Journal of Hydrology 396 (2011) 128–138

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer

Heesung Yoon [a], Seong-Chun Jun [b], Yunjung Hyun [a], Gwang-Ok Bae [a], Kang-Kun Lee [a,*]

[a] School of Earth and Environmental Sciences, Seoul National University, Seoul 151-747, Republic of Korea
[b] GeoGreen21 Co., Ltd., EnC Venture Dream Tower 2nd 901, Seoul 152-719, Republic of Korea

ARTICLE INFO

Article history:
Received 20 August 2009
Received in revised form 9 April 2010
Accepted 2 November 2010

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Luis E. Samaniego, Associate Editor

Keywords:
Groundwater level
Coastal aquifer
Artificial neural network
Support vector machine

SUMMARY

We have developed two nonlinear time-series models for predicting groundwater level (GWL) fluctuations using artificial neural networks (ANNs) and support vector machines (SVMs). The models were applied to GWL prediction of two wells at a coastal aquifer in Korea. Among the possible variables (past GWL, precipitation, and tide level) for an input structure, the past GWL was the most effective input variable for this study site. Tide level was more frequently selected as an input variable than precipitation. The results of the model performance show that root mean squared error (RMSE) values of ANN models are lower than those of SVM in model training and testing stages. However, the overall model performance criteria of the SVM are similar to or even better than those of the ANN in model prediction stage. The generalization ability of a SVM model is superior to an ANN model for input structures and lead times. The uncertainty analysis for model parameters detects an equifinality of model parameter sets and higher uncertainty for ANN model than SVM in this case. These results imply that the model-building process should be carefully conducted, especially when using ANN models for GWL forecasting in a coastal aquifer.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

In many countries, groundwater is a significant water resource for domestic, industrial, and agricultural activities. For the effective management of groundwater, it is important to predict groundwater level (GWL) fluctuations. Thus, there have been many researches on developing GWL prediction models.

Typically, physically based numerical models are used for characterizing a groundwater flow system and predicting the GWL fluctuation. These models establish a governing equation simplifying the physics of flow in the subsurface and solve it with proper initial and boundary conditions using numerical methods. Spatial and temporal GWL distributions can be simulated within a given domain. In this approach, for reliable predictions, a large quantity of precise data are required to assign physical properties of the domain and model parameters and to calibrate the model simulations. In practice, however, sufficient data for model development are not readily obtained because of limitations of cost and time, resulting in model uncertainties and poor model performance.

Another approach for predicting GWL fluctuations uses data-based time-series models. These models require time-series data of the GWL and relevant input variables only, but are limited to forecasting temporal variations at a fixed location. Many studies used Box and Jenkins transfer function noise models, assuming a linear relationship between the input and output series, to predict the GWL fluctuations (Box and Jenkins, 1976; Tankersley et al., 1993; van Geer and Zuur, 1997; Bierkens et al., 1999). In recent years, an artificial neural network (ANN) has been applied to solving various water resource problems including time-series forecasting (Zealand et al., 1999; Sudheer et al., 2002; Cigizoglu, 2003; Almasri and Kaluarachchi, 2005; Yoon et al., 2007). The ANNs are considered as standard nonlinear estimators and their abilities have been verified in a variety of fields. In hydrology, the ANN models have been satisfactorily applied to the prediction of nonlinear hydrologic processes such as rainfall runoff, stream flow, precipitation, and water quality modeling in the 1990s (ASCE, 2000b). Maier and Dandy (2000) reviewed 43 papers of ANN applications to water resource variables that had been published until the end of 1998. Among these papers, surface water flow and quality was the topic in 28 and rainfall forecasting in 13. Only two papers were associated with water tables using synthetic data. In the 2000s, studies on the prediction of GWL fluctuations in the real environment have increased. Coulibaly et al. (2001) assessed the performance of three types of functionally different ANN models for prediction of GWL fluctuations using hydrometeorological data such as past GWL, rainfall, river stage, and temperature. Coppola et al. (2003, 2005) used ANNs to build a GWL prediction model under variable pumping and climate conditions. A few

* Corresponding author. Tel./fax: +82 2 873 3647.
E-mail addresses: oolahee1@snu.ac.kr (H. Yoon), skybeast@hanmail.net (S.-C. Jun), yjhyun@snu.ac.kr (Y. Hyun), gokbae@snu.ac.kr (G.-O. Bae), kklee@snu.ac.kr (K.-K. Lee).

0022-1694/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2010.11.002

recent studies applied ANN models to predicting the GWL in coastal aquifers. Nayak et al. (2006) and Krishna et al. (2008) successfully predicted the GWL fluctuation in coastal aquifers using ANN models with input variables such as meteorological information and GWL data.

Support vector machines (SVMs), which were introduced by Vapnik (1995), are a relatively new structure in the data-driven prediction field. The SVM is based on the structural risk minimization (SRM), instead of the empirical risk minimization (ERM) of ANNs, which can cause the solution to be captured in a local minimum and the network overfitted. The SRM minimizes the empirical error and model complexity simultaneously, which can improve the generalization ability of the SVM for classification or regression problems in many disciplines. Recently, the SVMs have been applied to the prediction of water resource variations forward in time, mainly for surface water. Many researchers have verified the SVM performance for the prediction of rainfall runoff (Dibike et al., 2001), stream flow or stage (Liong and Sivapragasam, 2002; Asefa et al., 2006; Yu et al., 2006), and lake water level (Asefa et al., 2005; Khalil et al., 2006; Khan and Coulibaly, 2006). Most surface water researches have showed that SVM performance is comparable or even superior to ANN's. Despite the growing applications and successes of SVM in the surface water problems, there have been only a few studies related to subsurface water. Asefa et al. (2004) and Khalil et al. (2005) used the SVM to capture the spatial distribution features of groundwater head and quality, respectively. Gill et al. (2006) predicted the soil moisture using SVMs, based on past measurements of soil moisture and meteorological data. More recently, Gill et al. (2007) compared the performance of ANN and SVM for predicting GWL under conditions of incomplete data that were assumed to be missing at random.

In most of the previous researches, the precipitation has been used as a natural exogenous variable for the prediction of daily, weekly or monthly GWL. In a coastal aquifer, however, tidal level variations should be also considered as well as precipitation which is the main driving forces of groundwater recharge. In order to predict GWL as a rapid response to the precipitation and tide, it is necessary to consider a time step shorter than a day. The objective of this study was to develop and compare data-based time-series forecasting models for short-term GWL fluctuations in a coastal aquifer due to recharge from precipitation and tidal effect. The GWL prediction in a coastal aquifer should be performed prudently because it is related to quality problems caused by salt water intrusion as well as quantity problems. Six-hourly GWL predictions considering the rapid and sharp fluctuations in a coastal aquifer were conducted in this study. Time-series forecasting models were developed using ANN and SVM, then the performances of the two models were compared for various combinations of input variables including past measurements of precipitation, tide level, and GWL. We examined the influence of multistep lead times, which stands for a degree of time step to be predicted, on the model performance. We also evaluated generalization ability and model parameter uncertainty of the developed models.

2. Methods

2.1. Artificial neural networks

The ANN is a flexible mathematical structure patterned after the biological nervous system. In general, a common ANN structure, called a multilayer perceptron network (MLPN), consists of input, hidden, and output layers with their nodes and activation functions. For a network training method, the back-propagation algorithm (BPA) introduced by Rumelhart et al. (1986) can effectively train the network for nonlinear problems, which has stimulated a torrent of research in neural networks.

