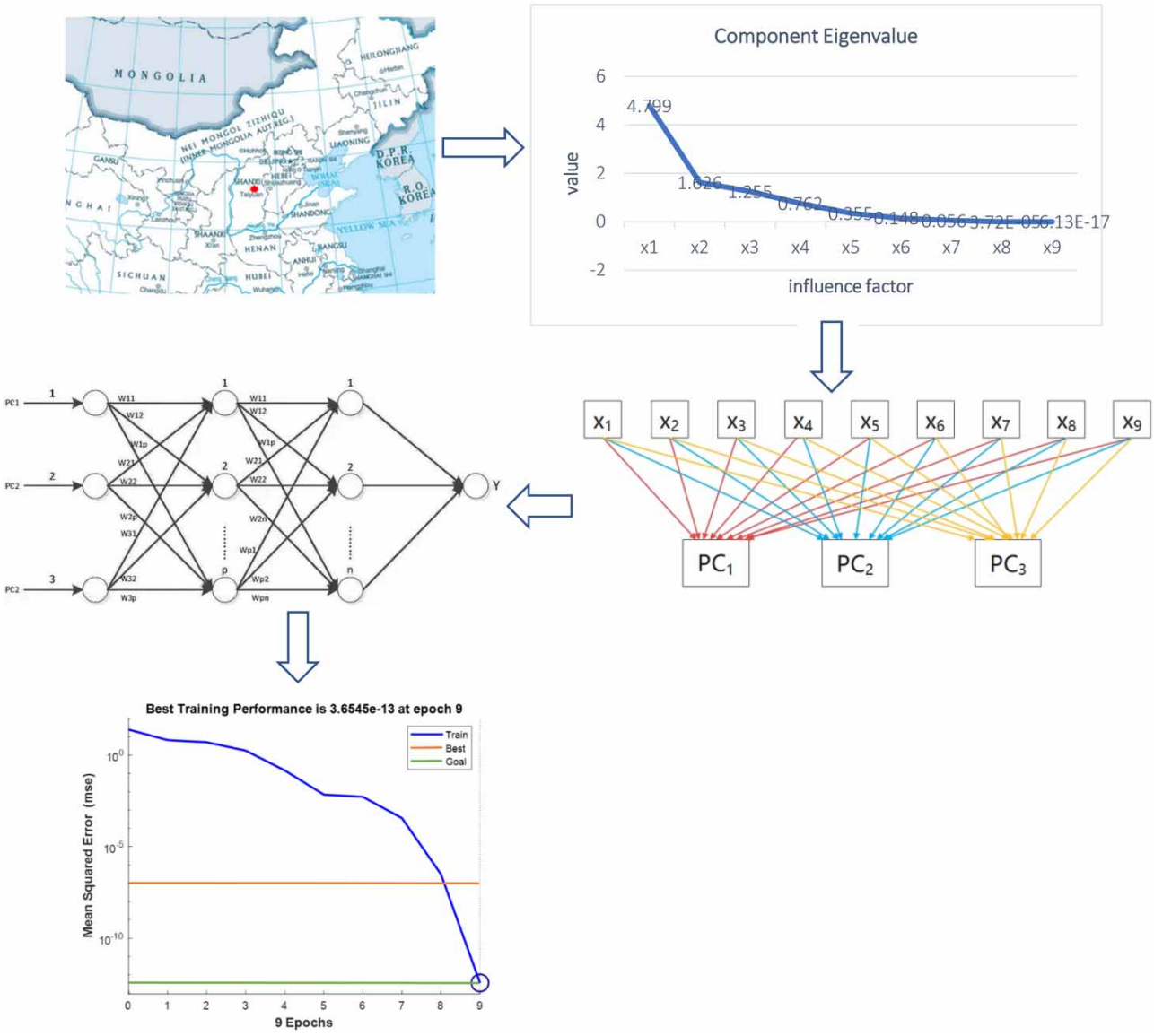
Water is a fundamental natural and strategic economic resource that plays a vital role in promoting economic and social development. With the accelerated urbanization and industrialization in China, the potential demand for water resources will be enormous. Therefore, accurate prediction of water resources demand is important for the formulation of industrial and agricultural policies, development of economic plans, and many other aspects. In this study, we develop a model based on principal component analysis (PCA) and back propagation (BP) neural network to predict water resources demand in Taiyuan, Shanxi Province, a city with severe water shortage in China. The prediction accuracy is then compared with PCA-ANN, ARIMA, NARX, Grey–Markov, serial regression, and LSTM models, and the results showed that the PCA-BP model outperformed other models in many evaluation factors. The proposed PCA-BP model reduces the dimensionality of high-dimensional variables by PCA and transformed them into uncorrelated composite data, which can make them easier to compute. More importantly, BP and weight threshold adjustment in model training further improve the prediction accuracy of the model. The model analysis will provide an important reference for water demand assessment and optimal water allocation in other regions.

Key words: BP neural network, principal component analysis, water demand forecast, water resources planning

HIGHLIGHTS

- A water demand forecasting model based on principal component analysis (PCA) and back propagation (BP) neural network is proposed.
- PCA is used to reduce the dimensionality of the data to reduce the computational complexity.
- Compared with other existing models, the prediction accuracy of the PCA-BP model has been significantly improved.

GRAPHICAL ABSTRACT



1. INTRODUCTION

According to China's second water resource statistics, the total amount of surface water resources in China is 2,738.8 billion m^3. The relative amount of groundwater resources is about 821.8 billion m^3. After removing the mutual transformation between groundwater resources and surface water, the whole amount of water resources is about 2,841.2 billion m^3. However, due to the enormous population, the total quantity of water resources per capita is minimal. In 2017, per capita water consumption in China dropped to 436 m^3, far below the world's average level of 8,000 m^3 per person (Manju & Sagar 2017). Moreover, because of the regional severe imbalance of water resources in China, especially in megacities, there is a serious issue of water shortage. Tianjin, Beijing, Taiyuan, and many other cities in North China use less than 300 m^3 of water per capita. At the same time, from 2000 to 2018, China's total population soared from 1.23 to 1.39 billion, and the urbanization rate boosted rapidly from 31.9 to 59.6%. With the rapid increase of urbanization rate and the large flow of population to cities, the total amount of domestic water will continue to rise rapidly, which also makes the weak urban water supply system worse (Kandissounon *et al.* 2018; Xiang & Jia 2019). In recent years, one of the essential requirements of the Chinese

government, and a critical link in its goal for the construction of an ecological civilization, has been a clean, safe, and sufficient water supply, which has, in turn, made the accurate prediction of domestic water supply become the primary prerequisite for the formulation of optimal provision and allocation of the resource.

