Self-optimizer data-mining method for aquifer level prediction

Omid Bozorg-Haddad, Mohammad Delpasand and Hugo A. Loáiciga

ABSTRACT

Groundwater management requires accurate methods for simulating and predicting groundwater processes. Data-based methods can be applied to serve this purpose. Support vector regression (SVR) is a novel and powerful data-based method for predicting time series. This study proposes the genetic algorithm (GA)–SVR hybrid algorithm that combines the GA for parameter calibration and the SVR method for the simulation and prediction of groundwater levels. The GA–SVR algorithm is applied to three observation wells in the Karaj plain aquifer, a strategic water source for municipal water supply in Iran. The GA–SVR's groundwater-level predictions were compared to those from genetic programming (GP). Results show that the randomized approach of GA–SVR prediction yields R^2 values ranging between 0.88 and 0.995, and root mean square error (*RMSE*) values ranging between 0.13 and 0.258 m, which indicates better groundwater-level predictive skill of GA-SVR compared to GP, whose R^2 and *RMSE* values range between 0.48–0.91 and 0.15–0.44 m, respectively.

Key words | genetic algorithm, groundwater level, prediction, simulation, support vector regression

Omid Bozorg-Haddad (corresponding author)
Mohammad Delpasand
Department of Irrigation & Reclamation Engineering, Faculty of Agricultural Engineering & Technology, College of Agriculture & Natural Resources,
University of Tehran,
Karaj, Tehran,
Iran
E-mail: *obhaddad@ut.ac.ir*

Hugo A. Loáiciga
Department of Geography,
University of California,
Santa Barbara, California 93106,
USA

INTRODUCTION

Groundwater is a vital source of municipal, industrial, and agricultural water use worldwide. One important aspect of groundwater management is the prediction of groundwater levels. This is mostly done by simulating groundwater flow with numerical models. This paper proposes an alternative approach to the prediction of groundwater levels, namely using a data-based methodology that relies on time series of observed groundwater levels and other relevant variables such as precipitation, recharge, and aquifer discharge. This paper shows that the application of data-based groundwater-level prediction models constitutes an alternative method to predicting groundwater levels bypassing the implementation of numerical groundwater simulations when aquifer characteristics (hydraulic conductivity, storativity, recharge, boundary, and initial conditions, etc.) are poorly known.

Box & Jenkins (1976) proposed a linear relation between input and output series to predict groundwater levels. Ever since, new data-based prediction methods have emerged with vastly improved capacities, such as artificial neural network (ANN), adaptive neural fuzzy inference system (ANFIS), genetic programming (GP), and support vector regression (SVR). Several papers reported the use of ANN for groundwater prediction and simulation in the past, but its high sensitivity to trained data, overlearning (overfitting) problems and its dependence on hidden neurons are significant drawbacks (Maier & Dandy 2000; Hsu *et al.* 2002; Coppola *et al.* 2003; Rao *et al.* 2003; Zaheer & Bai 2003; Affandi & Watanabe 2007; Nourani *et al.* 2008; Tsanis *et al.* 2008; Banerjee *et al.* 2009; Mohanty *et al.* 2010; Adamowski & Chan 2011; Trichakis *et al.* 2011; Wu & Chau 2011; Maheswaran & Khosa 2013; Mohanty *et al.* 2013; Shiri *et al.* 2013; Emamgholizadeh *et al.* 2014; Jha & Sahoo 2014). ANFIS was successfully employed to predict and simulate groundwater levels by Affandi & Watanabe

doi: 10.2166/ws.2019.204

(2007), Shiri & Kisi (2011), Moosavi *et al.* (2013), Shiri *et al.* (2013), and Emamgholizadeh *et al.* (2014). Fewer applications of GP for groundwater-level prediction and simulation are known, and the studies reported by Shiri & Kisi (2011), Shiri *et al.* (2013), Fallah-Mehdipour *et al.* (2013), and Fallah-Mehdipour *et al.* (2014) appear to be the only ones published to date. Fallah-Mehdipour *et al.* (2013) predicted and simulated the groundwater level of the Karaj aquifer, Iran, with GP and ANFIS. Fallah-Mehdipour *et al.* (2014) applied GP as a groundwater modeling tool in the Karaj aquifer. They generated 100 random combinations of aquifer recharge and discharge followed by numerical simulation of groundwater level of the cells in the groundwater network. The aquifer recharge and discharge calculated with the numerical model were then input to the GP method. Their results demonstrated there is no need to simulate the entire aquifer to obtain the recharge and discharge in each cell for changed conditions in the aquifer because the trained (i.e., calibrated) GP has the capacity to estimate the groundwater level at any desirable location in the aquifer.