In this study, an MLPN with one hidden layer trained by BPA was used to build the ANN model. The activation function consists of a log-sigmoid function in the hidden layer and a linear function in the output layer. It has been reported that ANNs with this configuration are the most commonly used form, as they have improved extrapolation ability (ASCE, 2000a; Maier and Dandy, 2000). The mathematical expression of the MLPN is as follows:

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

where x_i is the ith nodal value in the previous layer, y_j is the jth nodal value in the present layer, b_j is the bias of the jth node in the present layer, w_{ji} is a weight connecting x_i and y_j, N is the number of nodes in the previous layer, and f is the activation function in the present layer. Fig. 1a shows a schematic diagram of the MLPN used in this study. Detailed mathematical descriptions of the BPA can be found in Hagan et al. (1996) and ASCE (2000a).

To prevent the solution from being captured in some local minimum, a momentum term was added in the weight updating process. The momentum has an averaging effect, and diminishes drastic fluctuations in weight changes over consecutive iterations (Rumelhart et al., 1986). Because improper initial weights can cause the local minimum problem, the ANN was trained using 100 sets of random initial weights to select the best initial weights.

2.2. Support vector machines

The SVM, a relatively new learning system, is based on statistical learning theory (Vapnik, 1995, 1998). The structure of an SVM is not determined *a priori*. Input vectors supporting the model structure are selected through a model training process described below. Given by a set of N samples of $\{\mathbf{x}_k, y_k\}_{k=1}^{N}$, $\mathbf{x} \in R^m$, $y \in R$, where $\mathbf{x}$ is an input vector of m components and y is a corresponding output value, an SVM estimator (f) on regression can be expressed as:

$$f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b, \tag{2}$$

where $\mathbf{w}$ is a weight vector, and b is a bias. 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. Vapnik (1995) introduced the following convex optimization problem with an ε-insensitivity loss function to obtain the solution to Eq. (2):

$$\begin{aligned} &\underset{\mathbf{w},b,\xi,\xi^*}{\text{minimize}} \quad \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{k=1}^{N}(\xi_k + \xi_k^*) \\ &\text{subject to} \quad \left\{\begin{array}{l} y_k - \mathbf{w}^T\phi(\mathbf{x}_k) - b \leqslant \varepsilon + \xi_k \\ \mathbf{w}^T\phi(\mathbf{x}_k) + b - y_k \leqslant \varepsilon + \xi_k^* \\ \xi_k,\ \xi_k^* \geqslant 0 \end{array}\right\} \quad k = 1,\ 2,\ \cdots,\ N \end{aligned} \tag{3}$$

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. (3) is usually solved in a dual form using Lagrangian multipliers and imposing the Karush–Kuhn–Tucker (KKT) optimality condition. The input vectors that have nonzero Lagrangian multipliers under the KKT condition support the structure of the estimator and are called support vectors. The architecture of SVM is shown in Fig. 1b.

A number of algorithms have been suggested for solving the dual optimization problem of the SVM. An overview of these

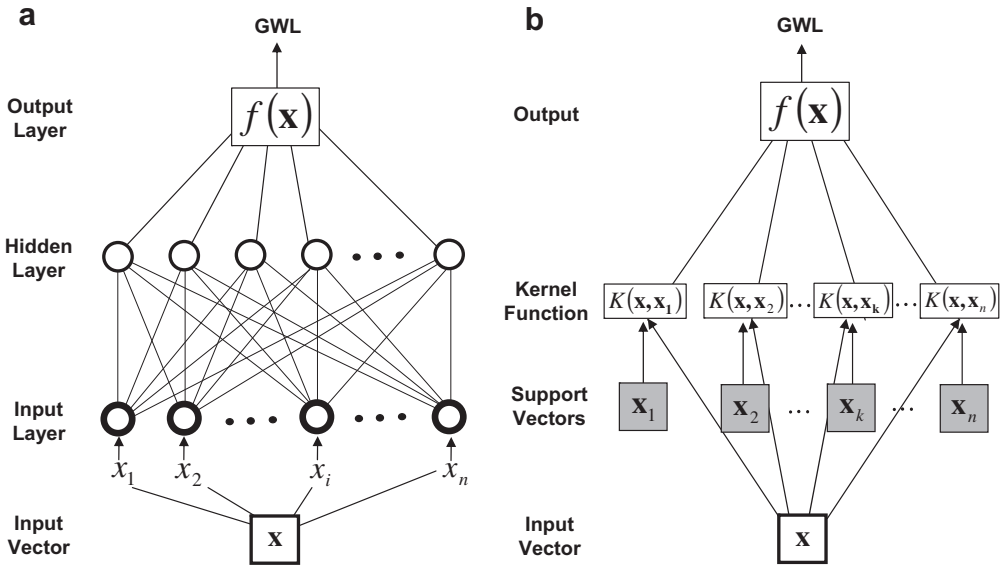

Fig. 1. A schematic diagram of the models used in this study: (a) ANN and (b) SVM.

algorithms is found in Shevade et al. (2000) and Schölkopf and Smola (2002). Conventional quadratic programming algorithms require extremely large memory for the kernel matrix computation and have difficulties in their implementation. Therefore, they are not suitable for large problems. To overcome this problem, subset selection methods have been developed. The optimization problem is solved in a selected subset to give a set of support vectors, and then a new subset is selected using these support vectors. This process continues until all the input vectors satisfy the KKT conditions. The sequential minimal optimization (SMO) algorithm, introduced by Platt (1999), puts the subset selection algorithm to the extreme by selecting a subset of size two and optimizing the estimation function with respect to them. The main advantage of the SMO is that an analytical solution of a subset can be obtained directly without invoking a quadratic optimizer. In this study, the SMO algorithm was employed to train the SVM model for GWL predictions. The detailed procedures of the SMO algorithm are found in Platt (1999) and Schölkopf and Smola (2002).

3. Study site and data descriptions

The study site is located at a beach of the coastal town of Mukho in Donghae city, Korea (Fig. 2). Average air temperature ranges from 2.3 °C in the winter season (December–February) to 21.9 °C in the summer season (June–August). The average annual precipitation is 1507 mm over the past 10 years. To study the groundwater behavior in this coastal aquifer, a total of 11 monitoring wells were installed, as shown in Fig. 2. Five of them (OFs and OBs) were installed near the seashore in 2003 but they were all washed out by high waves in late 2004. Six new wells (BHs) were installed for groundwater monitoring in 2004. Automated data loggers (Diver®, DI263 and DI250 models, Van Essen Instruments, Netherlands) were installed at depths of 4, 7.5, and 11 m in well OB2, and 5, 10, and 15 m in well BH5 to measure GWL, groundwater temperature, electrical conductivity, and air pressure. All data were gathered from June 2004 to the end of 2004 at well OB2 and from April 2005 at well BH5.

