

Aquifer Water Level Prediction Using Support Vector Machines Method

Mohsen Behzad¹, Keyvan Asghari²

¹Graduate Student

²Assistant Professor

Isfahan University of Technology, Civil Engineering Department

E-mail: m_behzad@cv.iut.ac.ir

Abstract

In this research, a new data-driven model called Support Vector Machine (SVMs) uses the initial water level measurements, production well extractions, and climate conditions to forecast the final water level elevation in multi-time scale (i.e. daily, weekly, bi-weekly, monthly and bi-monthly) at a specific monitoring well. Due to the fact that SVMs approach does not require the explicit characterization of the physical conditions and input parameters, simulation is made based on the easily quantifiable and measurable variables. This study will demonstrate the prediction capability of SVMs compared to that of ANNs in forecasting the aquifer Water Level Elevation (WLE).

Keywords: Support vector machines, ANNs, Aquifer water level elevation, Climate conditions.

Introduction

Physical-based numerical flow models of groundwater systems have been used for simulation and analysis of groundwater flows. They have been applied to problems ranging from aquifer safe yield analysis to groundwater environmental issues. Regarding the diversity of available user-friendly groundwater flow models, the most challenging part of the flow modeling is the parameter estimation and definition of boundary conditions for models [1]. Since most of the flow models need to be discretized in space and time domain, parameters involving aquifer characteristics and boundary conditions would be required to simulate.

For large-scale management problems for which accurate simulation of localized behavior is not essential, numerical models are extensively used. There are some situations that more precise simulation is needed, therefore, we have to avoid simplifying physical and mathematical assumptions of numerical models. Moreover, field data are typically not accessible for applying in numerical modeling of localized problems [1, 2]. Thus, the problem of inherent complexity and data uncertainty of groundwater systems mostly limits physical-based modeling simulation accuracy.

In this article, we investigate a state of the art modeling tool namely Support Vector Machine (SVM) as an alternative approach to physical-based model, by applying it to predict the Water Level Elevation (WLE) of a specific monitoring well. SVMs are developed based on the Statistical Learning Theory (SLT). The most advantageous aspect of these learning machines is employing of the Structural Risk Minimization (SRM) Principle from which SVMs are able to generalize well to unseen data. According to this principle, SVM has two outstanding features which lead to being a promising prediction method. The first one is its excellent generalization, and the second one is sparse representation. Moreover, for implementing SVM, a convex quadratic constrained optimization problem must be solved; hence, the solution is always unique and globally optimal. Owing to this principle, SVMs also seem to be powerful surrogates of Artificial Neural Networks (ANNs) by eliminating some of the basic weaknesses associated with ANNs modeling while retain all strengths of ANNs [3]. Hence in this study, the results of an ANN model used to examine the performance of SVM method.

Support Vector Machines

Support Vector Machines (SVMs), in the regression and classification forms, were introduced by Vapnik (1998) and his colleagues as a robust and significant learning tool [4]. Since then, there have been a growing number of researches on the SVMs applications. Recently, SVMs have been used in water resources and hydrological areas as a novel approach of learning. Liong and Sivapragasam (2002) successfully employed SVM in the flood stage

forecasting [5]. Asefa *et al.* (2004, 2005b) applied SVMs for various water resources modeling including optimal design of the groundwater monitoring networks for both the head observation and contamination detection networks [6, 7]. They also examined them for snow-runoff modeling and learning of chaotic time series [8, 9]. Khalil *et al.* (2005a, 2005b) exploited this learning method for a variety of applications comprise environmental, such as estimating the groundwater nitrate contamination level, and hydrological cases, as involving the real-time operation of a reservoir [10, 11]. Lin *et al.* (2006) used SVM for long-term discharge prediction and compared the performance of SVM with two alternative methods, ANN and ARMA models [12]. The latest SVMs applications in the hydrological modeling field involve the temporal prediction of rainfall. These articles emphasized the optimal selection of the SVMs parameters by applying stochastic search methods such as Genetic Algorithm and Simulated Annealing Algorithm [13, 14]. It should be mentioned that the use of ANNs in complex groundwater modeling and management problems specifically prediction of water level elevation has been investigated by Coppola *et al.* within several articles [1, 2, and 15].