The SVR method introduced by Vapnik (1995) has been applied to predict the groundwater level. Jin *et al.* (2009) used a least square support vector regression (LS–SVR) based on chaos dynamic and a radial basic function (RBF) as kernel to predict groundwater levels. Behzad *et al.* (2010) predicted the groundwater level under different pumping and weather conditions applying ANN and SVR methods. The RBF kernel is used in their study. The SVR method parameters were identified with the grid search method. Their results indicated that the SVR method is more accurate and general in comparison to ANN methods, especially when the data are insufficient. Yoon *et al.* (2011) compared ANN and SVR methods in groundwater-level prediction of two wells at a coastal aquifer in Korea. Their results showed the average resulted error in the SVR method was lower than that in the ANN method, that the SVR method was superior in learning complex relations between data and eliminating data noise, and that the model-building process should be carefully conducted, especially when using ANN models for groundwater-level forecasting in a coastal aquifer. Shiri *et al.* (2013) made a comparison between the ANN, ANFIS, GP, SVR, and auto regressive moving average methods in groundwater-level prediction. Their results showed GPs had the best predictive skill for groundwater levels up to 7 days beyond the recorded data. Gong *et al.* (2016) evaluated validity of three nonlinear time-series intelligence models – adaptive neuro fuzzy inference system, support vector machines, and artificial neural networks – in the prediction of the groundwater level. They developed and applied these three models for two wells close to Lake Okeechobee in Florida. The latter authors employed data sets of temperature, precipitation, groundwater level, and others as input data to forecast groundwater levels. They also calculated five quantitative standard statistical evaluation measures, correlation coefficient, normalized mean square error, root mean square error (*RMSE*), Nash–Sutcliffe efficiency coefficient, and the Akaike information criteria to evaluate the performances of these models. Their results established the SVM and ANFIS models' predictions were more accurate than the ANN model. Zhou *et al.* (2017) applied a data-base prediction model combining support vector machine and discrete wavelet transform preprocess for groundwater-level forecasting. They applied regular SVM and regular ANN, and wavelet preprocessed ANN models to monthly groundwater-level records over a period of 37 years from ten wells in Mengcheng County, China. The latter authors' results indicate that wavelet preprocess improved the training and test performance of the SVM and ANN models. The SVM model provided the most accurate and reliable groundwater-level prediction compared to the SVM, ANN, and SVM models, and the prediction results of the SVM model were superior to the ANN model's in generalization ability and precision.

A literature review shows an increasing trend in the use of data-driven methods in groundwater prediction and simulation. The ANN and ANFIS models have been more frequently applied than GP and SVR models, partly explained by the more recent origin of the latter two methods. The application of SVR to groundwater prediction and simulation has been hindered by the reliance on a linear objective function (Jin *et al.* 2009) and by the lack of a systematic approach to tune its parameters (Behzad *et al.* 2010; Yoon *et al.* 2011; Shiri *et al.* 2013; Su *et al.* 2013). A proper combination of the SVR parameters is necessary to achieve maximum predictive accuracy. The LS–SVR method solves a linear objective function instead of a quadratic one without any regularization, thus hindering its predictive accuracy (Sujay Raghavendra

& Deka 2014). Meta-heuristic optimization algorithms have overcome the parameter-tuning challenges (Sujay Raghavendra & Deka 2014), as demonstrated by several water-resources studies (Bozorg-Haddad *et al.* 2008; Noory *et al.* 2012; Sabbaghpour *et al.* 2012; Bozorg-Haddad *et al.* 2014b). Meta-heuristic algorithms such as the GA can optimize the SVR parameters, and the calibrated SVR is then applied to prediction problems. The coupled GA–SVR methodology offers flexibility in parameter estimation and a robust predictive algorithm.