The precipitation and tide are considered as exogenous factors affecting the GWL in this area. The precipitation data were collected at the Donghae meteorological station of the Korea Meteorological Administration and the tide-level data were collected at the Mukho tidal station of the National Oceanographic Research Institute. The stations are located about 8.0 km and 3.7 km away from the study site, respectively. In this study, six-hourly measured time-series data of the precipitation, tide, and GWL were used to predict future GWL fluctuations on well OB2 in 2004 and well

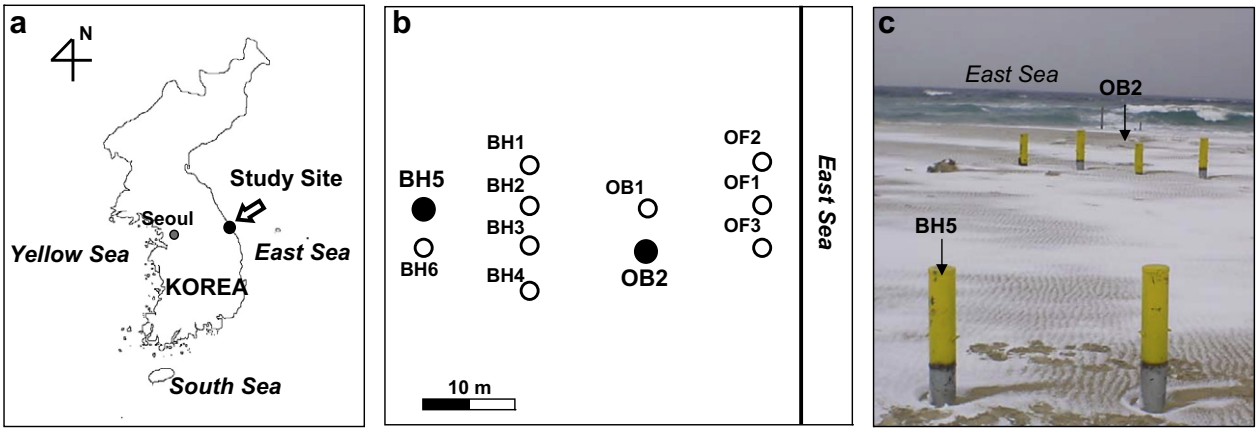

Fig. 2. Description of the study site: (a) location of the study site, (b) distribution of wells, and (c) picture of the study site.

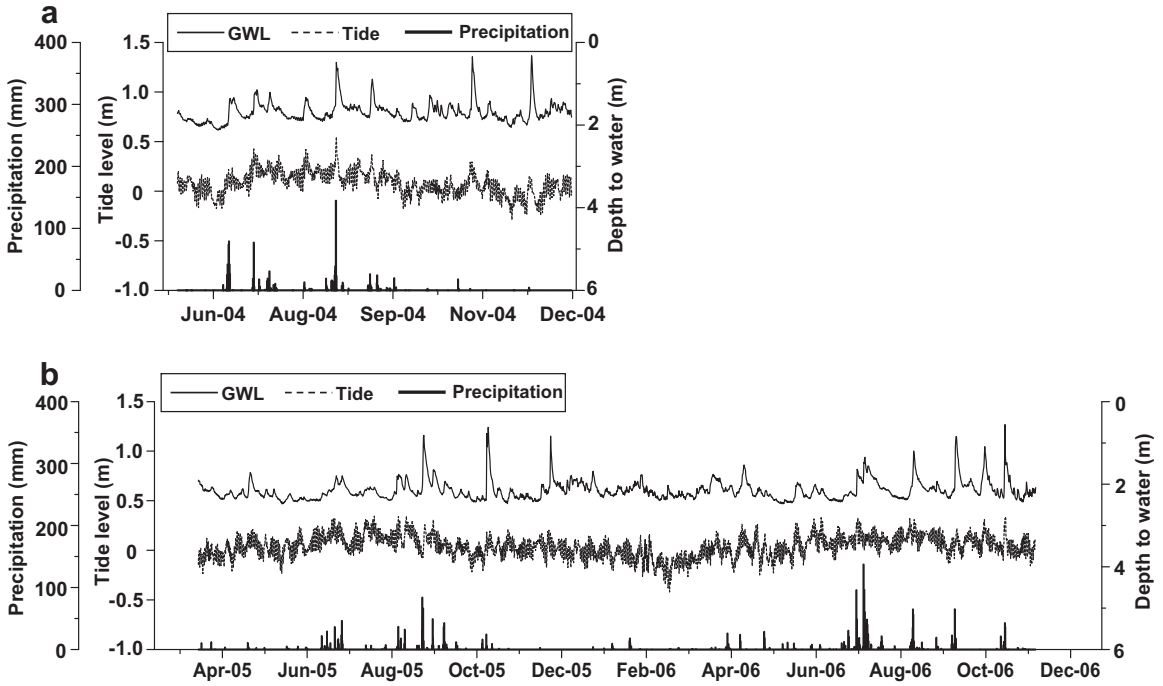

Fig. 3. Time-series data collected at the study site: (a) well OB2, (b) well BH5.

BH5 from 2005 to 2006. Fig. 3 illustrates the time-series data of precipitation, tide level, and GWL collected in the study site. The GWL shows dynamic fluctuations with many high peaks and the variation can exceed 1.0 m in a day. Moreover, its fluctuation is affected not only by precipitation but also by tide level with a fast response and a nonlinear relationship. Therefore, to understand the complex behavior of GWL and manage the groundwater resource effectively in this coastal aquifer, it is necessary to predict the GWL precisely. The present study evaluates and compares the model performances under various conditions for an accurate GWL prediction.

4. Model development

4.1. Data preprocessing

The ANN and SVM models were developed using the C programming language for predicting the GWL in wells OB2 and BH5 separately. Well OB2 provided GWL data for 7 months and BH5 well for 20 months. Each time-series dataset was divided into three parts in accordance with the three stages in the model-building process: training, testing, and predicting. In the training stage, weights were updated using the weight updating rules described above: BPA for ANN and SMO for SVM. Several model parameters were selected: the number of hidden nodes, learning rate (γ), momentum (β) and initial weights for the ANN model construction, and the tradeoff parameter (C), error tolerance (ε), and kernel parameter (σ) for the SVM. The weight updating procedure stops when the errors in the testing stage starts to increase for ANN and all input vectors satisfy the KKT conditions for SVM. The model parameters were selected in the testing stage to minimize errors in the test dataset by a trial and error approach. After the model-building process was completed through the calibration, namely training and testing stages, the developed model was validated in the predicting stage. Table 1 describes the number and span of data used in each stage for wells OB2 and BH5. All the time-series data were normalized using the minimum (X^{TR}_{min}) and maximum (X^{TR}_{max}) values of the training data set as described in Eq. (4), so that the variables (X) in the training data set ranged from 0 to 1. After the training, overall results of the normalized X are inverted to X using:

$$\text{normalized } X = \frac{X - X^{TR}_{min}}{X^{TR}_{max} - X^{TR}_{min}} \quad (4)$$

4.2. Input structure

We considered precipitation (P), tide level (T) and GWL (G) as input variables. Five combinations of input variables were generated: P + T, G, P + G, T + G and P + T + G. The input structure with the minimum error in the testing stage was selected for building a model. Lag times of precipitation and tide level were determined to be 4 time steps (1 time step stands for 6 h) by cross-correlation analyses between precipitation or tide level and GWL. The correlation coefficient value between precipitation and GWL was 0.39 and that between tide level and GWL was 0.54 for well OB2. They were 0.31 and 0.39 for well BH5. Input selection method using statistical properties such as cross-correlation has been often employed for the research of time-series data forecasting (Sajikumar and Thandaveswara, 1999; Coulibaly et al., 2000; Sudheer et al., 2002; Nayak et al., 2006). The present study focused on examining the relative importance of input variables. Therefore, the lag time of GWL was set equal to other input variables. For the ANN, the input combinations with related lag times define the nodes in the input layer, whereas they define the components of the input vectors for the SVM. In this study, multiple lead-time predictions of 1, 2, 4, 6 and 8 time steps were also performed for each input structure. These input structures and lead times constituted a total of 25 cases of a model design.

4.3. Performance criteria

In the model-building process, root mean squared error (RMSE) was used for the model calibration through the training and testing

Table 1
The number and span of data allocated to each stage of the two wells.