Shanxi Province is a critical energy and heavy chemical base in China, and one of the provinces with severe water shortages. As the capital of Shanxi Province, the current situation of water resources in Taiyuan is also an emergency. According to the Shanxi Water Resources Bulletin, Taiyuan's water resources per capita is 177 m^3, less than 9% of the national per capita amount, and is in a state of extreme water shortage as determined by the United Nations. Due to the serious shortage of water resources, Taiyuan City has to rely on more than 0.7 billion m^3 of groundwater every year to support economic and social development and water for residents. Over the years, serious overdrafting of groundwater has led to a significant drop in the groundwater level in the area, the aquifer has been continuously drained, a large number of shallow wells have been scrapped, and the depth of water extraction wells has been shifting downward, coupled with coal mining seepage not being effectively curbed, which means groundwater resources in Taiyuan are in great danger. Surface water pollution further exacerbates the shortage of water resources in Taiyuan City, making the contradiction between the supply and demand of water resources more prominent. Currently, with economic development and population growth, Taiyuan City has a growing demand for water resources, posing a serious threat to achieving sustainable development. Therefore, it has great practical significance to study and forecast the water resources demand in Taiyuan City, which can provide useful theoretical references for the rational allocation of water resources and the coordinated development of regional economy in Taiyuan City.

The following content of this paper is divided into four parts: the current research status and shortcomings of water resources prediction are introduced in Section 2. The two models of principal component analysis (PCA) and back propagation (BP) neural network are discussed in Section 3. In Section 4, the relevant data of Taiyuan City, with a large population and relative lack of water resources, are taken as an example. The demand of water resources is analyzed and predicted by using the PCA-BP model. Moreover, the prediction results are evaluated and compared with other models. The comparison results with other models show that the PCA-BP model outperforms other models in all indicators. The last section concludes a summary of this study and gives relevant suggestions.

2. CURRENT STATUS OF RESEARCH ON WATER RESOURCES PREDICTION

How to organize and utilize the water resources effectively and rationally, in reality, is essential to solving the sustainable development of urban water resources. Precise prediction of water demand is the primary task of the optimal allocation of water resources. Consequently, analyzing the model and method of forecasting water demand accurately is highly necessary. Meanwhile, water resources demand is affected by many uncertain factors such as climate, population, and economic level, which makes it a great challenge to precisely predict water consumption. At present, according to the periodicity of prediction variables and the size of the prediction range, the existing methods and models usually have different application scenarios (Donkor *et al.* 2014). Nowadays, the water demand forecast is divided into three terms: the water demand forecast of more than 2 years is defined as a long-term forecast, while the water demand forecast of 3 months to 2 years belongs to a medium-term forecast, and the water demand forecast of fewer than 3 months is classified as a short-term forecast by Billings & Jones (2008).

Short-term water resources demand can be predicted by the time-series model. Based on the historical data, Wong *et al.* (2010) developed a time-series model of daily urban water consumption based on rainfall and temperature. They applied it to the prediction of water consumption in Hong Kong, China. Faced with the nonlinear problem of data information change, Adamowski *et al.* (2012) used the artificial neural network (ANN) to predict short-term urban water demand. The results show that the ANN is superior to the linear regression technology in urban water demand prediction. Besides ANNs, support vector machine (SVM) is another machine learning technology for forecasting short-term water demand. Herrera *et al.* (2010) established a support vector regression model to predict the future urban water demand of southeast Spain. The model's predictions yield more accurate results compared with ANNs and other machine learning methods. Additionally, the support vector regression model, developed by Braun *et al.* (2014) for the 24-h water demand prediction of residential areas in Berlin performs, is better than the seasonal autoregressive model. Since then, support vector regression has been widely used in urban water demand forecasting, and the effectiveness of the method has been further verified (Wang *et al.* 2015; Shabani *et al.* 2017; Antunes *et al.* 2018; Bata *et al.* 2020; Deng *et al.* 2020a). For the medium-term forecast, an ANN method was proposed to forecast Bangkok's 6-month water demand (Babel & Shinde 2011). Ziervogel *et al.* (2010)

used the information of seasonal climate change to predict and plan water resources in South Africa. Lv (2014) established the precipitation forecast model of Zhengzhou City by using the time-series analysis method, and the 3-month precipitation forecast results were given. Polebitski & Palmer (2010) developed a water demand regression model, which could accurately predict a single household's water resources demand in a bimonthly time step.

Ordinarily, nonlinear model, statistical analysis model, and grey forecasting model can be utilized to make long-term forecasting models (Hernandez *et al.* 2014). The nonlinear model mainly includes ANN (Ren *et al.* 2020), SVM (Zhu & Wei 2013), time-series analysis model (Angelopoulos *et al.* 2019), and Markov chain model (Tsiliyannis 2018). The Grey model is a prediction model based on the grey theory, which is mainly applied in the uncertain background of fewer data and information. Through data processing and analysis, it can achieve the establishment of a prediction model, predict the development trend, and make a reasonable evaluation (Ding 2018; Deng *et al.* 2020e). Wang *et al.* (2021b) used the optimized Grey–Markov model to forecast the domestic water consumption in the Shaanxi Province of China. The results presented that the accuracy of the model was significantly improved compared with the general grey model and unbiased grey model.

Nowadays, some new water resources prediction models have been proposed to predict the demand for water resources more accurately. Tian *et al.* (2016) used the simulation method and the newly developed retrospective weather forecast of numerical weather forecast to improve the short-term forecasting ability of urban water demand. The model results demonstrate that the simulation method based on numerical prediction has a good application prospect in improving the prediction accuracy. Sanchez *et al.* (2020) established the indexes of social economy, environment, and landscape pattern and used a geographically weighted regression model to predict the urban water demand. The model takes the water demand of North Carolina and South Carolina as the empirical objects to evaluate the influence of population density and climate warming on future water demand. The research also reveals that the prediction results are impacted directly by the value of parameters in the water demand prediction model. Therefore, the reasonable calculation of model parameters is the key to the accuracy of water demand prediction (Rehman *et al.* 2017). According to the historical water demand data, Oliveira *et al.* (2017) used the harmony search algorithm to optimize the parameters of the autoregressive integrated moving average (ARIMA) model, which improved the short-term water demand forecasting efficiency. Mohammad & Pezhman (2019) used the extended ARIMA model and the nonlinear autoregressive exogenous (NARX) method to predict Teheran water consumption successfully. In recent years, stochastic optimization algorithm based on biological evolutionary mechanisms has become the mainstream tool for solving some complex problems that the initial values are difficult to choose and the objective functions are difficult to meet the accuracy requirements (Deng *et al.* 2021a, 2021b; Li *et al.* 2021; Wang *et al.* 2021a). The algorithm, as a population intelligence model, achieves search and optimization of the solution space by information exchange and sharing with the ideas of simulating animal foraging behavior and population optimization for survival (Deng *et al.* 2020b, 2020c, 2020d). It has the advantages of simple design, fast convergence speed, and few control parameters. For example, the particle swarm optimization (PSO) algorithm is a representative meta-heuristic swarm intelligence algorithm that integrates the social and historical cognition of particles into behavior, which opens up a new way to solve the optimization problem based on the survival principle of animal evolution (Lalwani *et al.* 2019; Deng *et al.* 2020b; Wu *et al.* 2021). Since then, many scholars have proposed swarm intelligence algorithms with different mechanisms, such as moth flame optimization algorithm (Mirjalili 2015a), ant lion optimizer (Mirjalili 2015b), grey wolf optimization algorithm (Teng *et al.* 2019), Harris hawk optimization (Heidari *et al.* 2019), SALP colony algorithm (Braik 2021). Swarm intelligence optimization algorithm has been used to predict water demand in recent years because of the excellent performance of the intelligent algorithms in resolving optimization problems. Bai *et al.* (2014) proposed an urban water demand estimation method based on multi-scale measurement. With this method, the adaptive chaotic PSO algorithm was used to search the optimal weighting factor of the relevance vector regression model. Wang *et al.* (2018) proposed a hybrid model based on linear and exponential models, using the firefly algorithm to solve the weight operators. Guo *et al.* (2020) proposed an improved whale optimization algorithm based on social learning and wavelet mutation strategy, which used the latest CEC2017 benchmark function to verify the superiority of the algorithm. Du *et al.* (2021) used PCA to reduce the dimensionality of factors affecting urban water demand, followed by discrete wavelet transform to eliminate the noisy part of water demand data. Then, LSTM was used to predict water demand with satisfactory results.

