Application and comparison of three data-driven models for groundwater level dynamics prediction in Shijiazhuang Plain

Jiyang Tian[1], Jia Liu[1,2], Chuanzhe Li[1], Shuanghu Cheng[3], Yang Wang[1], Fuliang Yu[1]

1. State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, No.1 Fuxing Road, Haidian District, Beijing 100038, China; 2. State Key Laboratory of Hydrology-Water Resource and Hydraulic Engineering, Hohai University, No.1 Xikang Road, Nanjing 210098, China; 3. Hydrology and Water Resources Survey Bureau of Hebei Province, Shijiazhuang 050031，China.

Abstract: Evaluation and prediction of groundwater levels via models can help manage groundwater resources. To investigate and predict the variation of groundwater dynamics in the plain of Shijiazhuang, the capital city of Hebei, three different data-driven models for researching the dynamics are assessed, including Multiple Linear Regression (MLR), Back-Propagation Artificial Neural Network (BPANN) and Support Vector Machines (SVM). Groundwater depth, precipitation, evaporation, groundwater exploitation, grain yield and Gross Domestic Product (GDP) records from 1984 to 2013 are used. We discuss the modeling process and accuracy of the three methods in the assessment of their relative advantages and disadvantages, based on Absolute Error (ABE), Relative Error (RE), Maximum Error (ME) and Average Error (AVE). The results showed that both SVM and BPANN models had sufficiently high accuracy in reproducing groundwater levels, while SVM performed better. This may provide a method and reference for forecasting of groundwater resources in this region.
Key wards: groundwater depth; prediction; MLR model; SVM model; BPANN model.

1. Introduction

In arid and semi-arid regions like North China, groundwater plays a more important role in agricultural, industrial and environmental uses than other regions. Because groundwater can cover the shortage of water which is caused by uneven distribution of surface water in time and space (Cao et al, 2013). However, the groundwater level has decreased dramatically because of over-exploitation in North China, which has been one of the most serious overdraft areas (Yuan and Shen, 2013). This phenomenon has attracted researchers' extensive attentions.

Groundwater dynamic forecasting is an important method for groundwater management. Physical descriptive model and data-driven model are two classes of dynamic prediction models (Knotters and Bierkens, 2000). However, considering the detailed data requirements in the application of the physical descriptive model, the data-driven model is a good choice when the data are in short and the groundwater system is complex. Multiple Linear Regression (MLR), Back-Propagation Artificial Neural Network (BPANN) and Support Vector Machines (SVM) are popular data-driven models used for forecasting the groundwater level (Coppola et al, 2003; Lin et al, 2006; Shiri et al, 2013). Among these methods, MLR is a linear regression model and the simplest one, while the other two models are nonlinear regression model and more complex. Some studies indicate that BPANN model and SVM model perform better, especially the SVM model (Nayak et al, 2006; Shirmohammadi et al, 2013), but the predictions of these two models are not always good for all places due to the complexity of hydrogeological conditions and groundwater flow in different regions (Ping et al, 2013).

Shijiazhuang is the capital city of Hebei province in North China. The groundwater level declines about 30m during the past three decades in this city. There are some studies focusing on the reason why the groundwater level declines continuously and the negative effects caused by the decrease of groundwater level in Hebei province and North China plain. However, there is little research done on the selection of the data-driven models for groundwater dynamical prediction in Shijiazhuang plain.

In this study, MLR, BPANN and SVM models are used to forecast the groundwater depth and the prediction accuracy of these three models is compared. Natural, anthropic, biological and economic factors which may influence the groundwater depth are considered. It can provide a new method for the groundwater dynamical prediction in Shijiazhuang plain, and it can also be treated as a complement for research methods of the groundwater forecasting.

