Prediction of Groundwater Level for Sustainable Water Management in an Arid Basin Using Data-driven Models

Mutao Huang[1, *] and Yong Tian[2]
[1]College of Hydropower & Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[2]South University of Science and Technology of China, Shenzhen, 518055, China
*Corresponding author

***Abstract*—Arid and semi-arid regions face major challenges in the management of scarce freshwater resources under economic development and climate change. Groundwater is commonly the most important water resource in these areas. Accurate prediction of groundwater level is an essential component of suitable water resources management. Physically based model are often employed to perform groundwater simulation and predications. However, they are not applicable in many arid and semi-arid regions due to data limitations. Data-driven methods have proven their applicability in modeling complex and non-linear hydrological processes. The focus of this study is the application and comparison of three data-driven models for forecasting short-term groundwater levels. The purpose is to develop a new data-based method for highly accurate groundwater level forecasting that can be used to help water managers, engineers, and stake-holders manage groundwater in a more effective and sustainable manner. A set of popular data-driven models are evaluated and compared, including Artificial Neuron Networks (ANNs), Support Vector Machines (SVMs), and M5 Model Tree. The feasibility and capability of these models are demonstrated through a case study of forecasting five-days ahead groundwater level in an arid and semi-arid basin located in northwestern China. The encouraging simulation results show that the methodologies can simplify and improve the procedure of groundwater level forecast.**

Keywords- data-driven; groundwater level forecasting; ANN; SVM; model tree

I. INTRODUCTION

In many areas, groundwater is often one of the major sources of water supply for domestic, urban, agricultural and industrial purposes, especially in arid and semi-arid areas[1]. However, many problems occurs due to overexploitation of groundwater and unsustainable groundwater use and management, such as major water-level declines, drying up of wells, water-quality degradation, increased pumping costs, land surface subsidence, loss of pumpage in residential water supply wells, and aquifer compaction[2]. These problems are becoming a serious issue globally, especially in developing countries. To secure water for the future, the sustainable management of groundwater resources in conjunction with surface water has urgently become the need of the hour. Accurate and reliable prediction of groundwater levels is a crucial component for achieving this goal, especially in watersheds in arid and semi-arid regions that are more susceptible to hydrological extreme events in the form of droughts[3].

The data-driven models attempt to identify a direct mapping between the inputs and outputs of the system without reaching an understanding of the internal structure of the physical process. After enjoying much success in numerous hydrologic and water environment applications, such as rainfall-runoff modeling and water quality forecasting, data-driven models are now being applied more and more to solve problems in the area of groundwater[4]. Examples of the most common methods used in data-driven modeling of groundwater levels include: artificial neural networks (ANNs), support vector machines(SVMs), genetic programming (GP) and fuzzy rule-based systems like M5 model tree. Despite the growing applications and successes of data-driven approaches in the surface water problems, there have been only a few studies related to groundwater in arid and semi-arid regions. This provides an impetus for the current work. The focus of this study is the application and comparison of three data-driven models (i.e., ANN,SVM and M5 model tree) for forecasting short-term groundwater levels. The purpose is to develop a new data-based method of highly accurate groundwater level forecasting that can be used to help water managers, engineers, and stake-holders manage groundwater in a more effective and sustainable manner. In this research, the ability and accuracy of the three data driven models are investigated by applying them to forecast groundwater level in the Shule river basin situated in Gansu province, China.

II. METHODOLOGY

A. Modeling Techniques

ANNs are parallel architectures that comprise nonlinear processing nodes connected by synapses assigned with fixed or variable weights. The most popular ANN architecture used for regression or prediction is the Multi-Layer Perceptron (MLP) network. The MLP network has a layered architecture that is comprised of an input layer, followed by at least one hidden layer and an output layer. A layer typically consists of a number of neurons. Directed synapses connect each neuron in one layer to every neurons in the next layer. Each synapse is assigned with a weight. The "knowledge" about the data

behavior of a training set is stored in terms of the synapses' weights.

The concept of SVMs is initially introduced by Vapnik[5]. In the literature, support vector regression (SVR) was used to describe regression with SVM. The following briefly describes the rationale behind SVR methodology. Given the training set $\boldsymbol{G} = [(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_K, y_K)]$, where $\boldsymbol{x}_i$ is an input vector and y_i is its corresponding output data, SVR maps $\boldsymbol{x}_i$ into a high dimensional feature space via a nonlinear function $\phi(x)$,and then performs a linear regression in this feature space to find a function $f(x)$ that can best approximate the actual output y with an error tolerance ε, and at the same time, is as flat as possible.