Nevertheless, previous research models often focus on the model design itself, and moreover, they do not fully consider the functional relationship between water resources demand and various influencing factors. In addition, there are many variables restricting water resource requirements, and the influence degrees are different. Therefore, it is necessary to quantify

and compare the influence degrees of various factors and select important vital indicators as the input information of the prediction model. Through the in-depth study of historical data information, it is more feasible and practical to seek regular changes and forecast future water resources demand.

3. BASIC PRINCIPLE AND METHOD STEPS OF THE MODEL

This section is divided into three parts. The basic principle of PCA is given in the first subsection. Next, the principle and steps of BP neural network are shown in the second subsection. Finally, the methods to test the accuracy of the model are listed in the third subsection.

3.1. Basic principle of PCA

PCA was first proposed by K. Pearson in 1901, then improved by Hotelling in 1933, and extended to random vector. The mathematical model of PCA can be expressed as follows:

$$\begin{cases} Z_1 = C_{11}X_1 + C_{12}X_2 + \cdots + C_{1p}X_p \\ Z_2 = C_{21}X_1 + C_{22}X_2 + \cdots + C_{2p}X_p \\ Z_3 = C_{31}X_1 + C_{32}X_2 + \cdots + C_{3p}X_p \\ \cdots \\ Z_p = C_{p1}X_1 + C_{p2}X_2 + \cdots + C_{pp}X_p \end{cases} \tag{1}$$

For any i, all $C_{i1}^2 + C_{i2}^2 + C_{i3}^2 + \cdots + C_{ip}^2 = 1$.

The specific operation steps of PCA are as follows:

1. Standardize the original data to prevent the dimensional difference between different indicators from affecting the results.
2. Establish the correlation coefficient matrix between variables R.
3. Calculate the eigenvalues λ_j and eigenvectors of the correlation coefficient matrix R.
4. Get the principal component factors and calculate the comprehensive score.

Calculate the information contribution rate and cumulative contribution rate of eigenvalue λ_j ($j = 1, 2, \cdots, m$), respectively. Among them, $b_j = \lambda_j / \sum_{k=1}^{m} \lambda_k$ is the information contribution rate of the main component, and $a_p = \sum_{k=1}^{p} \lambda_k / \sum_{k=1}^{m} \lambda_k$ is the cumulative contribution rate of each principal component. When a_p reaches 0.85, it shows that the influencing factors have been able to explain the original variables. Therefore, the first p variables are selected as the principal component factors.

3.2. Principle and steps of BP neural network

ANN has always been a hot academic frontier research and learning field in the international academic community. It has been widely used in various fields, such as power load forecasting (Xie *et al.* 2020; Wang *et al.* 2021c), prediction of river runoff (Ghose *et al.* 2018), and prediction of return rate of capital market (Galeshchuk 2016). The common structures of the neural network are RBF neural network (Huang & Yang 2020), particle swarm optimization neural network (Li *et al.* 2017), BP neural network (Ma *et al.* 2017), and genetic neural network (Ding *et al.* 2014).

Figure 1 shows the mathematical model of a single neuron.

A typical neural network generally consists of three to four layers, which are input layer, hidden layer for data processing, and output layer for result output (Figure 2). The relationship between input and output can be expressed by the following formula:

$$net_i = \sum_{j=1}^{n} w_{ij}x_j - \theta = \sum_{j=0}^{n} w_{ij}x_j \tag{2}$$

where θ is the threshold value, $\sum_{j=1}^{n} w_{ij}x_j$ is the net activation amount and the sum of each neuron input multiplied by its weight. In the neural network, the activation function is the mapping function between the net activation quantity and the output, which the formula is $y_i = f(net_i)$. Some common activation functions are linear function $y = kx + c$, S-shape function $f(x) = 1/1 + e^{-ax}$, and bipolar S-shape function $f(x) = 2/1 + e^{-ax} - 1$.

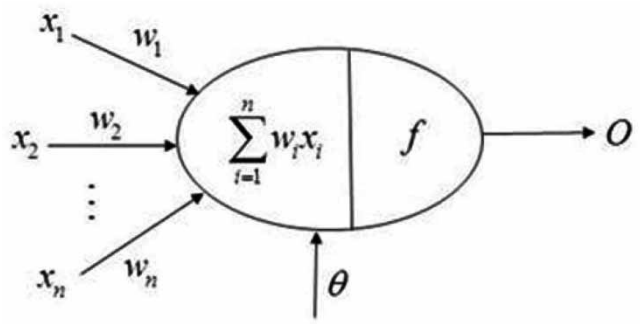


Figure 1 | Mathematical model of single neuron.

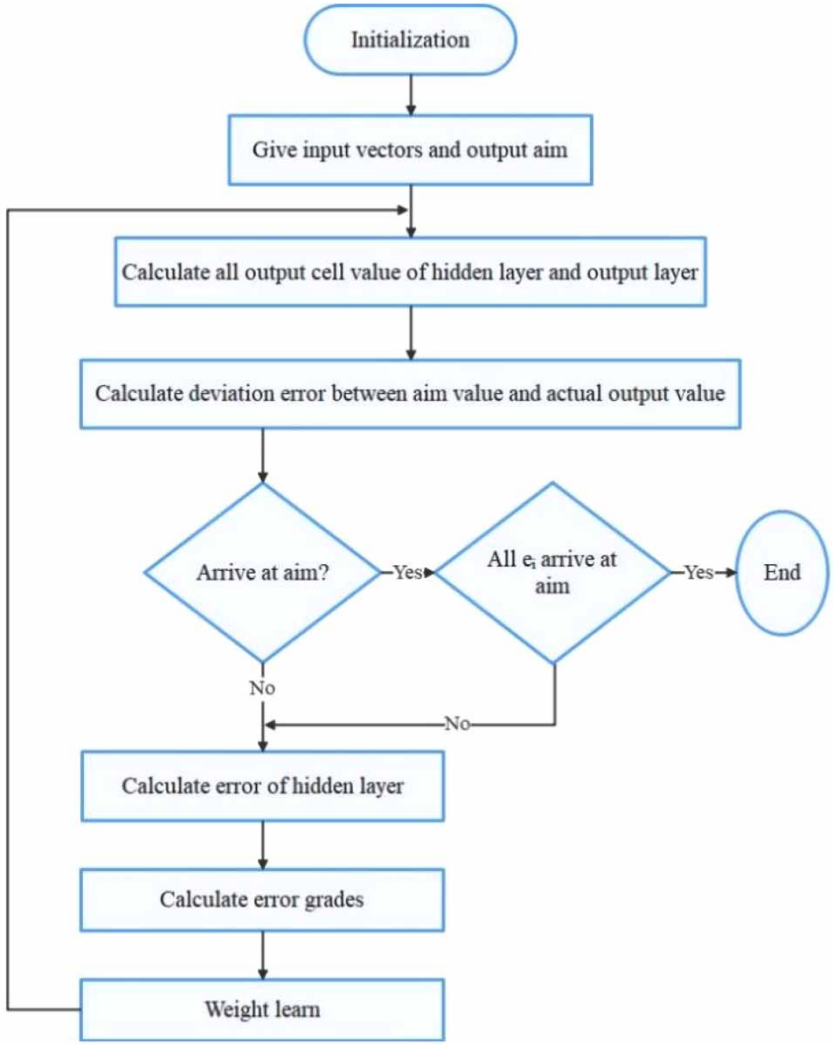


Figure 2 | Flow chart of BP neural network model.