2. Study area and methods

2.1 Study area

Shijiazhuang plain (from lat 37°33′ to lat 38°42′N and from long 113°18′ to long 114°41′E) is located in central and eastern Hebei province and belongs to the Bohai Sea Economic Zone. The total area is 8157 km^2. The average annual prediction is about 490 mm and average air temperature ranges from -0.8℃ in the winter season to 25.9℃ in the summer season. The plain comprises Shijiazhuang city and 13 counties, including Xingtang, Luquan, Yuanshi, Gaoyi, Xinle, zhengding, Luancheng, Zhaoxian, Wuji, Gaocheng, Jinzhou, Xinji and Shenze. The location of the study area and the distribution of the monitoring wells and meteorological stations are showed in figure 1.

2.2 Data and indices

14 monitoring wells covering the whole study are are chosen to study the groundwater dynamic variations. The observational data of annual groundwater depth for the 14 monitoring wells, annual precipitation and evaporation data at 13 meteorological stations from 1984 to 2013 are used in this study. Data of the annual groundwater exploitation, grain yield and Gross Domestic Product (GDP) for the whole study area are obtained from 1984 to 2013. The observational data of precipitation and evaporation are provided by the National Meteorological Information Center of China Meteorological Administration (available at http://www.nmic.gov.cn/).

Prediction, evaporation, groundwater exploitation, grain yield and GDP are chosen to be the key factors, which are closely related to groundwater dynamic changes. Prediction and evaporation are main natural factors which influence groundwater recharge and loss. Groundwater exploitation is main anthropic factor which is the major reason for groundwater descending in the study area in recent decades (Liu et al, 2001). Grain yield is the main biological factor which reflects the crop water consumption, especially groundwater consumption in the study area (Cao et al, 2014). GDP is a comprehensive index of reflecting the regional economic strength and pressure of groundwater used.

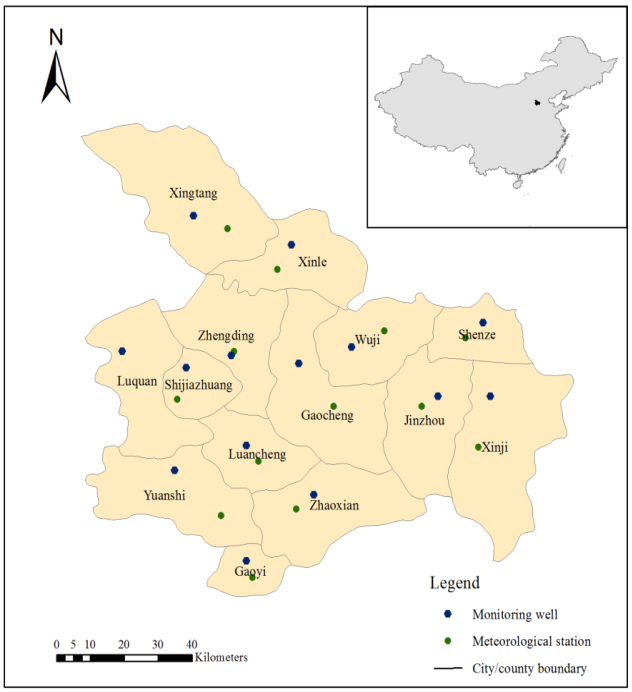

Fig. 1 Location of the study area and the distribution of the monitoring wells and meteorological stations

2.3 MLR model

The MLR model is often used for prediction by establishing the relationship between the forecast factors and the forecast objects. The general equation is as follows:

$$\hat{y} = p_0 + p_1 x_1 + \cdots + p_N x_N \quad (1)$$

Where p_k (k=0,…,N) are the parameters generally estimated by least squares and x_k (k=1,…,N) are the explanatory variables (forecast factors). $\hat{y}$ is the forecast object.