The M5model tree algorithm was initially introduced by Quinlan and then the idea was reconstructed and improved[6]. This machine learning technique combines a conventional decision tree with the possibility of generating linear regression functions at the leaves. The main idea behind the M5 model tree is: split the parameter space into areas (subspaces) and build in each of them a linear regression model which is an "expert" for that subspace. The result is a hierarchical tree (often a binary tree) with splitting rules at non-terminal nodes and the expert models at the leaves.

B. *Performance Criteria*

In the modeling exercises, the predictive accuracy of the various models were evaluated using four numerical indicators that were recommended for evaluating hydrological forecasting model performance by many studies, i.e. the correlation coefficient (R), the root mean squared error (RMSE), and the Nash–Sutcliffe efficiency coefficient (NSE).

III. STUDY AREA AND DATA

A. *Study Area*

The area for this study is the Shule River basin, located in the arid inland region of northwest China (see Figure 1)where the range of the surface basin is outlined. The Shule River basin is one of the extremely dry regions in China with the characteristics of very arid continental climate, low precipitation, low runoff coefficient, sandstorm, and high evaporation capacity. The evaporation could be as much as 3042.6 mm annually, meanwhile, the annual precipitation is less than 70 mm. The annual amount of surface water resource is about 1.08 billion m^3. In last 15 years, large amounts of groundwater is being exploited. So far more than 2000 wells have been developed, and pumping is unregulated, resulting in overexploitation of the aquifer and the consequential decrease in groundwater levels over time. This has led to the increase of problems associated with exhaustion of the water supply for both agriculture and environmental purposes. Therefore, there is a strong need for predications of the future trends in the groundwater levels. Particularly, short-term groundwater level forecast is important for irrigation scheduling in this region, especially in the period when the water consumptive uses of plants are high and the surface water is inadequate.

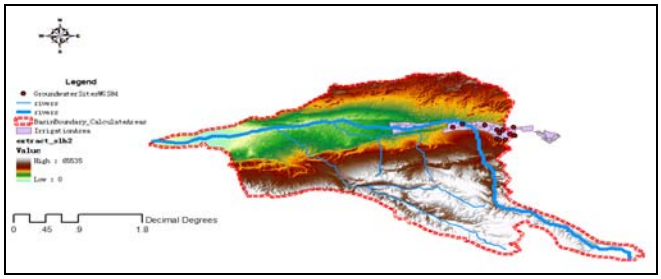


FIGURE I. STUDY AREA.

B. *Data*

According to past studies, groundwater level may be impacted by hydrological variables (e.g., surface stream-flow, precipitation, etc.) and meteorological variables (temperature, evaporation, etc.). However, observations data regarding these variables were limited for the study area. Only part of hydrological variables were available. More specially, the data used in this study consisted of daily river flow and average groundwater level of five days. The daily river flow was measured at a controlling hydrology station named Changma station. In the irrigation area there are about fifty perennial observations sites at which groundwater level were measured per five days. The observations of groundwater level from two representative sites were selected to illustrate the modeling methodology of this study: Huanghuaying and Shuguang. Both of the sites are located in the Changma irrigation system. All the data was obtained from Administration of Water Resources of Shule River Basin.

TABLE I. STATISTICAL PARAMETERS OF THE DATA SET OF DEPTH TO WATER TABLE AT EACH WELL.

Data Set	Well	Statistical parameter					
		Mean	***Max***	***Min***	S_d	C_v	C_{sx}
Training	Huang-huaying	5.20	5.45	4.94	0.113	0.022	-0.097
	Shu-guang	1.90	2.65	1.30	0.330	0.173	0.367
Testing	Huang-huaying	5.19	5.43	4.99	0.107	0.021	0.226
	Shu-guang	1.70	2.80	1.32	0.353	0.207	1.189

The available data was covered a period of 7 years from 2003 to 2009. For each site, the data of the first five years (2003-2007) were used for model calibration (or training), and the remaining two years of the data were used for model testing (or predication).The statistical parameters of the groundwater level data are given in Table 1, (The statistical parameters Mean, Max, Min, S_d, C_v and C_{sx} denote the mean, maximum, minimum, standard deviation, coefficient of variation and skewness, respectively).The data of Shuguang site show more scattered distribution than those of the Huanghuaying site (see the C_{sx} values in Table 1). It can be seen that the testing data set includes highly variable data with both low and high flows, which is important for the proper model testing.