This paper introduces a coupled GA–SVR algorithm and tests its accuracy in predicting groundwater levels. This paper's approach selects the training (i.e., calibration) and testing groundwater-level data chronologically and randomly. The GA is applied to optimize the SVR parameters and the parameter-optimized SVR is implemented for groundwater-level prediction. The prediction results obtained with the GA–SVR and the GP methods are compared to assess their relative merits.

THE SVR METHOD

The SVR method was introduced by Vapnik (1995) and has been applied to solve regression problems and to predict time series in a wide range of fields. The main idea behind the SVR method is to first select a nonlinear mapping algorithm, called the support vector kernel function, through which the input vectors are mapped onto a high-dimensional feature space and related to the output vectors. This method skillfully solves multi-dimensional prediction problems with generality. The SVR method steps are detailed next:

The function f represents the nonlinear relation between inputs (x) and outputs (y) of an arbitrary process. Equation (1) shows the function f:

$$y = f(x, \omega) = \omega \cdot \phi(x) + b \tag{1}$$

where x = input vector, where x belongs to an n-dimensional space ($x \in \Re^n$, x is a vector $n \times 1$); y = output ($y \in R$), [y a real scalar]; ω = the regression weight vector [ω: $1 \times n$]; b = model bias (scalar), $\phi(x)$ = nonlinear mapping. b and ω are calculated through with Equations (2) and (3). ϕ maps the nonlinear regression between inputs and outputs onto a high-dimensional space whereby a simpler regression is achieved that replaces the complex nonlinear regression of the original input space.

The kernel function K is needed in the mapping process with the equation $K(x, x') = [\phi(x)' \cdot \phi(x)]$, where x' is the transpose of x. This mapping replaces x in the original space by $\phi(x)$. It is not necessary to know the explicit expression of the nonlinear mapping ϕ throughout the solution processes. The kernel is appropriately chosen iteratively by the SVR method.

The selection of the kernel function *K*

There is no general principle to obtain the kernel function K (Li *et al.* 2010). The linear, polynomial, sigmoid, and the RBF are widely applied kernel functions in the SVR method. Previous findings indicate the best choice of kernel is problem-specific. The RBF has had the largest number of reported successful applications among the competing choices. This function has several advantages such as the low number of parameters compared with other kernel functions (Su *et al.* 2013). The RBF kernel function was applied in this work due to the large effect of the number of kernel function parameters on the SVR method's complexity. The RBF is given by Equation (4):

$$K(x, x_t) = e^{(-\gamma \| x - x_t \|^2)} \tag{2}$$

where $\| x - x_t \|^2$ = the Euclidean norm of x–x_t; x_t = the tth input vector, γ = the RBF parameter.

OPTIMIZATION OF THE SVR PARAMETERS WITH THE GA

The SVR method accuracy is largely dependent on the tuning SVR parameters (C, γ and ε). The standard SVR method does not specify an algorithm for selecting its parameters. Meta-heuristic optimization algorithms have been used for this purpose successfully in recent studies (Bozorg-Haddad *et al.* 2011; Su *et al.* 2013; Bozorg-Haddad *et al.* 2014a, 2014b). This work employs the GA for selecting the optimal values of the SVR method.

The GA is a heuristic search method inspired by evolutionary processes observed in nature, and is a widely employed algorithm for solving optimization problems in many fields of research (Chiu & Chen 2009). The GA poses an optimization problem as the selection of sequences of populations whose members are potential solutions of the problem at hand. The fitness of the members of generated populations in the GA improves through its iterations by the application of selection rules that mimic the principle of survival of the fitness of evolutionary theory. The best (fittest) members of a current population are selected for reproduction to generate a new population by means of crossover and mutation rules. Population members evolve from generation to generation. The GA ends after a large number of user-specified iterations or according to other user-specified termination criteria, at which time the remaining members of the last population are optimal in the sense that they constitute a solution that is very near the global optimal solution of the problem being solved. The GA is employed in this study to obtain optimal parameters of the SVR method. The flowchart of the coupled GA–SVR method is depicted in Figure 1.