Well		Training	Testing	Predicting	Total
OB2	Number	368	244	231	843
	Span	June 2004–August 2004	September 2004–October 2004	November 2004–December 2004	June 2004–December 2004
BH5	Number	976	604	790	2370
	Span	May 2005–November 2005	December 2005–April 2006	May 2006–November 2006	May 2005–November 2006

stages. The RMSE is a quadratic scoring rule that measures the average magnitude of the error between target (t) and output (o) values. Because the errors are squared before they are averaged, large errors take a relatively high weight. Therefore, RMSE is a useful error index when large errors are particularly undesirable. For the model validation in the predicting stage, the RMSE, the mean error (ME), the mean absolute percentage error (MAPE), the correlation coefficient (CORR), the Nash–Sutcliffe efficiency (NS) and the Akaike Information Criterion (AIC) were evaluated as model performance criteria. The ME measures the bias of overall errors, MAPE the relative magnitude of errors, which enables a comparison of errors that differ in level, CORR the strength and direction of a linear relationship between target and output values, NS the predictive power of hydrological model, and AIC the goodness of fit of a model considering the penalty of model complexity. Mathematical expressions of the performance criteria are given by:

$$\mathrm{ME} = \frac{1}{l}\sum_{i=1}^{l}(t_i - o_i), \tag{5}$$

$$\mathrm{MAPE} = 100 \times \frac{1}{l}\left(\sum_{i=1}^{l}\frac{|t_i - o_i|}{t_i}\right), \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(t_i - o_i)^2}, \tag{7}$$

$$\mathrm{CORR} = \frac{\frac{1}{l}\sum_{i=1}^{l}(t_i - \overline{t_i})(o_i - \overline{o_i})}{\sqrt{\frac{1}{l}\sum_{i=1}^{l}(t_i - \overline{t_i})^2}\sqrt{\frac{1}{l}\sum_{i=1}^{l}(o_i - \overline{o_i})^2}}, \tag{8}$$

$$\mathrm{NS} = 1 - \frac{\sum_{i=1}^{l}(t_i - o_i)^2}{\sum_{i=1}^{l}(t_i - \overline{t})^2}, \tag{9}$$

$$\mathrm{AIC} = 2k + l\ln\left(\frac{1}{l}\sum_{i=1}^{l}(t_i - o_i)^2\right) \tag{10}$$

where l is the total number of data points, k is the number of model parameters, and $\overline{t}$ and $\overline{o}$ are mean values of t and o, respectively.

5. Results and discussion

5.1. Model calibration

The model calibration and validation results for 25 cases of the model design are analyzed. Table 2a and 2b shows RMSE values in training and testing stages for all cases. In the training stage, the mean RMSE values for ANN and SVM models are 0.107 and 0.150, respectively, at well OB2, and 0.109 and 0.160 at well BH5. In the testing stage, they are 0.130 and 0.139 at well OB2, and 0.129 and 0.133 at well BH5. The mean RMSE values of the ANN models are smaller than those of the SVM models in both the training and testing stages, which implies that the calibration

Table 2a
RMSE values in training (TR) and testing (TS) stages for well OB2.

		LT1		LT2		LT4		LT6		LT8	
		TR	TS	TR	TS	TR	TS	TR	TS	TR	TS
P + T	ANN	0.080	0.145	0.091	0.148	0.115	0.152	0.137	0.160	0.114	0.167
	SVM	0.173	0.147	0.179	0.155	0.169	0.170	0.198	0.169	0.221	0.169
G	ANN	0.062	0.056	0.101	0.098	0.147	0.145	0.196	0.169	0.101	0.175
	SVM	0.089	0.078	0.097	0.109	0.143	0.147	0.182	0.179	0.210	0.179
P + G	ANN	0.046	0.055	0.083	0.095	0.115	0.133	0.153	0.163	0.140	0.174
	SVM	0.044	0.058	0.076	0.100	0.162	0.144	0.170	0.174	0.213	0.183
T + G	ANN	0.057	0.053	0.091	0.089	0.132	0.143	0.079	0.167	0.110	0.170
	SVM	0.062	0.053	0.107	0.092	0.169	0.142	0.144	0.190	0.215	0.174
P + T + G	ANN	0.046	0.051	0.077	0.089	0.107	0.130	0.145	0.155	0.152	0.171
	SVM	0.047	0.065	0.087	0.106	0.132	0.145	0.195	0.165	0.220	0.172

Table 2b
RMSE values in training (TR) and testing (TS) stages for well BH5.

		LT1		LT2		LT4		LT6		LT8	
		TR	TS	TR	TS	TR	TS	TR	TS	TR	TS
P + T	ANN	0.106	0.196	0.074	0.152	0.080	0.151	0.089	0.155	0.095	0.163
	SVM	0.203	0.156	0.201	0.155	0.206	0.155	0.229	0.164	0.220	0.190
G	ANN	0.063	0.054	0.098	0.092	0.141	0.141	0.165	0.159	0.151	0.166
	SVM	0.060	0.053	0.098	0.090	0.248	0.187	0.179	0.153	0.205	0.154
P + G	ANN	0.053	0.052	0.085	0.091	0.134	0.130	0.151	0.164	0.165	0.166
	SVM	0.058	0.054	0.129	0.125	0.127	0.127	0.176	0.146	0.202	0.156
T + G	ANN	0.063	0.050	0.102	0.090	0.207	0.153	0.111	0.156	0.103	0.156
	SVM	0.074	0.080	0.145	0.127	0.150	0.124	0.182	0.150	0.201	0.170
P + T + G	ANN	0.051	0.050	0.084	0.090	0.122	0.134	0.122	0.147	0.114	0.158
	SVM	0.066	0.060	0.135	0.128	0.131	0.125	0.176	0.141	0.206	0.154

capability of the ANN model is better than that of the SVM model for the given data. The selected input structure that minimizes the testing error is described in Table 3. G is the most frequently selected input variable, implying that the use of autoregressive data is important in this time-series model building. It is noted that T is more frequently selected than P. The GWL fluctuation in this site is affected by the tide in a fast and direct manner because of its close proximity to the coast. Moreover, the tide-level data include a little information on precipitation as the cross-correlation analysis for precipitation and tide level in the rainy season of the training stage shows that the maximum correlation coefficient is about 0.3. This can cause the tide level to be a more influential input variable than precipitation for the study site, although in general precipitation is the main natural driving force for GWL fluctuations. These results indicate that the tidal effect on the GWL should be considered for GWL prediction in coastal aquifers.

Table 3
The selected input structures.

Lead times	OB2		BH5	
	ANN	SVM	ANN	SVM
LT1	P + T + G	T + G	P + T + G	G
LT2	T + G	T + G	T + G	G
LT4	P + T + G	T + G	P + G	T + G
LT6	P + T + G	P + T + G	P + T + G	P + T + G
LT8	P + T	P + T	T + G	P + T + G

5.2. Prediction

Fig. 4 shows examples of GWL prediction results for well BH5 for lead times of 1, 4 and 8, showing the tendency for errors to increase greatly with lead time. As described in Fig. 3, the GWL has a fast response to precipitation and tide level, which can cause poor prediction results at longer lead times. Fig. 4 also shows the superiority of SVM models to ANN for large lead times in well BH5. The model prediction results of the selected input structures for all lead times are shown in Table 4. ME values for well OB2 indicate an overestimation for both models, but an underestimation for well BH5. However, the magnitude of the ME values in well OB2 is higher than that in well BH5, implying a higher bias of the prediction results in well OB2. MAPE and RMSE values also show that the prediction results for well BH5 are better than those for OB2. CORR and NS values are not much different for wells and models. AIC values show the superiority of SVM models especially for well BH5. The best model performance is found for the SVM model for well BH5 for all criteria. Fig. 5 shows RMSE and CORR values with lead times. The prediction error tends to increase with the lead time. For well OB2, the performance of the two models is similar (Fig. 5a and c). However, for well BH5, the performance of the SVM model is better than that of the ANN model, especially for large lead times (Fig. 5b and d). The cumulative distribution functions (CDFs) of observed and estimated GWLs are given in Fig. 6. The degree of deviation from the CDF of observed GWLs for ANN models (Fig. 6a and c) become higher than SVM (Fig. 6b and d) with