2.4 BPANN model

The Artificial Neural Network (ANN) is an information processing system composed of many nonlinear and densely interconnected processing elements or neurons, which is patterned after the biological nervous system. This mathematical structure consists of input, hidden, and output layers with their nodes and activation functions. The back-propagation algorithm can effectively train the network (Rumelhart et al, 1986), and it does not require information about the complex nature of the underlying process under consideration to be explicitly described in mathematical form. If the neuron is the *j*th one in the present layer, while the inputs which it receives from the other n neurons are x_1, x_2,……,x_n, respectively in the previous layer. The connection weights between the *j*th neuron and the other n neurons are w_{1j}, w_{2j},……, w_{nj}, respectively. The mathematical expression is as follows:

$$y_j = f\left(\sum_{i=1}^{N} w_{ji} x_i + b_j\right) \tag{2}$$

In this study, the ANN model is built with one hidden layer trained by BPA .The activation function consists of a log-sigmoid function in the hidden layer and a linear function in the output layer. ANNs with this configuration are the most commonly used form, which have improved the extrapolation ability. The input layer, hidden layer and output layer have respectively five nodes, thirty nodes and one node. To avoid being captured in some minimum, a momentum term is added in the weight updating process, which diminishes drastic fluctuation in weight changes over consecutive iterations.

2.5 SVM model

The SVM model is based on Vapnik-Chervonenkis (VC) demension and structural risk minimum principle (Vapnik, 1995; Vapnik, 1998). SVM provides a new approach to solve the nonlinear and high dimension problem with small sample set. The basic idea of SVM is to use linear model to implement nonlinear class boundaries through some nonlinear mapping of the input vector into the high-dimensional feature space. Given a set of N samples of $\{\mathbf{x}_k, y_k\}_{k=1}^{N}$ ($\mathbf{x}_k$ is the input vector, y_k is the corresponding output value), and the regression function of SVM can be expressed as:

$$y = f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b \tag{3}$$

Where ϕ denotes a nonlinear transfer function that maps the input vectors into a high-dimensional feature space in which theoretically a simple linear regression can cope with the complex nonlinear regression of the input space, **w** is a weight vector and b is a bias. Vapnik (1995) introduced the convex quadratic optimization question to ensure that extreme solution is optimal, and a ε-insensitively loss function is added to Eq. (3):

$$\begin{aligned} &\min \phi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{k=1}^{N}\left(\xi_k + \xi_k^*\right) \\ &\text{subject to } \begin{cases} y_k - \mathbf{w}^T\phi(\mathbf{x}_k) - b \le \varepsilon + \xi_k \\ \mathbf{w}^T\phi(\mathbf{x}_k) + b - y_k \le \varepsilon + \xi_k^* \\ \xi_k, \xi_k^* \ge 0 \end{cases} \qquad k = 1,2,\cdots,N \end{aligned} \tag{4}$$

where ξ and ξ^* are slack variables that penalize training errors by the loss function over the error tolerance ε, and C is a positive trade-off parameter that determines the degree of the empirical error in the optimization problem. Eq. (4) is solved in a dual form using Lagrangian multipliers and Karush-Kuhn-Tucker (KKT) optimality condition. The input vectors are called support vectors. The dual Lagrangian form is:

$$W(\alpha, \alpha^*) = -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) + \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) \tag{5}$$

with the constraints,

$$\begin{cases} \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C \end{cases} \qquad (i = 1,2,\cdots,N) \tag{6}$$

where α and α^* are Lagrange multipliers, and the optimal desired weight vector of the regression hyperplane is:

$$\mathbf{w} = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x, x_i) \tag{7}$$

where $K(x, x_i)$ is the kernel function. The Eq. (3) can be expressed as:

$$y = f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x, x_i) + b \tag{8}$$

In general, Radial basis function (RBF) is used as kernel function (Liu et al, 2009; Safavi and Esmikhani, 2013):

$$K(x, x_i) = \exp(-\gamma \| x - x_i \|) \tag{9}$$

In this study, we define y as the groundwater depth which is an index of groundwater dynamic variation and the forecast object, and x_k (k=1,2,3,4,5) as the factors which may influence the dynamic variation of groundwater for MLR, BPANN and SVM models.

Twenty groups of data from 1984 to 2003 are used to build and train the models, while ten groups of data from 2004 to 2013 are used to test the models. All the data are normalized to a range of 0~1 to avoid disturbance of dimension. Absolute Error (ABE), Relative Error (RE), Maximum Error (ME) and Average Error (AVE) are used to assess the accuracy of three methods based on true value (TV) and simulation value (SV):

$$\begin{cases} \mathrm{ABE} = \mathrm{SV} - \mathrm{TV} \\ \mathrm{RE} = \mathrm{ABE} / \mathrm{TV} \\ \mathrm{ME} = \max(\mathrm{ABE}) \\ \mathrm{AVE} = |\mathrm{ABE}| / N \end{cases} \tag{10}$$

where N is the number of groups.