In all of the aforementioned applications, SVMs presented overall superior performance when compared with other data-driven models such as ANNs. This encouraging performance is because of high generalization property of SVMs and it motivates the researcher to work on the further applications.

In the regression form, SVMs have been employed to find an estimation function $f(\mathbf{x})$ in order to estimate the real function $y(\mathbf{x})$ (for simplicity in deriving the estimation function, we just consider linear dependency between \mathbf{x} and y) with minimum error and only based on the following independent and identically distributed data;

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \subseteq (\mathbf{X} \subseteq \mathbf{R}^n \times Y \subseteq \mathbf{R}) \quad (1)$$

For achieving this goal, SVMs applied Structural Risk Minimization Principle. It declares that for finding the most proper estimation function with high generalization performance, simultaneous control of both capacity of function f and empirical error, between the real and estimation functions, have to be minimized. This principle leads to following convex constrained quadratic optimization problem [16]:

$$\begin{aligned} \text{minimize} \quad & \tau(\mathbf{w}, \xi, \xi^*) = \frac{1}{2} \|\mathbf{w}\|^2 + C \cdot \frac{1}{m} \sum_{i=1}^m (\xi_i + \xi_i^*), \\ \text{subject to} \quad & (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - y_i \leq \varepsilon + \xi_i, \\ & y_i - (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq \varepsilon + \xi_i^*, \\ & \xi, \xi_i^* \geq 0. \quad i = 1, \dots, m. \end{aligned} \quad (2)$$

here the \mathbf{w} are the support vectors weights and the angle bracket denote the dot product. The b is bias value and the ε indicates the level of accuracy to which the errors are bearable and ξ, ξ^* represent the slack variables determining the penalty value for errors more than ε (Fig. 1). These constrains imply the ε -insensitive loss function which is employed for calculating the empirical error. According to this function errors being smaller than ε value are regarded as noises and they have to vanish. This function is presented in the following form;

$$|y - f(\mathbf{x})|_\varepsilon = \begin{cases} 0 & \text{if } |y - f(\mathbf{x})| \leq \varepsilon \\ |y - f(\mathbf{x})| - \varepsilon & \text{otherwise} \end{cases} \quad (3)$$

Solving the optimization problem results in the estimation function;

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b. \quad (4)$$

The obtained estimation function has minimum complexity (i.e. the flattest estimation function within the linear hypothesis space) along with minimum risk according to the value of ε . In other words, only the data points outside the ε boundary (Fig. 1) have to be penalized by the amount of slack variables. The constant C determines the compromise between the complexity of the function f and the amount over which the errors more than ε are penalized. Usually, the optimization problem (2) is transformed to the Lagrangian form in which the problem is more convenient to solve and also it has some advantageous especially in dealing with nonlinear functions. By using this representation the following optimization problem will be obtained [16]:

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2} \sum_{i,j=1}^m (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \varepsilon \sum_{i=1}^m (\alpha_i^* + \alpha_i) + \sum_{i=1}^m y_i (\alpha_i^* - \alpha_i) \\ \text{subject to} \quad & \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0, \quad \alpha_i, \alpha_i^* \in [0, \frac{C}{m}]. \end{aligned} \quad (5)$$

When solving the new representation of objective function called dual representation the weight vector and consequently the estimation function derived as follows;

$$\mathbf{w} = \sum_{i=1}^m (\alpha_i^* - \alpha_i) \mathbf{x}_i, \text{ and } f(\mathbf{x}) = \sum_{i=1}^m (\alpha_i^* - \alpha_i) \langle \mathbf{x}_i, \mathbf{x} \rangle + b. \quad (6)$$

For estimating a non-linear dependency between input vector and output scalar, the input space (\mathbf{X}) is mapped to the feature space (\mathbf{F}) by a mapping function (i.e. $\phi(\mathbf{x}): \mathbf{X} \rightarrow \mathbf{F}$). Thus, by resolving the above-mentioned optimization problem for deriving estimation function the following expression will be accomplished [17];

$$f(\mathbf{x}) = \sum_{i=1}^m (\alpha_i^* - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) + b. \quad (7)$$

where the kernel function $K(\mathbf{x}_i, \mathbf{x})$ has the following description;

$$K(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle \text{ for all } \mathbf{x}, \mathbf{z} \in \mathbf{X} \quad (8)$$

this is the kernel trick in which one does not require knowing the mapping function explicitly.