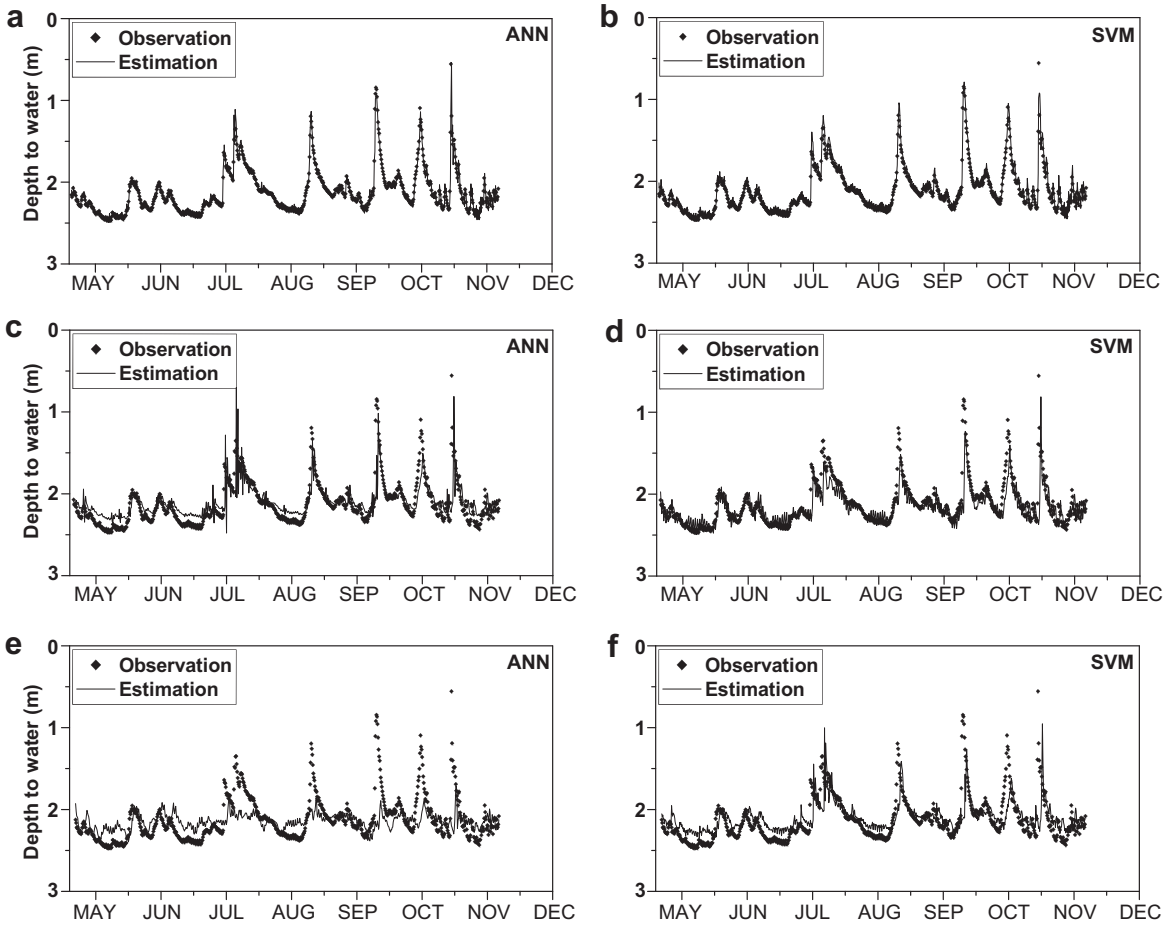

Fig. 4. Examples of model prediction results of ANN and SVM models in well BH5 for lead times of 1, 4, and 8: (a) ANN model for lead time 1, (b) SVM model for lead time 1, (c) ANN model for lead time 4, (d) SVM model for lead time 4, (e) ANN model for lead time 8, and (f) SVM model for lead time 8.

Table 4
The model prediction errors for the selected input structures.

Well	Model	ME ($\times 10^{-3}$ m)	MAPE (%)	RMSE (m)	CORR	NS	AIC ($\times 10^{3}$ m)
OB2	ANN	−16.6	4.25	0.186	0.768	0.576	−3.74
	SVM	−26.6	4.42	0.185	0.775	0.582	−3.76
BH5	ANN	8.06	3.42	0.184	0.733	0.534	−13.2
	SVM	3.40	2.76	0.165	0.793	0.628	−21.4

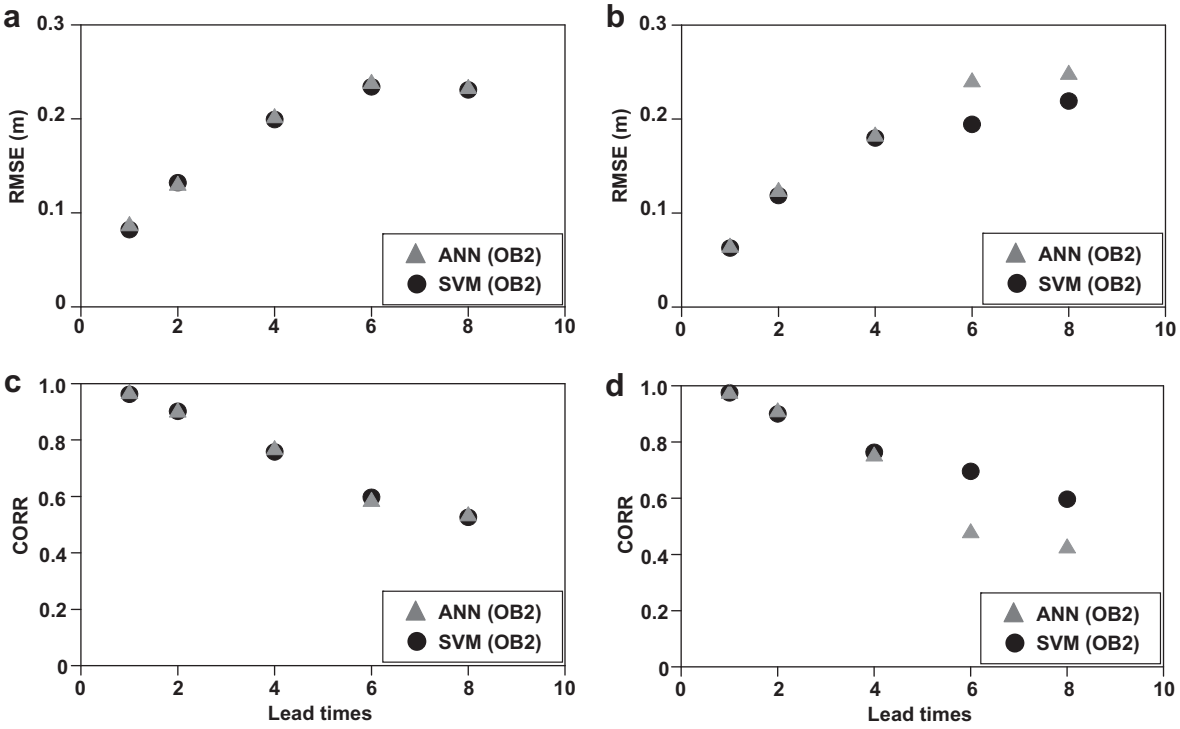

Fig. 5. Comparison of the model prediction results for the selected input structures: (a) RMSE values for well OB2, (b) RMSE values for well BH5, (c) CORR values for well OB2, and (d) CORR values for well BH5.

increasing lead times; and the degree of deviation for OB2 well is higher than BH5 well.