3. Results

3.1 Fitting results and errors

The equation of MLR model can be expressed as follows based on the twenty groups of data from 1984 to 2003:

$$\hat{y} = 0.0631 + 0.1015x_1 + 0.3349x_2 - 0.2435x_3 + 0.3515x_4 + 0.9132x_5 \tag{13}$$

where x_1 is precipitation, x_2 is groundwater exploitation, x_3 is evaporation, x_4 is grain yield and x_5 is GDP. $\hat{y}$ is the fitting results for groundwater depth.

Table 1 Fitting results and errors of MLR, BPANN and SVM models

Train groups' number	TV/m	MLR model			BPANN model			SVM model		
		SV/m	ABE/m	RE/%	SV/m	ABE/m	RE/%	SV/m	ABE/m	RE/%
1	12.99	15.03	2.04	15.70	13.69	0.69	5.34	13.51	0.52	3.97
2	13.87	15.64	1.77	12.78	11.37	-2.5	-18.03	14.94	1.07	7.71
3	15.18	14.61	-0.57	-3.76	14.89	-0.29	-1.92	16.01	0.83	5.45
4	16.52	15.13	-1.38	-8.36	16.56	0.05	0.3	17.23	0.71	4.33
5	16.37	16.28	-0.09	-0.56	15.93	-0.44	-2.67	17.48	1.11	6.78
6	17.01	16.46	-0.55	-3.21	17.72	0.71	4.16	17.95	0.94	5.53
7	16.24	17.98	1.74	10.74	16.56	0.32	1.98	17.47	1.23	7.60
8	16.91	17.13	0.22	1.28	17.51	0.6	3.53	18.3	1.39	8.19
9	18.62	17.32	-1.30	-6.98	19.06	0.44	2.36	19.84	1.22	6.54
10	19.74	17.96	-1.78	-9.00	20.85	1.11	5.64	21.21	1.47	7.45
11	21.18	19.01	-2.17	-10.26	20.12	-1.06	-5.02	21.03	-0.15	-0.71
12	20.61	18.51	-2.09	-10.16	20.29	-0.32	-1.55	21.76	1.15	5.59
13	17.86	19.47	1.61	9.00	15.41	-2.45	-13.72	18.99	1.13	6.33
14	18.56	21.21	2.65	14.27	16.03	-2.53	-13.62	19.01	0.45	2.44
15	20.35	21.95	1.60	7.89	21.21	0.87	4.26	20.03	-0.32	-1.55
16	21.93	23.20	1.27	5.81	22.88	0.95	4.34	22.58	0.65	2.98
17	22.90	22.31	-0.59	-2.57	23.8	0.9	3.95	23.53	0.63	2.77
18	24.33	24.58	0.25	1.03	25.57	1.24	5.1	25.97	1.64	6.73
19	25.81	24.01	-1.80	-6.98	25.52	-0.29	-1.12	24.17	-1.64	-6.35
20	26.29	25.45	-0.84	-3.18	26.93	0.64	2.44	24.43	-1.86	-7.06

The other two models are also trained based on the twenty groups of data from 1984 to 2003, and the optimal fitting results are adopted. The fitting results and errors of the three models are showed in Table1 and Figure2. The MEs of MLR model, BPANN model and SVM model are 2.65, -2.53 and -1.86, respectively. The AVEs of MLR model, BPANN model and SVM model are 1.32, 0.92 and 1.01. The three data-driven models all perform well. The SVM model has the best ability of generalization in the sample learning process, because the SVs fluctuate within a narrow range around TVs and all the REs are below 10%. The BPANN model performs better than MLR model as a whole and the BPANN model has the lowest AVEs, but the SVs fluctuate within a large range around TVs in some individual groups. The stability of fitting results for MLR model is worse than other two models. Good fitting results do not mean good prediction results for the data-driven models, so it's necessary to examine their prediction capabilities.