It should be noted that α and α^* (Lagrangian coefficients) are derived from optimization problem (5). Further, these coefficients are zero for the data points that lie inside the ε boundary, but the data pairs with absolute error larger than ε have non-zero coefficients and they only involve the estimation function for future prediction. This is the sparseness property of SVMs which has significant computational implication. The input vectors corresponding to non-zero coefficients named Support Vectors (SVs) and hence the name Support Vector Machines [17].

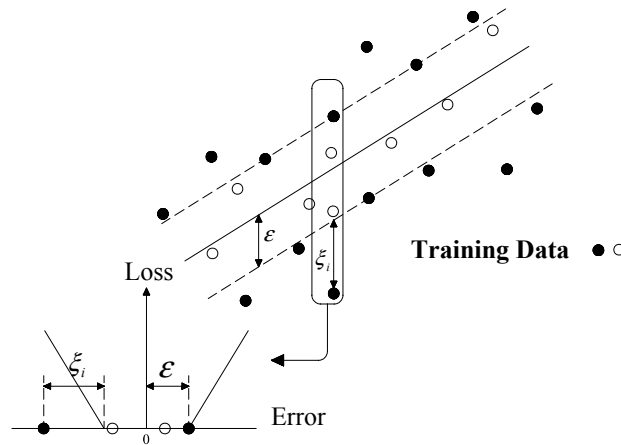


Figure 1- ε -insensitive Loss Function

Study Area

The Towaco aquifer locates in the northern part of Morris County, New Jersey, US. The Towaco Valley contains a small, buried glacial valley cut into bedrock and filled with unconsolidated glacial and lacustrine deposits of Quaternary age. The glacial deposits encompass clay and silt mixed with sand and boulders, lacustrine fine-grained silty sand, and outwash coarse-grained sand and gravel aquifer materials. These unconsolidated glacial deposits comprise the highly prolific Towaco Valley aquifer, which is the Township's only public drinking water source by extracting from three productions wells. In the vicinity of these wells, two monitoring wells (Cooks Lane and Indian Lane monitoring wells) installed to measure the rising and falling of the water level elevation (WLE) of the Towaco aquifer in order to supervise and control the aquifer behavior. Hydrogeologic investigations (Fig. 2) reveal an unconfined aquifer encompassing coarse materials leading to high permeability. Directly beneath this aquifer there is a discontinuous semi-confining layer which has silt and clay and therefore lower permeability [1]. This layer overlies the highly transmissive semi-confined Towaco aquifer. The foregoing layers make the entire system very impractical to simulate with traditional modeling tools.

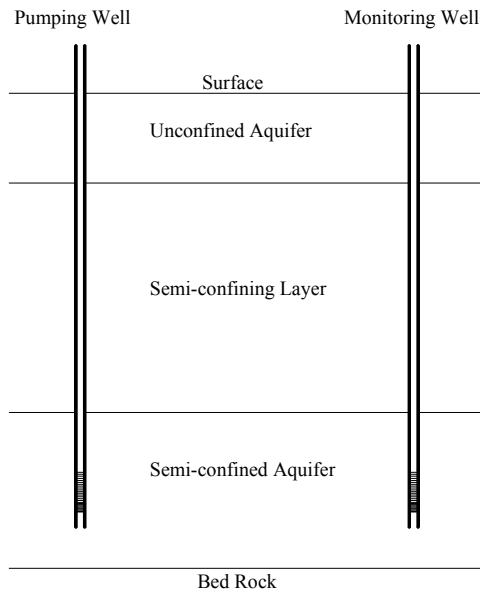


Figure 2- Hydrogeologic cross-section of Towaco Valley

Methodology

In this research, the prospective influences of the pumping extraction and climate changes on the Water Level Elevation (WLE) of Cooks monitoring well, located in the semi-confined aquifer were assessed by SVMs and ANNs method.