The structure of ANN is shown in the literatures (Adesanya *et al.* 2021; Cicek & Ozturk 2021). The neural networks can be divided into two states: learning state and working state. The learning state is to adjust the weight of neural network to make the output close to the real value, while the working state is to use the established network to classify and predict without changing the weight of neural network. The learning mode of neural network is tutored learning. Meanwhile, the weight of the network is adjusted by the difference between the actual output and the expected output of the network to make

the model fit as accurately as possible. The BP neural algorithm used in this study is mainly composed of two parts: the forward propagation of signal and the BP of error. The basic idea of the algorithm is to use the gradient search technique to minimize the mean square error between the actual output value and the expected output value, and according to the error propagation layer by layer, the error estimation of each layer can be obtained. Then, the weight of each layer is modified until it reaches the acceptable range.

Suppose: there are n neurons in the input layer, p neurons in the hidden layer, and q neurons in the output layer, and parameter and function representation of BP model are shown in Table 1.

The steps of BP neural algorithm are as follows.

Step 1. Calculate the input and output of each layer:

$$hi_h(k) = \sum_{i=0}^{n} w_{hi} x_i(k) \tag{3}$$

$$ho_h(k) = f(hi_h(k)) \tag{4}$$

$$yi_o(k) = \sum_{h=0}^{p} w_{oh} ho_h(k) \tag{5}$$

$$yo_o(k) = f(yi_o(k)) \tag{6}$$

Step 2. Use the expected output and the actual output of the network to calculate the partial derivatives of the error function to the neurons in the output layer:

$$\frac{\partial yi_o(k)}{\partial w_{oh}} = \frac{\partial\left(\sum_{h}^{p} w_{oh} ho_h(k)\right)}{\partial w_{oh}} = ho_h(k) \tag{7}$$

Step 3. Using the connection weight from hidden layer to output layer, the output layer $\delta_0(k)$, and the output of the hidden layer to calculate the partial derivative of the error function to each neuron in the hidden layer $\delta_h(k)$:

$$\frac{\partial hi_h(k)}{\partial w_{hi}} = \frac{\partial\left(\sum_{i=0}^{n} w_{hi} x_i(k)\right)}{\partial w_{hi}} = x_i(k) \tag{8}$$

Step 4. Using the error function of each neuron in the output layer and the output of each neuron in the hidden layer to correct the connection weight $w_{oh}(k)$:

$$w_{oh}^{N+1} = w_{oh}^{N} + \mu \delta_o(k) ho_h(k) \tag{9}$$

The parameter μ is the set learning rate.

Table 1 | Parameter and function representation of the BP model

Input vector	$x = (x_1, x_2, \cdots, x_n)$	Hidden layer input vector	$hi = (hi_1, hi_2, \cdots, hi_p)$
Hidden layer output vector	$ho = (ho_1, ho_2, \cdots, ho_p)$	Output layer input vector	$yi = (yi_1, yi_2, \cdots, yi_q)$
Output layer output vector	$yo = (yo_1, yo_2, \cdots, yo_q)$	Expected output function	$d_o = (d_1, d_2, \cdots, d_q)$
Error function	$e = \frac{1}{2}\sum_{o=1}^{q} (d_o(k) - yo_o(k))^2$		

Step 5. Calculate the global error:

$$E = \frac{1}{2m}\sum_{k=1}^{m}\sum_{o=1}^{q}(d_o(k) - y_o(k))^2 \tag{10}$$

Step 6. Judge whether the network error meets the set accuracy requirements. When the error reaches the preset accuracy or the number of learning times is greater than the set maximum number, the algorithm ends. Otherwise, the next learning sample and the corresponding expected output must be selected to return to the next round of learning.

3.3. Test of model accuracy

The methods to test the accuracy of the model generally include mean absolute error test, mean relative error test, residual test, and posterior variance test. In this study, the average absolute error and the average relative error are used to test the prediction results of the research model. The absolute error calculation formula is as follows:

$$\Delta_k = |x_k - L_k| \tag{11}$$

The corresponding mean absolute error is as follows:

$$\overline{\Delta} = \frac{1}{n}\sum_{k=1}^{n}\Delta_k \tag{12}$$

The formula for the relative error δ_k is as follows:

$$\delta_k = \frac{|x_k - L_k|}{L_k} \times 100\% \tag{13}$$

The formula for calculating the average relative error $\overline{\delta}$ is as follows:

$$\overline{\delta} = \frac{1}{n}\sum_{k=1}^{n}\delta_k \times 100\% \tag{14}$$

In the above formulas, x_k and L_k are the predicted value and the true value of the kth period, respectively, and n is the number of periods of the test period.

4. ANALYSIS OF WATER RESOURCES PREDICTION MODEL IN TAIYUAN CITY

This section is divided into six parts. The first subsection introduces the physical geography and social economy of Taiyuan City, which enables to have a further understanding of the research object's situation. The second part is an introduction to the dataset. The third part is the analysis of the results of the PCA model. Next, according to the selected influencing factors, the details of the BP neural network model are provided in the fourth subsection. In the following subsection, the prediction effect of the proposed model is compared with other latest models. Finally, the future water consumption of Taiyuan City is reasonably predicted in the last subsection. Similarly, the data processing flow in the study is as follows: firstly, we analyze the influencing factors of domestic water demand in Taiyuan and find nine closely related influencing factors. Next, the influencing factor dataset was selected from the Taiyuan Water Resources Bulletin as well as the Statistical Bulletin on the National Economic and Social Development of Taiyuan City. Through PCA, the nine influencing factors are reduced to obtain the data of the first three main components. Then, the first three main components are input into the BP neural network model for corresponding training and verification so as to obtain the final prediction results. The relevant flow chart of the paper is shown in Figure 2.

4.1. Local physical geography and social economy

4.1.1. Physical geography

Taiyuan City is located in the central part of Shanxi Province in China, with an average altitude of about 800 m. The terrain is high in the north and low in the south. It is adjacent to Taihang Mountain in the west, Luliang Mountain in the east, Houlianyunzhong mountain, and Xizhou mountain in the east. The climate type is the north temperate continental climate. The annual precipitation distribution is very uneven, and the temperature varies greatly during the day. The annual precipitation is mainly concentrated in summer, and the winter is long, cold, and dry. While in spring, the temperature soars rapidly, and there are more gale days. At this time, the rain belt has not yet moved to northern China, evaporation is high, and there will be a spring drought with an annual average precipitation of 468.4 mm.