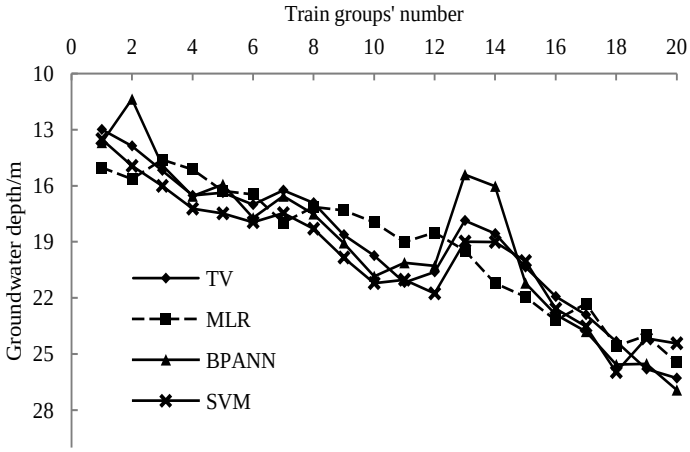

Fig. 2 Simulation results of groundwater depth with MLR, BPANN and SVM models

3.2 Prediction results and errors

Three models are tested based on the ten groups of data from 2004 to 2013. The Prediction results and errors are showed in Table 2 and Figure 3. The MEs of MLR model, BPANN model and SVM model are 4.91, -5.09 and 3.39, respectively. The AVEs of MLR model, BPANN model and SVM model are 3.16, 2.49 and 1.48. The three data-driven models all perform well. The SVM model has the best ability of generalization in the sample learning process, because the SVs fluctuate within a narrow range around TVs and most of the REs are below 10%. The BPANN model performs better than MLR model as a whole except the test group1, and the SVs also fluctuate within a large range around TVs.

Table 2 Prediction results and errors of three models

Test groups' number	TV/m	MLR model			BPANN model			SVM model		
		SV/m	ABE/m	RE/%	SV/m	ABE/m	RE/%	SV/m	ABE/m	RE/%
1	26.91	23.99	-2.92	-10.85	21.82	-5.09	-18.9	29.22	2.31	8.58%
2	27.87	25.77	-2.1	-7.53	25.52	-2.35	-8.43	29.14	1.27	4.56%
3	28.32	25.48	-2.84	-10.03	26.89	-1.44	-5.07	29.13	0.81	2.86%
4	29.43	25.92	-3.51	-11.93	26.85	-2.57	-8.74	29.18	-0.25	-0.85%
5	29.74	24.91	-4.83	-16.24	30.14	0.4	1.33	30.23	0.49	1.65%
6	30.6	25.69	-4.91	-16.05	32.26	1.66	5.42	31.69	1.09	3.56%
7	31.37	27.21	-4.16	-13.26	34.57	3.2	10.2	32.87	1.5	4.78%
8	32.3	28.34	-3.96	-12.26	27.9	-4.4	-13.62	33.41	1.11	3.44%
9	32.52	31.18	-1.34	-4.12	31.29	-1.22	-3.76	35.13	2.61	8.03%
10	32.48	31.49	-0.99	-3.05	29.92	-2.56	-7.87	35.87	3.39	10.44%