For selecting the appropriate input variables, we should understand the underlying physical and other processes involving the hydrological behavior of groundwater system. For groundwater flow, the governing principle is the conservation of mass which is the basis of the physical equations of groundwater flow. Regarding this law, the net gains and losses in aquifer storage lead to changes of the potentiometric surface of the aquifer. Typical sources of water entering the aquifer include areal recharge, losing surface flow and other aquifers leakage. There are also some terms referring to water leaving the aquifer (i.e. sinks) consisting of evapotranspiration, pumping extraction, and gaining surface water. Since quantification of most of the above-mentioned sources and sinks are difficult, we use some directly measured variables as surrogates of those parameters in order to use them in developing the ANN and SVM models. For instance, we can apply precipitation and temperature variables instead of areal recharge, a problematic source in modeling of the groundwater flow. Coppola *et al.* (2005) had a comprehensive investigation on the use of some simple variables as substitution of those variables which are considered in groundwater flow modeling but are difficult to quantify ones in space and time domain [1].

According to studies done by Coppola *et al.* (2005) and the aforementioned reasons, we exploited the following attributes for input vector whose output is the final water level elevation at the end of the n days period in the Cooks monitoring well, to develop two data-driven models known as ANN and SVM. We then compare the results of the two methods to evaluate models performance. The following input parameters (attributes) were utilized in order to predict the final water level elevation in corresponding monitoring well.

1. Mean daily pumping rate of production well 1 over n days period.
2. Mean daily pumping rate of production well 2 over n days period.
3. Mean daily pumping rate of production well 3 over n days period.
4. Cumulative mean pumping rates of three production wells over n days period.
5. Total precipitation over n days period.
6. Mean daily temperature over n days period.
7. Initial water level measurement of monitoring well at the beginning of n days period.

In current study, the value of n includes 1, 7, 14, 30, and 60 days which imply daily, weekly, bi-weekly, monthly and bi-monthly prediction. Thus, we could investigate the different time-scale, from short-term to medium and long-term, for prediction of WLE in a more comprehensive manner. We applied the data from spring 2002 to mid-fall 2002 which consist of 122, 116, 109, 93 and 63 data pairs for daily, weekly, bi-weekly, monthly and bi-monthly time period, respectively.

For implementation of the SVM we must initially recognize appropriate values of optimal hyper parameters in advance of building the model. The optimal values, obtained by 10-fold cross validation, are presented in table

1. The Radial Basis Function was selected for proper kernel function. This function has been proved to be more efficient than other kernel functions according to many reported studies on the use of RBF [7, 8, 9, 18 and 19]. Also the data transformed over the range of [-1 1] by a linear transformation to be appropriate for learning machines. The SVM modeling was carried out by using the LIBSVM software, developed by Chang and Lin, (2001) [20].

For ANN modeling, a Multi-Layer Feed-forward (MLF) neural networks cooperating with the Levenberg-Marquardt technique of the back-propagation learning algorithm, with one hidden layer and sigmoid transfer function was used.

Table 1- The Optimal Value of the SVM Hyperparameters

	c	γ	ϵ
Daily	17	0.0114	0.0474
Weekly	12	0.1	0.03
Bi-weekly	0.5	0.3	0.001
Monthly	16.5	0.5	0.001
Bi-monthly	41	0.2	0.001

Results and Discussion

In the proposed application, SVM was able to simulate the variability of water level in a specific observation well, by using the pumping rate and climate conditions. We considered five different periods of daily, weekly, bi-weekly, monthly and bi-monthly in order to evaluate the capability of learning machine for prediction of WLE in distinct time periods. The input vector consists of the pumping rate from production wells, initial water level in observation well and local climate conditions. The water level at the end of time period is the output value.

A complex, non-homogeneous and non-isotropic aquifer was selected for this application. It should be noted that the numerical models of groundwater flow systems have to consider not only the hydrological characteristics and boundary conditions of the semi-confined aquifer but also the characteristic of the semi-confining layer, unconfined aquifer and bed rock formation too. Also there are some difficulties in characterizing the temporal and spatial variables involved in mathematical modeling such as ET rate and interaction of surface and underground flow.

Since data-driven models have capability of modeling complex hydrogeological systems, they could be modeling tools of great interest. In this research, we applied both SVM and ANN for predicting water level and compared the results with those of ANN.