4.1.2. Social and economic situation

Taiyuan is a city with a long history, connecting Jinzhong City in the south and Yangquan City in the east. As the second-largest tributary of the Yellow River, the Fenhe River flows through Taiyuan City from north to south. Taiyuan has always been the commercial center of our country. 'Shanxi Merchants' are famous all over the world, and it is also a heavy industry base and resource base, containing iron, copper, lead, manganese, and other metals. In 2017, the city's total population reached 4.3797 million, and the urbanization rate reached 84.7%. In 2018, Taiyuan's gross domestic product (GDP) was 388.448 billion yuan, with a growth rate of 9.2%. By the end of 2018, Taiyuan City has jurisdiction over six municipal districts, three counties, and a total of 54 streets.

4.2. Introduction to the dataset

The dataset is selected from the Taiyuan National Economic and Social Development Statistical Bulletin and Taiyuan Water Resources Bulletin, from which nine indicators (precipitation resident population, GDP, total water resources, total industrial output value, total agricultural output value, average temperature, annual average relative humidity, and annual sunshine hours) from 2012 to 2020, as the influencing factors of water demand in Taiyuan City, have been selected. The dataset is shown in Table 2.

4.3. Analysis of the results of the PCA model

In this study, we set the following variables to replace the indicator values. x_1 is the annual rainfall (m), x_2 is the permanent population (million people), x_3 is the gross regional product (100 billion yuan), x_4 is the total water resources (100 billion m^3), x_5 is the total industrial output value (10^4 yuan), x_6 is the total agricultural output value (10^4 yuan), x_7 is the annual average

Table 2 | Main factors affecting water resources demand in Taiyuan City

Year	Precipitation (m)	Resident population (million)	GDP (100 billion yuan)	Water resources (100 million m^3)	Industrial output value (10^4 yuan)	Agricultural output value (10^4 yuan)	Average temperature (°C)	Average relative humidity (%)	Annual sunshine hours (h)	Water consumption (100 million m^3)
2009	6.251	3.5018	1.5452	4.33	29,453,287	503,058	11.1	54	2,448.6	5.45
2010	3.766	4.2047	1.7780	3.54	35,382,271	549,368	11.3	52	2,413.4	5.66
2011	4.966	4.2354	2.0801	5.52	41,971,398	631,629	10.8	53	2,372.8	6.18
2012	4.278	4.2563	2.3114	5.05	45,368,757	674,967	10.7	51	2,618.6	6.17
2013	4.873	4.2777	2.4128	4.66	47,210,239	722,383	11.2	56	2,627.1	6.43
2014	4.287	4.2989	2.5311	4.49	47,639,725	726667	10.9	58	2,513.5	6.89
2015	4.036	4.3187	2.7353	4.27	52,029,777	703,090	11.3	58	2,710.3	7.45
2016	5.284	4.3444	2.9556	6.12	58,762,828	729,905	11.2	59	2,730.4	7.76
2017	5.309	4.3797	3.3821	5.42	68,420,760	755,015	11.5	56	2,511.8	7.78
2018	4.509	4.4215	3.8845	6.153	82,812,060	755,056	10.5	53	2,740.4	7.82
2019	3.861	4.4619	4.028	5.09	97,998,460	755,098	10.6	57	2,756.6	8.24
2020	5.116	5.3041	4.1533	5.66	113,040,360	755,130	10.3	54	2,748.3	8.14

temperature (centigrade), x_8 is the annual average relative humidity (%), x_9 is the annual sunshine hours (h). Taking these nine indicators as the influencing factors, SPSS 26 is used to standardize the data, and then PCA is carried out to calculate the eigenvector, eigenvalue, and cumulative variance contribution rate. Tables 3 and 4 show the results of the PCA of water demand affecting Taiyuan City.

From Table 3, it can be seen that the eigenvalue of the first principal component is 4.799, the eigenvalue of the second principal component is 1.626, and the eigenvalue of the third principal component is 1.255. The formula of variance percentage is the quotient of the variance of each element and the sum of the variances of all elements. According to the principle that the first k elements with eigenvalues greater than 1 or cumulative contribution percentages greater than 85% are selected as principal components, so the first three components are selected as principal ones.

From Table 4, it can be seen that the first principal component has the highest correlation with GDP, gross industrial output value, gross agricultural output value, higher correlation with the number of population and sunshine hours per year, and a smaller and negative correlation with precipitation. The second principal component has the highest correlation with precipitation and is positively correlated. The third principal component has a higher correlation with the average temperature per year.

4.4. Construction of the BP neural network model

After selecting the first principal component second principal component, third principal component as the main factors affecting the water demand of Taiyuan City, this work takes the historical data as the input sample of BP neural network

Table 3 | Explanation of total variance

		Initial eigenvalue		
Calculate order	**Component**	**Total**	**Variance percentage**	**Cumulative %**
Rescaling	x_1	4.799	53.322	53.322
	x_2	1.626	18.067	71.390
	x_3	1.255	13.944	85.333
	x_4	0.762	8.461	93.795
	x_5	0.355	3.941	97.735
	x_6	0.148	1.647	99.382
	x_7	0.056	0.617	100.000
	x_8	3.717E-5	0.000	100.000
	x_9	−6.129E-17	−6.810E-16	100.000

Table 4 | Composition matrix[a]

	Rescaling components		
New variable values after standardization	**1**	**2**	**3**
Annual precipitation (mm)	−0.112	0.985	−0.046
Resident population (million)	0.776	−0.546	−0.176
GDP (100 billion)	0.971	0.058	0.022
Total water resources (100 million m^3)	0.579	0.493	−0.601
Total industrial output value (10^4 yuan)	0.957	0.083	−0.006
Total agricultural output value (10^4 yuan)	0.947	−0.119	−0.160
Annual average temperature (°C)	0.356	0.166	0.837
Annual average relative humidity	0.706	0.252	0.352
Annual sunshine hours (h)	0.684	−0.019	0.101

Extraction method: PCA.

[a]Three principal components were extracted.

model and takes the annual water consumption of Taiyuan from 2009 to 2017 as the output sample. In applying traditional machine learning methods, the division of training and test sets is generally 7:3. In our study, there are 12 years in total, so the training set should generally contain eight to nine samples. To make the model training relatively adequate and the test sets comparable, we set the data from 2009–2017 as the training set and the data from 2018–2020 as the test set. The test set is used to observe whether the error between the output results of the neural network and the actual results reaches the expected results. Finally, the model is applied to the prediction of water resources demand in Taiyuan City from 2021 to 2030 to observe the changes of water resources demand in the future. Set the number of iterations to 10,000 and the target error as 10^{-7}, then Figure 3 shows the error image during the training period.

It can be seen from Figure 3 that the final minimum error of BP neural network is 3.6545×10^{-13}, which meets the expected requirements, and then we test the validation set. The results are shown in Table 5.