IV. MODEL DEVELOPMENT

A. Determination of Model Structure

In order to select an appropriate set of inputs for the data-driven models, the cross-correlation method, which may be the most popular analytical technique for selecting appropriate inputs in hydrology, was employed to determine the input structure for both studied sites. The cross-correlation analysis was utilized to calculate the strength of the relationship between each potential input time series and the output time series at various lags (1 lag stands for 5 days).Followed the correlation analysis, the groundwater level of the Huanghuaying well and that of the Shuguang well will be predicated according to the following formulations:

For Huanghuaying Well:

$$z_h(t+\Delta t) = f_{non}(z_h(t+\Delta t - n_{zh})) \quad (1)$$

For Shuguang Well:

$$z_s(t+\Delta t) = f_{non}(q(t+\Delta t - n_q), z_s(t+\Delta t - n_{sh})) \quad (2)$$

where f_{non} is the unknown non-linear mapping function; t stands for the time index; Δt is the lead time; $z_h(t)$ and $z_s(t)$ are the groundwater levels to be forecasted at time t; $n_{zh} = 1,2,3,4,5,6$; $n_{sh} = 1,2,3,4,5,6$; $n_q = 1,2,3,4,5$; $q(t-n_q)$ is the past observed river runoff at time $t-n_q$; $e(t)$ is the unknown mapping error to be minimized.

In this study, multiple lead-time predictions of 2 and 3 time steps (one step stands for 5 days) were also carried out for each study well. The available observations used as inputs to equation (1-2), were used to predict directly 1 step-ahead groundwater levels.

B. Data Normalization

The groundwater level and the runoff have different units and their values do not represent the same quantities, so normalization of data within a uniform range is essential. An additional reason for normalization is that normalization can prevent the data-driven models from being dominated by the variables with large values and thus improve the model performance. There is no one standard procedure for normalizing inputs and outputs. One way is to scale input and output variables (x_i) in interval $[\lambda_1, \lambda_2]$ corresponding to the range of the transfer function:

$$x_i^{'} = \lambda_1 + (\lambda_2 - \lambda_1)\left(\frac{x_i - x_i^{\min}}{x_i^{\max} - x_i^{\min}}\right) \quad (3)$$

where $x_i^{'}$ is normalized value of x_i ; $x_i^{\min}$ and $x_i^{\max}$ are the minimum and maximum values of x_i . In this study, the interval $[\lambda_1, \lambda_2]$ was set to [0.1,0.9], thus all the input and output data was normalized to the range from 0.1 to 0.9. The forecasting models were fed normalized data, then the final model output results were returned to their original scale through reverse calculations according to equation (1-2).

C. Model Implementation

The algorithm of ANN model was implemented using C# language. Two ANN models trained with BP algorithm were developed for the two study wells. As concerns a MLP network, the number of neurons in input and output layer was determined by the input and output vectors presenting to the network, thus the problem of selecting a suitable architecture for a MLP network relies on specifying the type of activation function to be used and the number of neurons in the hidden layer. When developing the ANN models, the sigmoid activation function was used for both the hidden and output neurons. The number of neurons in the hidden layer was optimized through a trial-and-error procedure.

The revised codes of LIBSVM written in C# was used to implement SVM forecasting models. As demonstrated by many authors, the selection of parameters greatly affects the performance of SVM model. The parameters dominating SVM model include the cost constant C, the radius of the insensitive tube ε, and kernel functions. In this study, the type of kernel function was selected at first. Much work on the use of SVM regression model has demonstrated that RBF kernel outperformed other kernel functions. After preliminarily testing SVM models with linear, polynomial, and RBF kernels on all training datasets, RBF was found to provide the best predictive performance.

A Matlab toolbox named M5 PrimeLab developed by Jekabsons was utilized to implement the M5 model tree model. The model has three parameters to be specified: the minimum number of training data instances one node may represent N_{ins}, the smoothing coefficient for the smoothing process and a threshold values. The parameter *k* is a kind of smoothing coefficient for the smoothing process, more smoothing is applied for largerk values, making the tree behave like containing just one leaf corresponding to the root node, while no smoothing is applied for value 0. The similar experimental steps adopted with the ANN techniques were carried out to seek optimal model parameters.

V. RESULTS AND ANALYSIS

In order to evaluate generalization (or forecasting) ability of the trained models, the three models were applied to predicate groundwater five days ahead for the testing period 2008-2009. Table 2 presents comparison of the forecasting performance of the three models. The overall performance of the models are found acceptable based on the high correlation efficiency. From the table, it's obvious to see that the SVM model shows inferior results compared with the other two models in the

training period, but the M5 model holds the best performance in the testing period.