The content of the left-dotted box in Figure 1 concerns the preprocessing of data. The SVR method with chosen kernel function is trained after identifying the train and test data set. The contents of the right-dotted box in Figure 1 describe the procedure of determining the optimal parameters (C, γ, and ε). The iteration number in the optimization process starts from zero ($i = 0$), at which time the initial population is randomly generated. The SVR method is trained (calibrated), employing the population of its members (potential solutions), and the objective function is calculated for each individual member to evaluate their fitness. Population, elitism, crossover rate, mutation rate, and the number of iterations of the GA were set to 10, 1, 0.8, 0.02, and 100 respectively. The *RMSE* was herein employed as the objective function to be minimized in the search for the SVR optimal parameters. The training data set was applied to calculate the objective function. The splitting of the data into training and testing data sets is discussed later. The *RMSE* and regression R^2 criteria were implemented to assess the error and accuracy of the hybrid GA–SVR method. Their *RMSE* and R^2 formulas

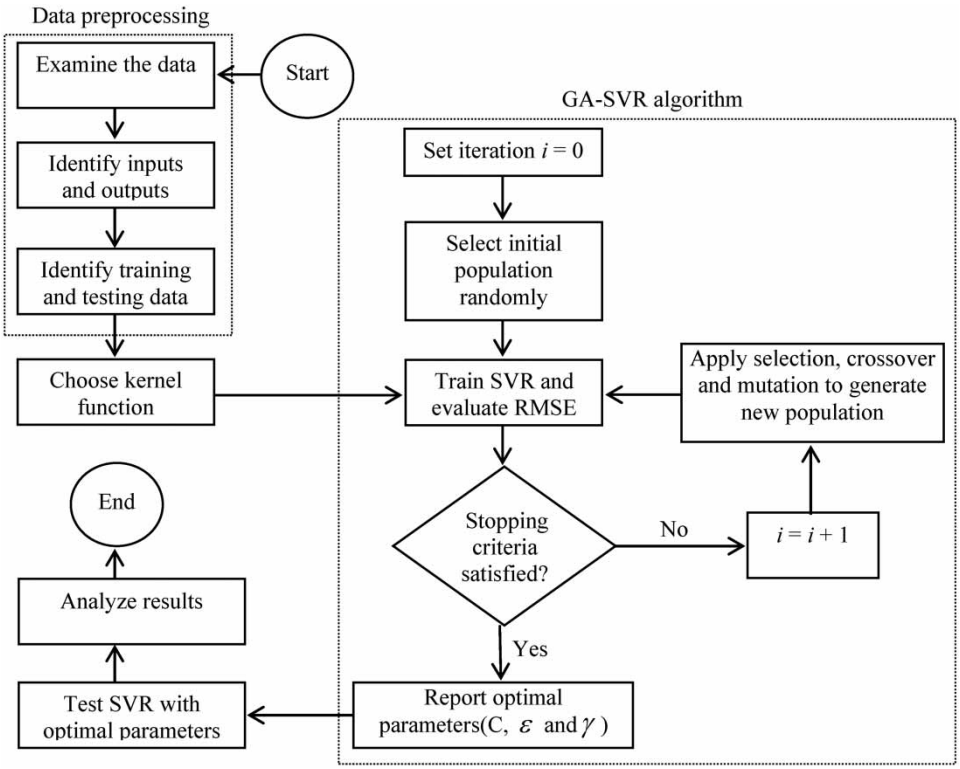

Figure 1 | The flowchart of the GA–SVR method.

are given respectively by Equations (5) and (6):