The volume of model calibration data for well BH5 is about 2.6 times as much as for well OB2. However, the correlation coefficient between input variables and GWL for well BH5 is lower than well OB2. This implies that the time-series data of well BH5 can have a higher nonlinearity than those of well OB2 and include more noisy data that hinder the model training process. The result of model performances for well BH5 indicates that the SVM model is more likely to catch the nonlinear relationship for the given data and to filter out the noise than the ANN model in this case.

5.3. Generalization ability

In the development of data-driven models, such as ANN and SVM, guaranteeing the model's generalization ability is one of the most important problems. In this study, to evaluate this ability, two indices for model calibration (CAL) and validation (VAL) are defined as follows:

$$\mathrm{CAL} = \frac{\text{RMSE in testing stage}}{\text{RMSE in training stage}}$$
$$\mathrm{VAL} = \frac{\text{RMSE in predicting stage}}{\text{RMSE in training stage}} \quad (11)$$

If a developed model learns a given system perfectly, then the CAL and VAL values are unity. However, if the model concentrates on learning the given training data rather than creating a more general system, then the index values exceed unity, meaning overtraining. The model might find a local solution of the optimization problem. Values of the indices less than unity indicate that the model is undertrained. To evaluate the generalization ability of the two models, prediction errors for all input structures that include unselected input structures are calculated. Fig. 7 shows CAL and VAL values of the models and wells calculated for all 25 cases. Sixty-eight percent of the CAL values for the ANN models and 26% for the SVM models exceed unity, and 100% of the VAL values for both models exceed unity. Although the minimum CAL values for the two models are close to each other, the CAL values of the ANN model include higher and more widely distributed values than those of the SVM model. On the other hand, the CAL values of the SVM models are close to unity and narrowly distributed, especially for well BH5.

For the VAL values, the discrepancy of the two models is higher than that for the CAL values. These results indicate that the generalization ability of the ANN model is more sensitive to the selection of the input structure and lead time than that of the SVM model. It is noted that the CAL and VAL values of ANN models for well BH5 are more widely distributed than those for well OB2, which is opposite to the SVM models. The overall generalization ability of the SVM model is superior to that of the ANN model in this study site. These results imply that the ERM learning theory embedded in ANN. minimizing errors of the given data can cause the model to be overtrained, whereas the SRM in SVM minimizing the empirical error and model complexity simultaneously can improve the generalization ability.

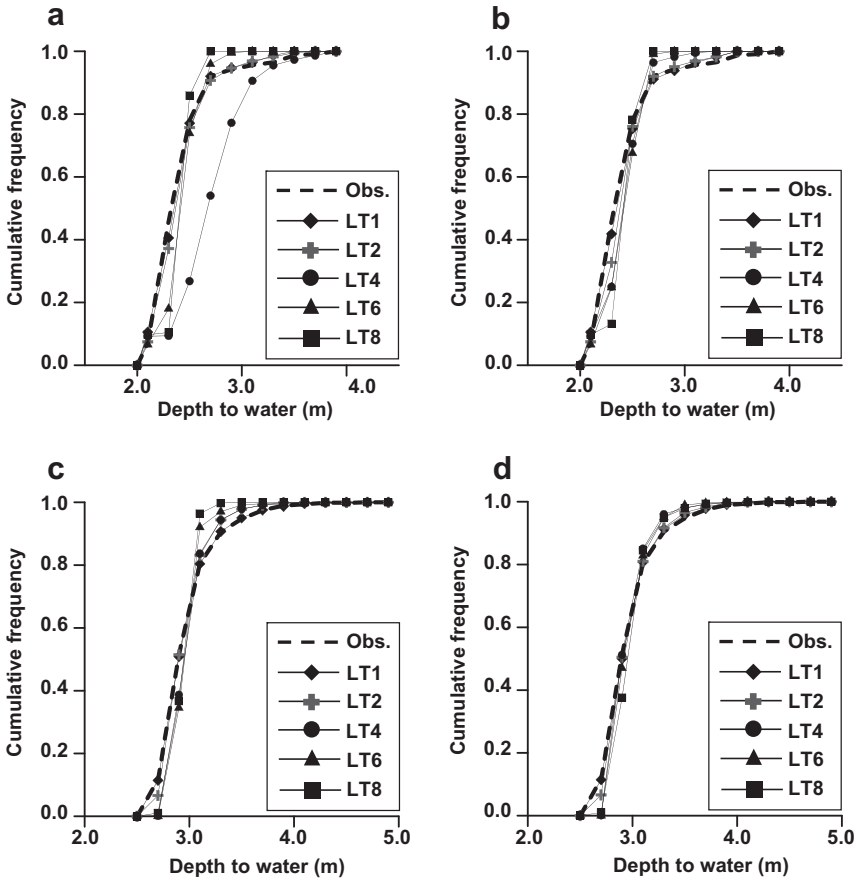

Fig. 6. Cumulative distribution functions (CDFs) of observed and estimated GWLs with lead times: (a) ANN model for well OB2, (b) SVM model for well OB2, (c) ANN model for well BH5, and (d) SVM model for well BH5.

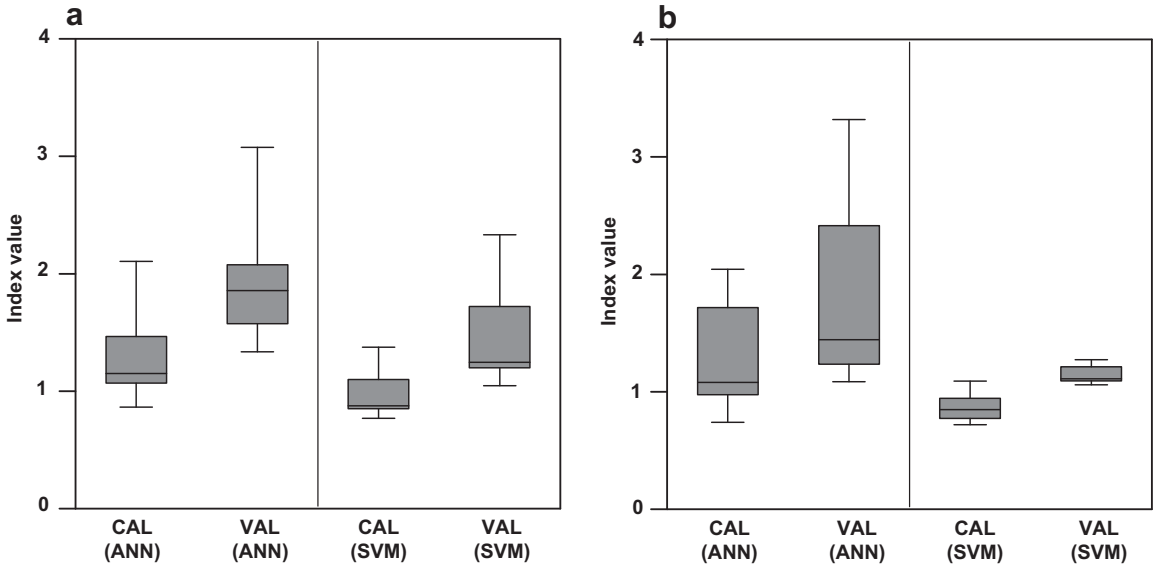

Fig. 7. Box plots of: (a) CAL and (b) VAL values of the models and wells for all 25 cases of the model design.