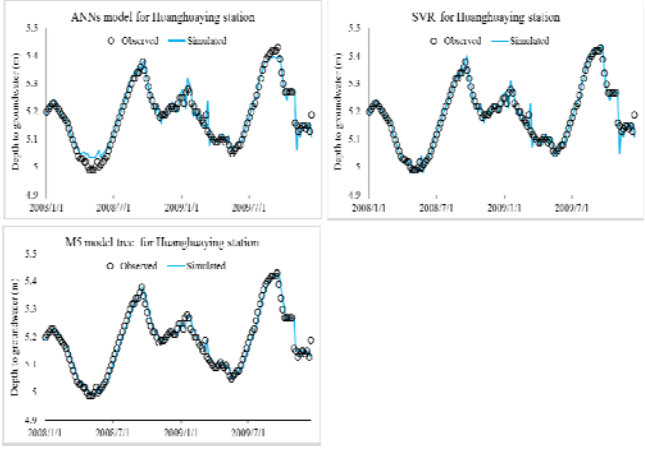
The forecasted groundwater levels produced by the three models versus the measured values at the two stations in the testing period are presented graphically using curves as shown in Figure 2. The following can be summarized from these two figures: it can be seen from Figure 2 that there is relatively good agreements between the simulated and observed groundwater level for all the three models. Thus, it is practically possible to develop groundwater forecasting models using these data-driven approaches. However, there are discrepancies in matching some of the peak events, where the events may be under predicted or over predicted values. From the overall exercises above, the possibility of using the ANN, SVR, M5 data-driven components for groundwater flow modeling was demonstrated. The fact that the three data-driven components ran independently and sufficiently good forecasting results were made prove that the system can be applied to the real world scenarios.

TABLE II. COMPARISON OF ANN, SVR AND M5 PERFORMANCES FOR THE TRAINING AND TESTING PERIOD AT HUANGHUAYING STATION.

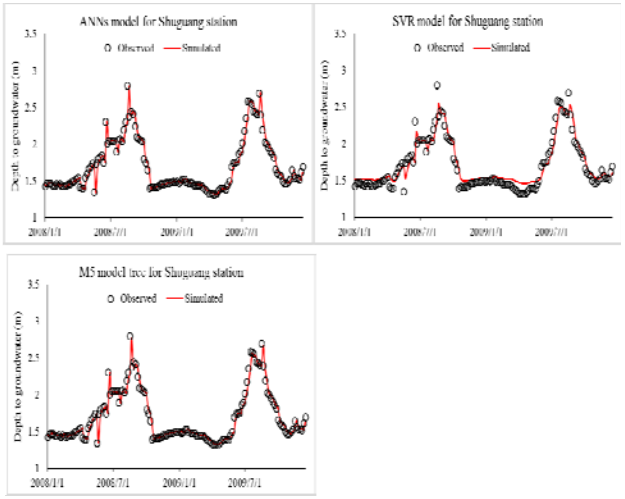
Model	Training			Testing		
	R	*RMSE* (m)	*NSE*	*R*	*RMSE* (m)	*NSE*
ANN	0.989	0.017	0.978	0.972	0.026	0.943
SVR	0.993	0.014	0.986	0.979	0.023	0.955
M5	0.987	0.018	0.974	0.981	0.021	0.962

TABLE III. COMPARISON OF ANN, SVR AND M5 PERFORMANCES FOR THE TRAINING AND TESTING PERIOD AT SHUGUANG STATION.

Model	Training			Testing		
	R	*RMSE* (m)	*NSE*	*R*	*RMSE* (m)	*NSE*
ANN	0.947	0.106	0.896	0.942	0.123	0.878
SVR	0.948	0.108	0.897	0.943	0.118	0.887
M5	0.948	0.106	0.896	0.943	0.119	0.886



(a) Huanghuaying station



(b) Shuguang station

FIGURE II. SIMULATED AND OBSERVED GROUNDWATER LEVEL DURING THE TESTING PERIOD USING ANN, SVR AND M5 AT HUANGHUAYING AND SHUGUANG STATIONS.

ACKNOWLEDGMENT

This research was financially supported by the National Science Foundation of China (NSFC) (No. 51579108 and No. 51209098) and the Fundamental Research Funds for the Central Universities (Program No. 2014TS153).

REFERENCES

[1] Adam, J. C., et al. (2009), Implications of global climate change for snowmelt hydrology in the twenty-first century, Hydrol Process, 23(7), 962-972.

[2] Banerjee, P., et al. (2009), Forecasting of groundwater level in hard rock region using artificial neural network, Environmental Geology, 58(6), 1239-1246.

[3] Yoon, H., et al. (2011), A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer, Journal of Hydrology, 396(1-2), 128-138.

[4] Solomatine, D. P., and A. Ostfeld (2008), Data-driven modelling: some past experiences and new approaches, J Hydroinform, 10(1), 3-22.

[5] Vapnik, V. N. (1995), The nature of statistical learning theory, 188 pp., Springer-Verlag New York, Inc.

[6] Quinlan, J. R. (1992), Learning With Continuous Classes, in 5th Australian Joint Conference on Artificial Intelligence, edited by Adams and Sterling, pp. 343–348, World Scientific, Singapore.