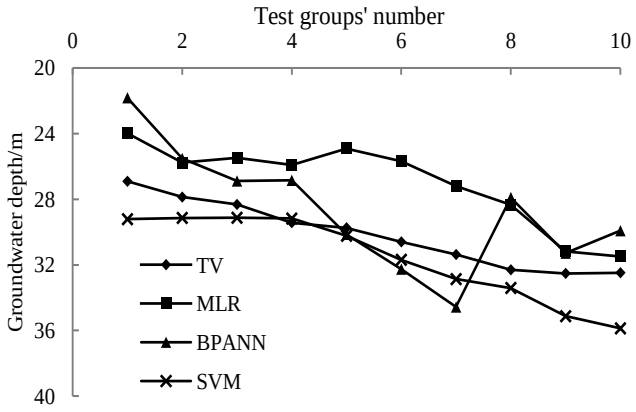

Fig. 3 Prediction results of three models

3.3 The influence of parameters for SVM

The parameters have a significant effect on the prediction results for the SVM model, especially the parameter C which is a positive trade-off parameter that determines the degree of the empirical error and the parameter γ which is the main parameter in kernel function of RBF. The optimal C and γ are 0.4 and 3.0, which are found by a lot of tests. The results are showed in figure 4 and figure 5. The minimum of the maximum RE and the minimum AVE are 10.44% and 1.43m in the tests. The RE and AVE changes significantly with the change of C values in the range of 0~1 and the change of γ values in range of 0~6.

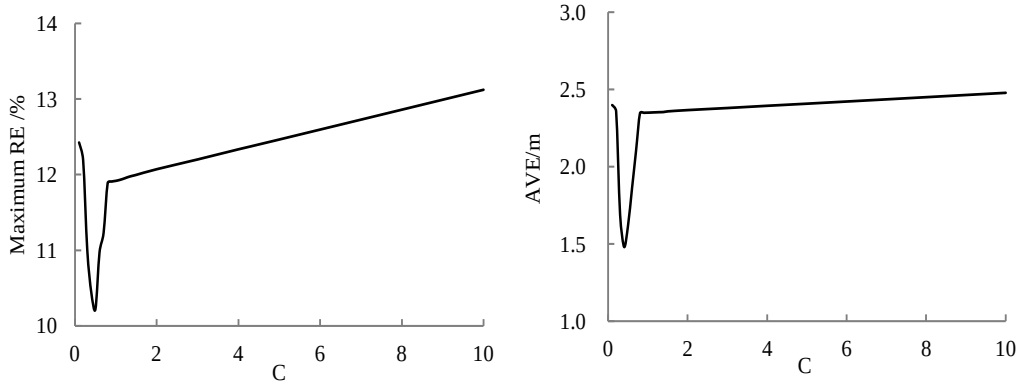

Fig. 4 Influence of parameter C on the results of prediction with γ=3

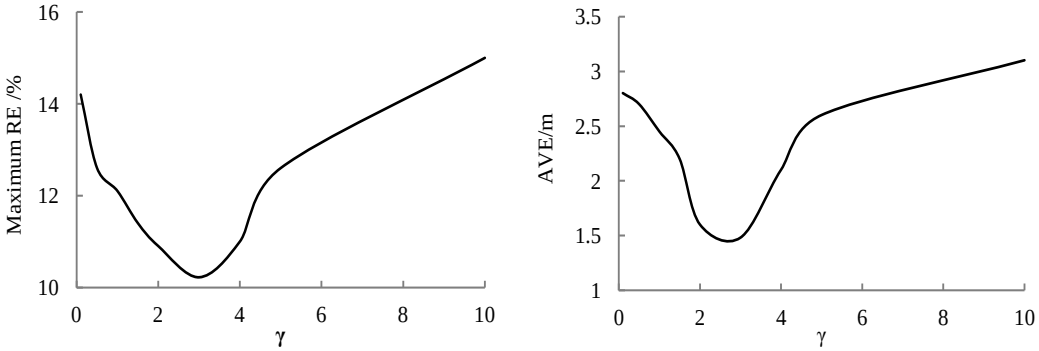

Fig. 5 Influence of parameter γ on the results of prediction with C=0.4

4. Conclusions

Three data-driven models are used to predict the groundwater depth in Shijiazhuang plain. Prediction, evaporation, groundwater exploitation, grain yield and GDP are chosen as the key factors which may influence the groundwater depth. These five factors involve natural, anthropic, biological and economic influence. The BPANN and SVM models perform well and can be used to forecast the groundwater depth in Shijiazhuang plain.