5.4. Model parameter uncertainty

There have been plenty of discussions on equifinality of model parameters in physically based hydrological modeling (Schulz et al., 1999; Beven and Freer, 2001; Pappenberger et al., 2005; Todni, 2007; Bardossy, 2007; Bardossy and Singh, 2008). In this study, uncertainty analysis was conducted for evaluating the equifinality of the data-driven model parameters. The developed 1-lead time models of well OB2 were used and 500 model parameter sets for each model were generated for this uncertainty analysis. The

model performance was estimated using the NS value for each parameter set.

Fig. 8 shows the variability of model performances for the parameter sets of ANN and SVM. In this case study, the overall uncertainty in model parameters of the ANN is higher than that of SVM. The model uncertainty tends to increase with momentum of the ANN and ε of SVM, but no specific trend is found for other parameters. Fig. 8 also shows relatively parallel maximum lines of box plots for their ranges. This structure of the plots indicates that single optimal parameter set could not be clearly determined. In order to identify the equifinality of model parameters, prediction results of five different parameter sets of highest NS values for each ANN and SVM are compared. Fig. 9 illustrates CDFs of estimated GWLs for the selected parameter sets with the observed. To evaluate the validity of two models for this uncertainty analysis, a two-sample Kolmogorov–Smirnov (K–S) test was applied. The K–S test measures maximum distance between two CDFs in order to examine the null hypothesis, that the observed and predicted GWL CDFs are not different. The p value of the K–S Z statistic for all selected parameter sets is greater than 0.05, which indicates that the null hypothesis of this study cannot be rejected at 5% significance level. Table 5 summarizes the result of K–S tests and closely distributed NS values of the selected parameter sets, implying the equifinality of model parameters in this modeling approach based on ANN and SVM.

Table 5
Results of the K–S test for five parameter sets of highest NS values.

Model	Statistics	Set 1	Set 2	Set 3	Set 4	Set 5
ANN	NS value	0.904	0.903	0.900	0.895	0.895
	K–S Z statistics	0.704	0.657	0.751	0.469	0.610
	p value	0.705	0.781	0.626	0.469	0.610
SVM	NS value	0.933	0.931	0.929	0.927	0.927
	K–S Z statistics	0.939	0.986	1.079	0.986	1.033
	p value	0.342	0.286	0.194	0.286	0.237

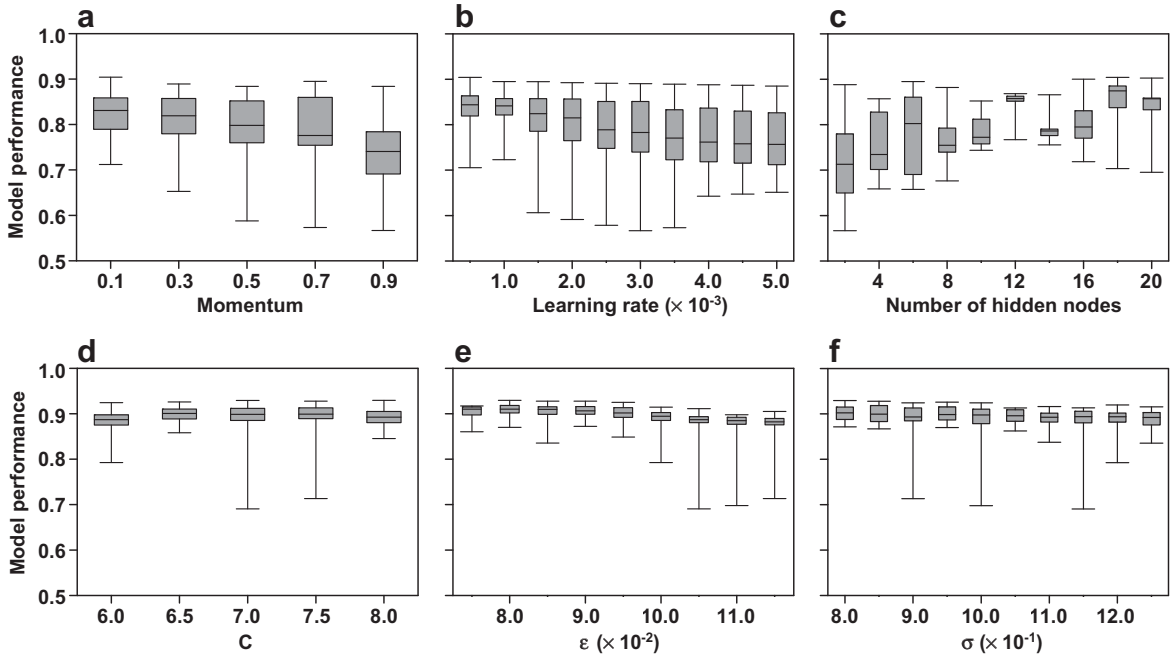

Fig. 8. Variability of model performances for 500 model parameter sets in the uncertainty analysis.

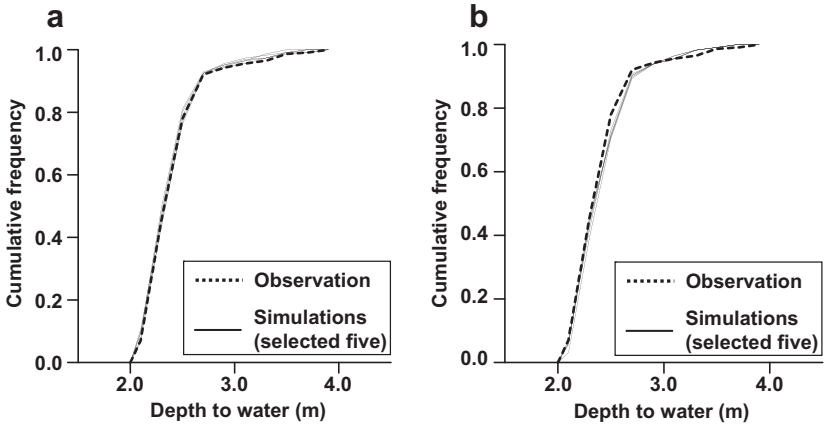

Fig. 9. CDFs of estimated GWLs for selected parameter sets in the uncertainty analysis.

6. Summary and conclusions

We have developed the time-series forecasting models for the short-term GWL fluctuation in a coastal aquifer using ANN and SVM, and we have compared the model performances for input structures and lead times. The RMSE analyses for the 25 cases of the model design show that the ANN model is effective in the training and testing stages. The past GWL data are the most frequently selected input variable in this time-series model building. The tide-level data are more frequently selected as an input variable than precipitation, which implies that tidal effects should be considered when developing time-series models in coastal aquifers. The prediction results of the selected input structures are similar for the two models in well OB2, while SVM performance is better than ANN's in well BH5, especially for larger lead times. The generalization ability of the SVM model is superior to that of the ANN model, as this model is more sensitive to the input structure and lead time than the SVM model. The uncertainty analysis which is conducted for well OB2 with 1-lead time detects the equifinality of model parameter sets and higher uncertainty for ANN models than SVM in this case. Thus, it is suggested that the model-building process should be carefully conducted when using ANN models for GWL forecasting in coastal aquifers. These results indicate that the SVM model of this study is more likely to learn the complex relationship for the given data and to filter out the noise than the ANN.