Table 2 presents evaluation of models on the test data set, based on the three different performance criteria, namely root mean squared error (*RMSE*), coefficient of determination (R^2) and coefficient of efficiency (E_1). The latter one has the following definition:

$$E_1 = 1.0 - \frac{\sum_{i=1}^N |O_i - P_i|}{\sum_{i=1}^N |O_i - \bar{O}|} \quad (9)$$

where O and P indicate the observed and predicted values, \bar{O} is the mean of the observed values and N is the number of data pairs.

Through comparison of the ANN and SVM results, it is concluded that by increasing the length of the time period the SVM gradually demonstrates its predominance over the ANN until the monthly prediction which has more tangible improvement in terms of accuracy. This conclusion implies that SVM method enhances the long-term prediction quality relative to that of ANN; while maintaining its suitability to model the short-term forecasting. This might be because of the ability of SVM to learn from scarce data in the context of the SRM principle. For long-term prediction, the data set shrinks owing to the longer time periods. This was the case of bi-monthly forecasting in which only 34 out of 63 data pairs encompasses the training data set. Consequently, SVM is more applicable and reliable than ANN, particularly in situation where limited data pairs are available.

As it was mentioned in the preceding section, the most advantageous property of the SVM is the high generalization performance which is derived from structural risk minimization principle. Table 3 reveals this characteristic by specifying the error for both training and test data sets. By examining the *MSE* values as error criterion, it could recognize that error growing from training to test set is not identical for both methods. In other words, in the SVM model the discrepancy between the test and training *MSE* is negligible in comparison with that of neural networks. This result indicates the outstanding generalization feature of the proposed modeling method. Although the *MSE* of ANN in the training is smaller than that of SVM, the improved efficiency of SVM

on the test set designates the overtraining phenomena occurring during the training of the ANN; in spite of cross-validation phase which was also exploited in the training stage.

Figures 3a to 3e show the various time period prediction of the WLE (or Depth to Water) in the monitoring well versus the observed values. It is obvious that both SVM and ANN simulate the fluctuation of the water level precisely. Note that the days are not necessarily consecutive in time.

Table 2- Evaluation of the Models on the Test Set

	SVM			ANN		
	<i>RMSE, m</i>	R^2	E_j	<i>RMSE, m</i>	R^2	E_j
Daily	0.186	0.9554	0.8526	0.177	0.9626	0.8391
Weekly	0.119	0.9783	0.8674	0.130	0.9722	0.8370
Bi-weekly	0.153	0.9577	0.8054	0.179	0.9588	0.7610
Monthly	0.134	0.9235	0.7343	0.158	0.8970	0.6893
Bi-monthly	0.219	0.6738	0.2535	0.315	0.3887	0.0227

Table 3- Investigating the Generalization Performance of the Models (Transformed MSE)

	SVM		ANN	
	Training	Test	Training	Test
Daily	0.014158	0.014606	0.002136	0.012614
Weekly	0.007988	0.006170	0.001413	0.007423
Bi-weekly	0.014875	0.010212	0.000646	0.015096
Monthly	0.008107	0.013727	0.003402	0.019025
Bi-monthly	0.061302	0.093269	0.044247	0.192154