4.5. Compared with other prediction algorithms

To prove that the BP neural network model has better accuracy in the prediction data, this study compares with the results of the Grey prediction model (Li *et al.* 2020), the optimized Grey–Markov model (Wang *et al.* 2019), the time-series prediction model (Sena & Nagwani 2016), ARIMA model (Jamil 2020), and the NARX method (Mohammad & Pezhman 2019) and takes the water consumption data of Taiyuan from 2009 to 2017 as the training sample and the water consumption data of Taiyuan from 2018 to 2020 as the prediction sample. Table 6 shows the results of different models for the forecast of water demand in Taiyuan City for the 3 years 2018, 2019, and 2020 and the error comparison.

As it can be seen from Table 6, the average absolute error of the PCA-BP model is 0.07, and the average relative error is 0.92%. Meanwhile, compared with the optimized Grey–Markov Chain model, Gray prediction model, serial regression model, ARIMA model, NARX model, LSTM model, NAR model, and ANN model, the average absolute error has decreased by 0.6277, 0.5785, 0.2913, 0.3869, 0.0179, 0.12, 0.554, and 0.163, and the average relative errors decreased by 7.71, 7.11, 3.56, 4.76, 0.16, 1.458, 6.71, and 1.99%, respectively. From the comparison results, it can be seen that the

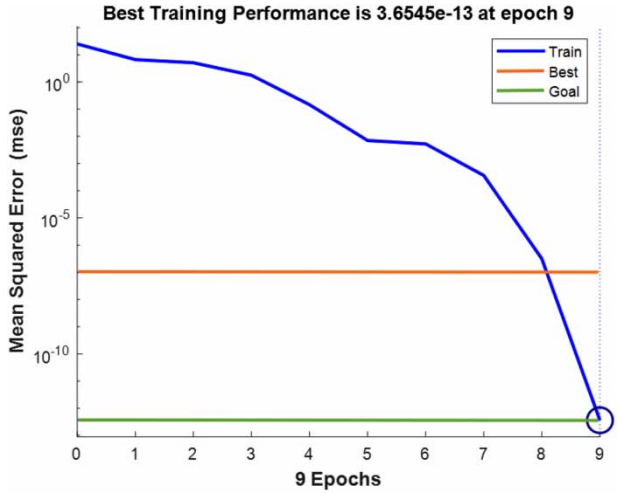


Figure 3 | Error image during training.

Table 5 | PCA-BP neural network model prediction value and error analysis

Year	Actual water consumption (billion m³)	Prediction (billion m³)	Relative error (%)	Absolute error	Average relative error (%)	Mean absolute error
2018	7.82	7.69	1.66	0.13	0.92	0.07
2019	8.24	8.18	0.73	0.06		
2020	8.14	8.17	0.37	0.03		

Table 6 | Prediction values of different models and error analysis

Different models	Year	Actual water consumption (billion m^3)	Prediction (billion m^3)	Relative error (%)	Absolute error	Average relative error (%)	Mean absolute error
Grey model	2018	7.82	8.3528	6.8	0.5328	8.63	0.6977
	2019	8.24	8.758	6.28	0.518		
	2020	8.14	9.183	12.81	1.043		
Optimized Grey-Markov model	2018	7.82	8.3221	6.42	0.5021	8.03	0.6485
	2019	8.24	8.6972	5.55	0.4572		
	2020	8.14	9.1261	12.11	0.9861		
Series regression model	2018	7.82	8.1456	4.16	0.3256	4.48	0.3613
	2019	8.24	8.4279	2.28	0.1879		
	2020	8.14	8.7103	7	0.5703		
ARIMA model	2018	7.82	8.3199	6.39	0.4999	5.68	0.4569
	2019	8.24	8.5902	4.25	0.3502		
	2020	8.14	8.6605	6.39	0.5205		
NARX method	2018	7.82	7.8931	0.935	0.0731	1.08	0.0872
	2019	8.24	8.3234	1.01	0.0834		
	2020	8.14	8.2451	1.29	0.1051		
LSTM method	2018	7.82	8.11	3.7	0.29	2.378	0.19
	2019	8.24	8.27	0.364	0.03		
	2020	8.14	8.39	3.07	0.25		
NAR method	2018	7.82	7.87	0.683	0.0534	7.63	0.624
	2019	8.24	7.19	12.74	1.05		
	2020	8.14	7.37	9.46	0.77		
PCA-ANN method	2018	7.82	8.13	3.96	0.31	2.91	0.233
	2019	8.24	8.20	0.0485	0.04		
	2020	8.14	8.49	4.29	0.35		

prediction results of the PCA-BP model are better than those of each of the latest models, which proves the superiority of the PCA-BP model, and the model can be used as an important method for future water demand prediction in Taiyuan City.

4.6. Forecast of future water consumption in Taiyuan City

On the basis of the test results of the PCA-BP model, the future water consumption in Taiyuan City can be reasonably predicted. According to the growth rate of GDP of Taiyuan City over the years, this study reasonably assumes that the growth rate of GDP of Taiyuan City over the years is 7% higher than that of the previous year, and the number of residents population (million) increases by 0.93% every year. The annual precipitation is the moving average of precipitation of Taiyuan City every 10 years, and the results are shown in Table 7.

According to the prediction results of the PCA-BP neural network model, the domestic water consumption of Taiyuan in 2025 and 2030 will be 962.85 and 1,053.59 billion m^3, respectively, which will increase by 18.3 and 29.4% compared with 814 million m^3 in 2020. In other words, it indicates that there will be a large gap in Taiyuan's water consumption in recent years.

Based on the above projections for water demand in Taiyuan City from 2021–2030, we conclude the following recommendations:

1. Strengthen the legal system and regulate water strictly in accordance with the laws.
2. Use groundwater correctly and effectively, and reasonably allocate water sources to different industries. After analyzing the proportion of groundwater use in various industries, redistribute the sources of water supply to different industries.
3. Strengthen the monitoring level of groundwater geological environment and improve the dynamic monitoring level of groundwater utilization in Taiyuan, so as to reduce or prevent environmental problems, and need to gradually improve the long-term monitoring process of groundwater quality to prevent water pollution.

Table 7 | Water demand forecast of Taiyuan City in the next 10 years

Year	Annual precipitation (m)	Resident Population (million)	GDP (100 billion)	Forecast demand (100 million m^3)
2021	4.6519	4.5624	4.4437	8.1700
2022	4.6205	4.6064	4.7547	8.4567
2023	4.6547	4.6492	5.0875	8.8848
2024	4.6329	4.6925	5.4436	9.2582
2025	4.6675	4.7360	5.8247	**9.6285**
2026	4.7307	4.7800	6.2323	9.9446
2027	4.6753	4.8244	6.6685	10.1300
2028	4.6110	4.8690	7.1352	10.2865
2029	4.6222	4.9142	7.6346	10.4240
2030	4.6983	4.9599	8.1690	**10.5359**

5. CONCLUSION

Water demand forecasting is a typical nonlinear problem with various influencing factors. The weight to be attributed to these variables is difficult to determine because they have complex relationships. This study proposes a water demand forecasting model based on principal component analysis and BP neural network (PCA-BP) and takes Taiyuan, a city in Shanxi Province, China, which has severe water shortage, as the study site. By using PCA, the main factors affecting the water demand in Taiyuan City are found to be annual precipitation, total local output, and local permanent population. After determining the main factors, the BP neural network model is selected to predict the water demand for the next few years and compares with optimized Grey–Markov chain model, Grey prediction model, serial regression model, ARIMA model, NARX model, LSTM model, NAR model, and ANN model, and the results show that the PCA-BP model has higher prediction accuracy and better performance.