This study focused on short-term GWL fluctuations using data measured at 6-h interval. Recently, many water resource utilities have implemented automatic data measurement systems (Coppola et al., 2005). The Korean government has constructed and operated the National Groundwater Monitoring Network (NGMN) throughout the country since 1995. A total of 320 groundwater monitoring stations were implemented by 2005, and the construction of 199 additional stations has been suggested (Lee et al., 2007). Every monitoring station measures the GWL, groundwater temperature, and electrical conductivity every 6 h and the data are automatically collected and transmitted to a host server. Therefore, the real time data for GWLs measured at 6-h interval are available in NGMN. The developed models and the results of this study may be useful for the management of the automatic data measurement system such as that of the NGMN of Korea, especially in coastal aquifers.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (# 2010-0001449) and partly by Korea Ministry of Environment as "The GAIA Project".

References

Almasri, M.N., Kaluarachchi, J.J., 2005. Modular neural networks to predict the nitrate distribution in ground water using the on-ground nitrate loading and recharge data. Environ. Modell. Software 20, 851–871.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000a. Artificial neural networks in hydrology I: preliminary concepts. J. Hydrol. Eng. 5, 115–123.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000b. Artificial neural networks in hydrology II: hydrologic applications. J. Hydrol. Eng. 5, 124–137.

Asefa, T., Kemblowski, M.W., Urroz, G., McKee, M., Khalil, A., 2004. Support vector-based groundwater head observation networks design. Water Resour. Res. 40, W11509. doi:10.1029/2004WR003304.

Asefa, T., Kemblowski, M., Lall, U., Urroz, G., 2005. Support vector machines for nonlinear state space reconstruction: application to the Great Salt Lake time series. Water Resour. Res. 41, W12422. doi:10.1029/2004WR003785.

Asefa, T., Kemblowski, M., McKee, M., Khalil, A., 2006. Multi-time scale stream flow predictions: the support vector machines approach. J. Hydrol. 318, 7–16.

Bardossy, A., 2007. Calibration of hydrological model parameters for ungauged catchments. Hydrol. Earth Syst. Sci. 11, 703–710.

Bardossy, A., Singh, S.K., 2008. Robust estimation of hydrological model parameters. Hydrol. Earth Syst. Sci. 12, 1273–1283.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology. J. Hydrol. 249, 11–29.

Bierkens, M.F.P., Knotters, M., van Geer, F.C., 1999. Calibration of transfer function-noise models to sparsely or irregularly observed time series. Water Resour. Res. 35, 1741–1750.

Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis – Forecasting and Control. Holden-Day, San Francisco, California, USA. 598 p.

Cigizoglu, H.K., 2003. Estimation, forecasting and extrapolation of river flows by artificial neural networks. Hydrol. Sci. J. 48, 349–361.

Coppola, E., Szidarovszky, F., Poulton, M., Charles, E., 2003. Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping, and climate conditions. J. Hydrol. Eng. 8, 348–360.

Coppola, E., Rana, A.J., Poulton, M.M., Szidarovszky, F., Uhl, V.V., 2005. A neural network model for predicting aquifer water level elevations. Ground Water 43, 231–241.

Coulibaly, P., Anctil, F., Bobee, B., 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J. Hydrol. 230, 244–257.

Coulibaly, P., Anctil, F., Aravena, R., Bobee, B., 2001. Artificial neural network modeling of water table depth fluctuations. Water Resour. Res. 37, 885–896.

Dibike, Y.B., Velickov, S., Solomatine, D., Abbott, M.B., 2001. Model induction with support vector machines: introduction and applications. J. Comput. Civil Eng. 15, 208–216.

Gill, M.K., Asefa, T., Kemblowski, M.W., McKee, M., 2006. Soil moisture prediction using support vector machines. J. Am. Water Resour. Assoc. 42, 1033–1046.

Gill, M.K., Asefa, T., Kaheil, Y., McKee, M., 2007. Effect of missing data on performance of learning algorithms for hydrologic predictions: implications to an imputation technique. Water Resour. Res. 43, W07416. doi:10.1029/2006WR005298.

Hagan, M.T., Demuth, H.B., Beale, M., 1996. Neural Network Design. PWS Publishing Company, Boston, Massachusetts, USA. 651 p.

Khalil, A., Almasri, M.N., McKee, M., Kaluarachchi, J.J., 2005. Applicability of statistical learning algorithms in groundwater quality modeling. Water Resour. Res. 41, W05010. doi:10.1029/2004WR003608.

Khalil, A.F., McKee, M., Kemblowski, M., Asefa, T., Bastidas, L., 2006. Multiobjective analysis of chaotic dynamic systems with sparse learning machines. Adv. Water Resour. 29, 72–88.

Khan, M.S., Coulibaly, P., 2006. Application of support vector machine in lake water level prediction. J. Hydrol. Eng. 11, 199–205.

Krishna, B., Satyaji Rao, Y.R., Vijaya, T., 2008. Modelling groundwater levels in an urban coastal aquifer using artificial neural networks. Hydrol. Process. 22, 1180–1188.

Lee, J.Y., Yi, M.J., Yoo, Y.K., Ahn, K.H., Kim, G.B., Won, J.H., 2007. A review of the national groundwater monitoring network in Korea. Hydrol. Process. 21, 907–919.

Liong, S.Y., Sivapragasam, C., 2002. Flood stage forecasting with support vector machines. J. Am. Water Resour. Assoc. 38, 173–186.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. Environ. Modell. Software 15, 101–124.

Nayak, P.C., Satyaji Rao, Y.R., Sudheer, K.P., 2006. Groundwater level forecasting in a shallow aquifer using artificial neural network approach. Water Resour. Manage. 20, 77–90.

Pappenberger, F., Beven, K., Horritt, M., Blazkova, S., 2005. Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations. J. Hydrol. 302, 46–69.

Platt, J.C., 1999. Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smolar, A.J. (Eds.), Advances in Kernel Methods—Support Vector Learning. MIT Press, Cambridge, Massachusetts, USA.

Rumelhart, D.E., McClelland, J.L.The PDP Research Group, 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, Massachusetts, USA. 516 p.

Sajikumar, N., Thandaveswara, B.S., 1999. A non-linear rainfall–runoff model using an artificial neural network. J. Hydrol. 216, 32–55.

Schölkopf, B., Smola, A.J., 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, Massachusetts, USA. 626 p.

Schulz, K., Beven, K., Huwe, B., 1999. Equifinality and the problem of robust calibration in nitrogen budget simulations. Soil Sci. Soc. Am. J. 63, 1934–1941.

Shevade, S.K., Keerthi, S.S., Bhattacharyya, C., Murthy, K.R.K., 2000. Improvements to the SMO algorithm for SVM regression. IEEE Trans. Neural Network 11, 1188–1193.

Sudheer, K.P., Gosain, A.K., Ramasastri, K.S., 2002. A data-driven algorithm for constructing artificial neural network rainfall–runoff models. Hydrol. Process. 16, 1325–1330.

Tankersley, C.D., Graham, W.D., Hatfield, K., 1993. Comparison of univariate and transfer function models of groundwater fluctuations. Water Resour. Res. 29, 3517–3533.

Todni, E., 2007. Hydrological catchment modelling: past, present and future. Hydrol. Earth Syst. Sci. 11, 468–482.

van Geer, F.C., Zuur, A.F., 1997. An extension of Box–Jenkins transfer/noise models for spatial interpolation of groundwater head series. J. Hydrol. 192, 65–80.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York, USA. 314 p.
Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York, USA. 736 p.
Yoon, H., Hyun, Y., Lee, K.K., 2007. Forecasting solute breakthrough curves through the unsaturated zone using artificial neural networks. J. Hydrol. 335, 68–77.
Yu, P.S., Chen, S.T., Chang, I.F., 2006. Support vector regression for real-time flood stage forecasting. J. Hydrol. 328, 704–716.
Zealand, C.M., Burn, D.H., Simonovic, S.P., 1999. Short-term streamflow forecasting using artificial neural networks. J. Hydrol. 214, 32–48.