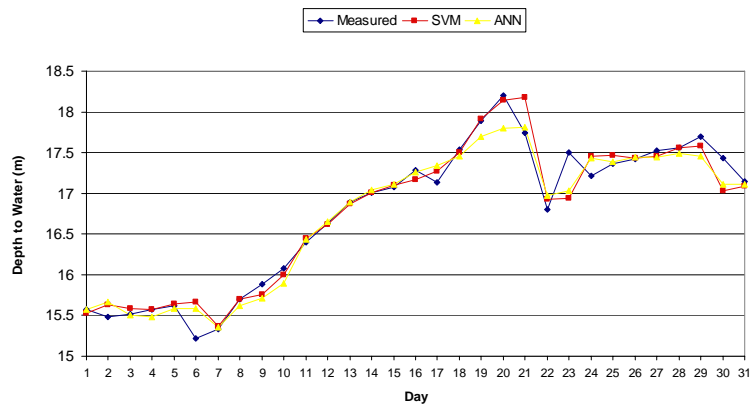


Figure 3a- Daily Prediction of WLE on the Test Set

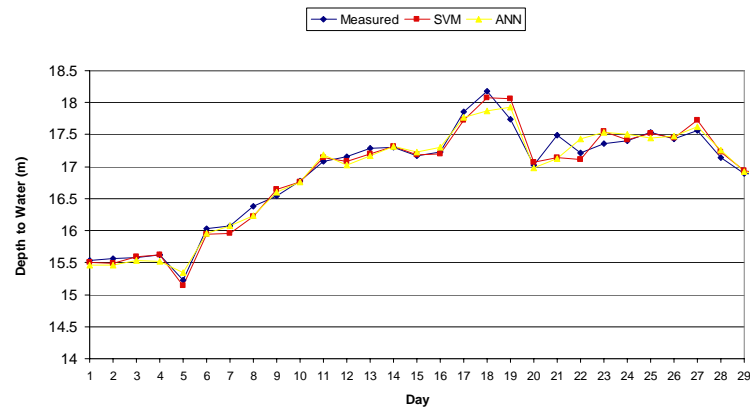


Figure 3b- Weekly Prediction of WLE on the Test Set

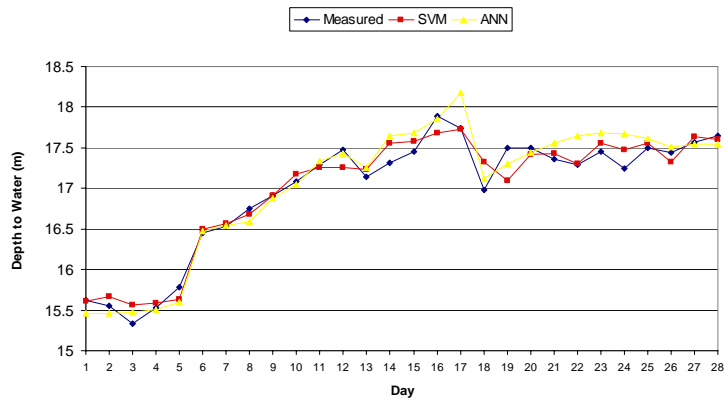


Figure 3c- Bi-weekly Prediction of WLE on the Test Set

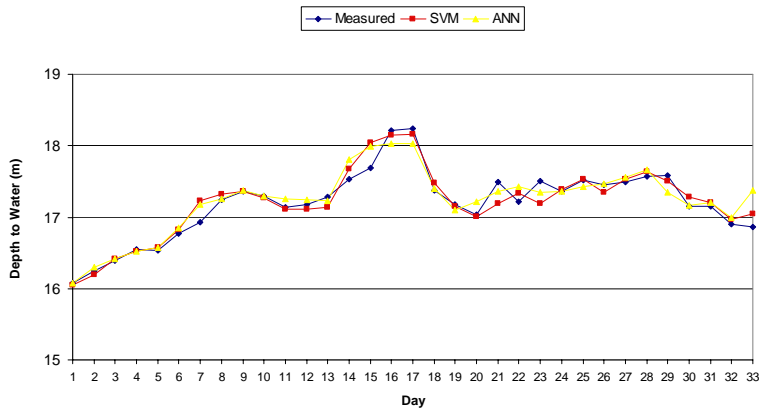


Figure 3d- Monthly Prediction of WLE on the Test Set

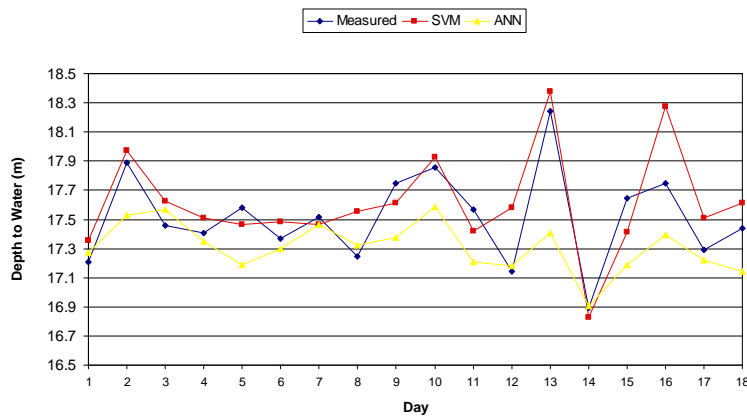


Figure 3e- Bi-monthly Prediction of WLE on the Test Set

References

1. Coppola, E., Rana, A., Poulton, M., Szidarovszky, F. and Uhl, V. (2005) A Neural Network Model for Predicting Aquifer Water Level Elevations. *Ground Water*, **43**(2), 231–241.
2. Coppola, E., Poulton, M., Charles, E., Dustman, J. and Szidarovszky, F. (2003a) Application of Artificial Neural Networks to Complex Groundwater Management Problems. *Natural Resources Research*, **12**(4), 303–320.
3. ASCE Task Committee on Application of the Artificial Neural Networks in Hydrology. (2000a) Artificial Neural Networks in Hydrology, I: Preliminary Concepts. *Journal of Hydrological Engineering* **5**(2): 115-123.
4. Vapnik, V.N. (1998) *Statistical Learning Theory*. John Wiley, New York.
5. Liong, S-Y and Sivapragasam, C. (2002) Flood stage forecasting with support vector machines. *Journal of American Water Resources Association*, **38**(1), 173-186.
6. Asefa, T., Kemblowski, M.W., Urroz, G., McKee, M. and Khalil, A. (2004) Support vector-based ground water head observation networks design. *Water Resources Research*, **40**, W11509, DOI: 10.1029/2004WR003304.
7. Asefa, T., Kemblowski, M.W., Urroz, G. and McKee, M. (2005a) Support vector machines (SVMs) for monitoring networks design. *Ground Water*, **43**(3), 413-422.
8. Asefa, T., Kemblowski, M.W., McKee, M. and Khalil, A. (2006) Multi-time scale stream flow prediction: The support vector machines approach. *Journal of Hydrology*, **318**, 7-16.
9. Asefa, T., Kemblowski, M.W., Lall, U. and Urroz, G. (2005b) Support vector machines for nonlinear state space reconstruction: Application to Great Salt Lake time series. *Water Resources Research*, **41**, W12422, DOI: 10.1029/2004WR003785.