ABE, RE, ME and AVE are used to assess the accuracy of three methods. The prediction results of BPANN model and SVM model, which are nonlinear regression models, are superior to the linear regression model MLR model. Although the SVs of BPANN model fluctuate within a large range around TVs in some train and test groups, the accuracy is better than MLR model in general. The prediction accuracy and stability of SVM model are better than the other two models. The ME and AVE values of the SVM model are -1.86 and1.01 for fitting results, while the values are 3.39 and1.48 for prediction results. The maximum RE of SVM model is 10.44% and minimum RE is -0.85%.

The values of parameter C and γ have a significant effect on the prediction results for the SVM model. The optimal C and γ is 0.4 and 3.0, respectively, which are found by a lot of tests. The minimum of the maximum RE and the minimum AVE are 10.44% and 1.43m in the tests.

The results of this study can provide a reference for the forecasting of groundwater resources using data-driven models in the study area. However, monthly prediction of the groundwater depth has not been considered in this study, and the contributions of the five factors to the groundwater depth changes should be further studied in depth.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant No. 51409270 and 51209225), the International Science and Technology Cooperation Program of China (Grant No. 2013DFG70990), the foundation of China Institute of Water Resources and Hydropower Research (1232), and the Open Research Fund Program of State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering (2014490611).

References

Cao G, Han D, Song X. Evaluating actual evapotranspiration and impacts of groundwater storage change in the North China Plain[J]. Hydrological Processes, 2014, 28(4): 1797-1808.

Cao G, Zheng C, Scanlon B R, et al. Use of flow modeling to assess sustainability of groundwater resources in the North China Plain[J]. Water Resources Research, 2013, 49(1): 159-175.

Coppola Jr E, Szidarovszky F, Poulton M, et al. Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping, and climate conditions[J]. Journal of Hydrologic Engineering, 2003, 8(6): 348-360.

Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

Knotters M, Bierkens MFP. Physical basis of time series models for water table depths[J]. Water Resources Research, 2000, 36(1): 181-188.

Liu CM, Yu JJ, Kendy E. Groundwater exploitation and its impact on the environment in the North China Plain[J]. Water International, 2001, 26(2): 265-272.

Lin J Y, Cheng C T, Chau K W. Using support vector machines for long-term discharge prediction[J]. Hydrological Sciences Journal, 2006, 51(4): 599-612.

Liu J, Chang J, Zhang W. Groundwater level dynamic prediction based on chaos optimization and support vector machine[C]//Genetic and Evolutionary Computing, 2009. WGEC'09. 3rd International Conference on. IEEE, 2009: 39-43.

Nayak P C, Rao Y R S, Sudheer K P. Groundwater level forecasting in a shallow aquifer using artificial neural network approach[J]. Water Resources Management, 2006, 20(1): 77-90.

Ping J, Qiang Y, Xixia M. A combination model of chaos, wavelet and support vector machine predicting groundwater levels and its evaluation using three comprehensive quantifying techniques[J]. Information Technology Journal, 2013, 12(15): 3158.

Rumelhart D E, McClelland J L. The PDP Research Group: Parallel distributed processing: Explorations in the microstructure of cognition[J]. Foundations, 1986, 1.

Safavi H R, Esmikhani M. Conjunctive use of surface water and groundwater: application of support vector machines (SVMs) and genetic algorithms[J]. Water resources management, 2013, 27(7): 2623-2644.

Shiri J, Kisi O, Yoon H, et al. Predicting groundwater level fluctuations with meteorological effect implications—A comparative study among soft computing techniques[J]. Computers & Geosciences, 2013, 56: 32-44.

Shirmohammadi B, Vafakhah M, Moosavi V, et al. Application of several data-driven techniques for predicting groundwater level[J]. Water Resources Management, 2013, 27(2): 419-432.

Vapnik V N, Vapnik V. Statistical learning theory[M]. New York: Wiley, 1998.

Vapnik V. The nature of statistical learning theory[M]. Springer Science & Business Media, 2013.

Yuan Z, Shen Y. Estimation of agricultural water consumption from meteorological and yield data: a case study of Hebei, North China[J]. PloS one, 2013, 8(3): e58685.