The water resources prediction model based on the BP neural network and PCA also has some shortcomings. First, due to the limited original calculation data collected, it cannot make a complete and accurate evaluation of the calculation effect of the model, which is necessary for a mature model. Second, although the computational efficiency of the model meets the accuracy requirements to a certain extent, there is still room for further improvement. Nevertheless, the model proposed in this study also has a certain promotion and reference significance for the prediction of water resources demand. We believe that with the further development of the research, the model will become more promising and specific, which will better assist the research topic of water resources demand to forecast. The future research work mainly focuses on the parameter adjustment and optimization of neural network so as to improve the accuracy of the model to a greater extent.

ACKNOWLEDGEMENTS

It was supported by the Open Research Fund of State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research (grant no. IWHR-SKL-201905).

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Adamowski, J., Chan, H. F., Prasher, S. O., Ozga-Zielinski, B. & Sliusarieva, A. 2012 Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resour. Manag.* **48** (1), 273–279.

Adesanya, E., Aladejare, A., Adediran, A., Lawal, A. & Illikainen, M. 2021 Predicting shrinkage of alkali-activated blast furnace-fly ash mortars using artificial neural network (ANN). *Cement Concrete Comp.* **124**, 104265.

Angelopoulos, D., Siskos, Y. & Psarras, J. E. 2019 Disaggregating time series on multiple criteria for robust forecasting: the case of long-term electricity demand in Greece. *Eur. J. Oper. Res.* **275** (1), 252–265.

Antunes, A., Andrade-Campos, A., Sardinha-Lourenço, A. & Oliveira, M. S. 2018 Short-term water demand forecasting using machine learning techniques. *J. Hydroinform.* **20** (6), 1343–1366.

Babel, M. S. & Shinde, V. R. 2011 Identifying prominent explanatory variables for water demand prediction using artificial neural networks: a case study of Bangkok. *Water Resour. Manag.* **25** (6), 1653–1676.

Bai, Y., Wang, P., Li, C., Xie, J. & Wang, Y. 2014 A multi-scale relevance vector regression approach for daily urban water demand forecasting. *J. Hydrol.* **517**, 236–245.

Bata, M., Carriveau, R. & Ting, D. S.-K. 2020 Short-term water demand forecasting using hybrid supervised and unsupervised machine learning model. *Smart Water* **5** (1), 1–18.

Billings, B. & Jones, C. 2008 *Forecasting Urban Water Demand*, 2nd edn. American Water Works Association, Denver, CO.

Braik, M. S. 2021 Chameleon swarm algorithm: a bio-inspired optimizer for solving engineering design problems. *Expert Syst. Appl.* **174**, 114685.

Braun, M., Bernard, T., Piller, O. & Sedehizade, F. 2014 24-hours demand forecasting based on SARIMA and support vector machines. *Procedia Eng.* **89**, 926–933.

Cicek, Z. & Ozturk, Z. K. 2021 Optimizing the artificial neural network parameters using a biased random key genetic algorithm for time series forecasting. *Appl. Soft Comput.* **102**, 107091.

Deng, W., Liu, H., Xu, J., Zhao, H. & Song, Y. 2020a An improved quantum-Inspired differential evolution algorithm for deep belief network. *IEEE Trans. Instrum. Meas.* **69** (10), 7319–7327.

Deng, W., Xu, J., Song, Y. & Zhao, H. 2020b An effective improved co-evolution ant colony optimisation algorithm with multi-strategies and its application. *Int. J. Bio-Inspir. Com.* **16** (3), 158–170.

Deng, W., Xu, J., Gao, X.-Z. & Zhao, H. 2020c An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems. *IEEE Trans. Syst. Man Cybern. Syst.* doi:10.1109/TSMC.2020.3030792.

Deng, W., Xu, J., Zhao, H. & Song, Y. 2020d A novel gate resource allocation method using improved PSO-Based QEA. *IEEE Trans. Intell. Transp. Syst.* doi:10.1109/TITS.2020.3025796.

Deng, A., Wang, Z., Liu, H. & Wu, T. 2020e A bio-inspired algorithm for a classical water resources allocation problem based on adleman-Lipton model. *Desalin. Water Treat.* **185**, 168–174.

Deng, W., Shang, S., Cai, X., Zhao, H., Zhou, Y., Chen, H. & Deng, W. 2021a Quantum differential evolution with cooperative coevolution framework and hybrid mutation strategy for large scale optimization. *Knowl. Based Syst.* **224**, 107080.

Deng, W., Xu, J., Song, Y. & Zhao, H. 2021b Differential evolution algorithm with wavelet basis function and optimal mutation strategy for complex optimization problem. *Appl. Soft Comput.* **100**, 106724.

Ding, S. 2018 A novel self-adapting intelligent grey model for forecasting China's natural-gas demand. *Energies* **162**, 393–407.

Ding, Y. R., Cai, Y. J., Sun, P. D. & Chen, B. 2014 The use of combined neural networks and genetic algorithms for prediction of river water quality. *J. Appl. Res. Technol.* **12** (3), 493–499.

Donkor, E. A., Mazzuchi, T. A., Soyer, R. & Roberson, J. A. 2014 Urban water demand forecasting: review of methods and models. *J. Water Resour. Plan. Manag.* **140** (2), 146–159.

Du, B., Zhou, Q., Guo, J., Guo, S. & Wang, L. 2021 Deep learning with long short-term memory neural networks combining wavelet transform and principal component analysis for daily urban water demand forecasting. *Expert. Syst. Appl.* **171**, 114571.

Galeshchuk, S. 2016 Neural networks performance in exchange rate prediction. *Neurocomputing* **172**, 446–452.

Ghose, D., Das, U. & Roy, P. 2018 Modeling response of runoff and evapotranspiration for predicting water table depth in arid region using dynamic recurrent neural network. *Groundw. Sustain. Dev.* **6**, 263–269.

Guo, W., Liu, T., Dai, F. & Xu, P. 2020 An improved whale optimization algorithm for forecasting water resources demand. *Appl. Soft Comput.* **86**, 105925.

Heidari, A. A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M. M. & Chen, H. 2019 Harris hawks optimization: algorithm and applications. *Future Gener. Comput. Syst.* **97**, 849–872.

Hernandez, L., Baladron, C., Aguiar, J. M., Carro, B., Sanchez-Esguevillas, A. J., Lloret, J. & Massana, J. 2014 A survey on electric power demand forecasting: future trends in smart grids, microgrids and smart buildings. *IEEE Commun. Surv. Tutor.* **16** (3), 1460–1495.

Herrera, M., Torgo, L., Izquierdo, J. & Pérez-García, R. 2010 Predictive models for forecasting hourly urban water demand. *J. Hydrol.* **387** (1), 141–150.

Huang, W. & Yang, Y. 2020 Water quality sensor model based on an optimization method of RBF neural network. *Comput. Water, Energy, and Environ. Eng.* **9** (1), 1–11.

Jamil, R. 2020 Hydroelectricity consumption forecast for Pakistan using ARIMA modeling and supply-demand analysis for the year 2030. *Renew. Energy* **154**, 1–10.