10. Khalil, A., Almasri, M.N., McKee, M. and Kaluarachchi, J.J. (2005a) Applicability of statistical learning algorithms in groundwater quality modeling. *Water Resources Research*, **41**, W05010, DOI: 10.1029/2004WR003608.
11. Khalil, A., McKee, M., Kemblowski, M.W. and Asefa, T. (2005b) Sparse Bayesian learning machine for real-time management of reservoir releases. *Water Resources Research*, **41**, W11401, DOI: 10.1029/2004WR003891.
12. Lin, J-Y, Cheng, C-T and Chau, K-W. (2006) Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal*, **51**(4), 599- 612.
13. Pai, P-F and Hong, W-C. (2007) A recurrent support vector regression model in rainfall forecasting. *Hydrological Processes*, **21**, 819-827.
14. Hong, W-C and Pai, P-F. (2007) Potential assessment of support vector regression technique in rainfall forecasting. *Water Resource Management*, **21**, 495-513, DOI: 10.1007/s11269-006-9026-2.
15. Coppola, E., Szidarovszky, F., Poulton, M. and Charles, E. (2003b) Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping, and climate conditions. *Hydrologic Engineering*, **8**(6), 348–359.
16. Schölkopf, B. and Smola, A.J. (2002) *Learning with Kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, Mass.
17. Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines*. Cambridge University Press, New York.
18. Dibike, Y.B., Velickov, S., Solomatine, D. and Abbott, M.B. (2001) Model Induction with support vector machines: Introduction and Application. *Journal of Computing in Civil Engineering*, **15**(3), 208-216.
19. Han, D. and Cluckie, I. (2004) Support vector machines identification for runoff modeling. In: Liong, S.Y., Phoon, K.K., Babovic, V. (Eds.), *Proceedings of the Sixth International Conference on Hydroinformatics*. 21–24 June, Singapore.
20. Chang, C-C and Lin, C-J. (2001) LIBSVM: a Library for Support Vector Machines (Version 2.82, April 2006). Software available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.