Kandissounon, G. A., Karla, A. & Ahmad, S. 2018 Integrating system dynamics and remote sensing to estimate future water usage and average surface runoff in Lagos, Nigeria. *Civil Eng. J.* **4** (2), 378–393.

Lalwani, S., Sharma, H., Satapathy, S. C., Deep, K. & Bansal, J. C. 2019 A survey on parallel particle swarm optimization algorithms. *Arab. J. Sci. Eng.* **44** (4), 2899–2923.

Li, M., Wu, W., Chen, B., Guan, L. & Wu, Y. 2017 Water quality evaluation using back propagation artificial neural network based on self-adaptive particle swarm optimization algorithm and chaos theory. *Comput. Water, Energy, and Environ. Eng.* **6** (3), 229–242.

Li, S., Zeng, B., Ma, X. & Zhang, D. 2020 A novel grey model with a three-parameter background value and its application in forecasting average annual water consumption per capita in urban areas along the Yangtze river basin. *J. Grey Syst.* **32** (1), 118–132.

Li, R., Chang, Y. & Wang, Z. 2021 Study on optimal allocation of water resources in Dujiangyan irrigation district of China based on improved genetic algorithm. *Water Suppl.* **21** (6), 2989–2999.
Lv, Z. 2014 Application of time series analysis on the annual precipitation of Zhengzhou city. *S. N. Water Transfers Water Sci. Technol.* **12**, 35–37.
Ma, D., Zhou, T., Chen, J., Qi, S., Shahzad, M. A. & Xiao, Z. 2017 Supercritical water heat transfer coefficient prediction analysis based on BP neural network. *Nucl. Eng. Des.* **320**, 400–408.
Manju, S. & Sagar, N. 2017 Renewable energy integrated desalination: a sustainable solution to overcome future fresh-water scarcity in India. *Renew Sust. Energ. Rev.* **73**, 594–609.
Mirjalili, S. 2015a Moth-flame optimization algorithm: a novel nature-inspired heuristic paradigm. *Knowl. Based Syst.* **89**, 228–249.
Mirjalili, S. 2015b The ant lion optimizer. *Adv. Eng. Softw.* **83**, 80–98.
Mohammad, E. B. & Pezhman, M. M. 2019 Extended linear and non-linear auto-regressive models for forecasting the urban water consumption of a fast-growing city in an arid region. *Sustain. Cities Soc.* **48**, 101585.
Oliveira, P. J., Steffen, J. L. & Cheung, P. 2017 Parameter estimation of seasonal ARIMA models for water demand forecasting using the harmony search algorithm. *Procedia Eng.* **186**, 177–185.
Polebitski, A. S. & Palmer, R. N. 2010 Seasonal residential water demand forecasting for census tracts. *J. Water Resour. Plan. Manag.* **136** (1), 27–36.
Rehman, S. A. U., Cai, Y., Fazal, R., Walasai, G. D. & Mirjat, N. H. 2017 An integrated modeling approach for forecasting long-term energy demand in Pakistan. *Energies* **10** (11), 1868.
Ren, X., Wang, X., Wang, Z. & Wu, T. 2020 Parallel DNA algorithms of generalized traveling salesman problem-Based bioinspired computing model. *Int. J. Comput. Intell. Syst.* **14** (1), 228–237.
Sanchez, G. M., Terando, A., Smith, J. W., García, A. M., Wagner, C. R. & Meentemeyer, R. K. 2020 Forecasting water demand across a rapidly urbanizing region. *Sci. Total Environ.* **730**, 139050.
Sena, D. & Nagwani, N. K. 2016 A time-series forecasting-based prediction model to estimate groundwater levels in India. *Res. Commun.* **111** (6), 1083–1090.
Shabani, S., Yousefi, P. & Naser, G. 2017 Support vector machines in urban water demand forecasting using phase space reconstruction. *Procedia Eng.* **186** (186), 537–543.
Teng, Z., Lv, J. & Guo, L. 2019 An improved hybrid grey wolf optimization algorithm. *Soft Comput.* **23**, 6617–6631.
Tian, D., Martinez, C. J. & Asefa, T. 2016 Improving short-term urban water demand forecasts with reforecast analog ensembles. *J. Water Resour. Plan. Manag.* **142** (6), 4016008.
Tsiliyannis, C. A. 2018 Markov chain modeling and forecasting of product returns in remanufacturing based on stock mean-age. *Eur. J. Oper. Res.* **271** (2), 474–489.
Wang, P., Bai, Y., Li, C., Wang, Y. & Xie, J. 2015 Urban daily water consumption forecasting based on variable structure support vector machine. *J. Basic Sci. Eng.* **23** (5), 895–901.
Wang, H., Wang, W., Cui, Z., Zhou, X., Zhao, J. & Li, Y. 2018 A new dynamic firefly algorithm for demand estimation of water resources. *Inform. Sci.* **438**, 95–106.
Wang, Z., Ren, X., Ji, Z., Huang, W. & Wu, T. 2019 A novel bio-heuristic computing algorithm to solve the capacitated vehicle routing problem based on Adleman-Lipton model. *BioSystems* **184**, 103997.
Wang, Z., Bao, X. & Wu, T. 2021a A parallel bioinspired algorithm for Chinese postman problem based on molecular computing. *Comput. Intell. Neurosci.* **2021**, 1–13.
Wang, Z., Wang, D., Bao, X. & Wu, T. 2021b A parallel biological computing algorithm to solve the vertex coloring problem with polynomial time complexity. *J. Intell. Fuzzy Syst.* **40** (3), 1–11.
Wang, Z., Wu, X., Wang, H. & Wu, T. 2021c Prediction and analysis of domestic water consumption based on optimized grey and Markov model. *Water Suppl.* **21** (7), 3887–3899. doi:10.2166/ws.2021.146.
Wong, J. S., Zhang, Q. & Chen, Y. D. 2010 Statistical modeling of daily urban water consumption in Hong Kong: trend, changing patterns, and forecast. *Water Resour. Res.* **46** (3), W03506.
Wu, X., Wang, Z. C., Wu, T. H. & Bao, X. G. 2021 Solving the family traveling salesperson problem in the Adleman-Lipton model based on DNA computing. *IEEE Trans. Nanobiosci.* doi:10.1109/TNB.2021.3109067.
Xiang, X. & Jia, S. 2019 China's water-energy nexus: assessment of water-related energy use. *Resour. Conserv. Recycl.* **144**, 32–38.
Xie, K., Yi, H., Hu, G., Li, L. & Fan, Z. 2020 Short-term power load forecasting based on Elman neural network with particle swarm optimization. *Neurocomputing* **416**, 136–142.
Zhu, B. & Wei, Y. 2013 Carbon price forecasting with a novel hybrid ARIMA and least squares support vector machines methodology. *OMEGA – Int. J. Manag. Sci.* **41** (3), 517–524.
Ziervogel, G., Johnston, P., Matthew, M. & Mukheibir, P. 2010 Using climate information for supporting climate change adaptation in water resource management in South Africa. *Clim. Change* **103** (3), 537–554.

First received 5 August 2021; accepted in revised form 19 October 2021. Available online 1 November 2021