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(y_t - \bar{y})^2(\hat{y}_t - \hat{\bar{y}})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2 \times \sum_{t=1}^{T}(\hat{y}_t - \hat{\bar{y}})^2} \tag{4}$$

in which $\hat{y}_t$ = the estimated (calculated) data by the SVR method, $\hat{\bar{y}}$ = the average of estimated data and $\bar{y}$ = the average of the observation data, and T = the number of data values. The GA iterates with the current population of potential solutions until finding optimized values of the SVR parameters C, ε, and γ, which are those that occur whenever the *RMSE* objective function is minimized. The selection, crossover and mutation operators of the GA are then applied to generate a new population of solutions after finding the SVR optimal parameters. The SVR method is trained (calibrated) with the new population. This process is repeated until finding the SVR parameters associated with the optimal population of solutions. At this juncture the SVR method is optimally trained (calibrated) and it is then ready for testing with the testing data set.

THE CASE STUDY: THE KARAJ PLAIN

This paper's study region is located in the northwestern Karaj plain in Iran. The drinking water for Karaj city is predominantly groundwater. The aquifer storage is recharged by precipitation and recharge wells. Three observation wells were selected in the current study as the indicator of groundwater-level fluctuations in the region. The Karaj plain region depicting the location of the observation wells is shown in Figure 2.

The Shah-Abbasi, Tarbiat-Moallem, and Mehr-Shahr wells are labeled as 1, 2, and 3, respectively, in Figure 2. The monthly precipitation (P_t) and evaporation (EV_t) were selected as the surface parameters most strongly affecting the groundwater level. The 2002–2008 statistical period was used in this study (84 months total). Two prediction and simulation models were defined using the various input sets. The model functions are presented in Equations (7)–(10). For ease of referencing the models are numbered as follows:

$$h_t^j = f_1^j(h_{t-1}^1,\ h_{t-1}^2,\ h_{t-1}^3) \quad j = 1,\ 2,\ 3 \quad \text{Model 1} \tag{5}$$

$$h_t^j = f_2^j(h_{t-1}^1,\ h_{t-1}^2,\ h_{t-1}^3,\ P_{t-1},\ EV_{t-1}) \quad j = 1,\ 2,\ 3 \quad \text{Model 2} \tag{6}$$

$$h_t^j = \begin{cases} f_3^1(h_{t-1}^1,\ h_{t-1}^2,\ h_{t-1}^3,\ h_t^2,\ h_t^3) & j = 1 \\ f_3^2(h_{t-1}^1,\ h_{t-1}^2,\ h_{t-1}^3,\ h_t^1,\ h_t^3) & j = 2 \\ f_3^3(h_{t-1}^1,\ h_{t-1}^2,\ h_{t-1}^3,\ h_t^1,\ h_t^2) & j = 3 \end{cases} \quad \text{Model 3} \tag{7}$$

$$h_t^j = \begin{cases} f_4^1(h_{t-1}^1, h_{t-1}^2, h_{t-1}^3, h_t^2, h_t^3, P_{t-1}, P_t, EV_{t-1}, EV_t) & j = 1 \\ f_4^2(h_{t-1}^1, h_{t-1}^2, h_{t-1}^3, h_t^1, h_t^3, P_{t-1}, P_t, EV_{t-1}, EV_t) & j = 2 \\ f_4^3(h_{t-1}^1, h_{t-1}^2, h_{t-1}^3, h_t^1, h_t^2, P_{t-1}, P_t, EV_{t-1}, EV_t) & j = 3 \end{cases} \quad \text{Model 4} \tag{8}$$

where h_t^j = groundwater level of the jth well at time step t; j = number of observation wells in which Shah-Abbasi ($j = 1$), Tarbiat-Moallem ($j = 2$), and Mehr-Shahr ($j = 3$); t = monthly time step; h_{t-1}^1 = groundwater level of well number 1 at time step of $t-1$; h_{t-1}^2 = groundwater level of well number 2 at time step $t-1$; h_{t-1}^3 = groundwater level of well number 3 at time step $t-1$; P_{t-1} = precipitation depth at time step $t-1$; P_t = precipitation depth at time step t; EV_{t-1} = evaporation depth at time step $t-1$; EV_t = evaporation depth at time step t; f_1^j = prediction model for the jth well with groundwater data (Model 1); f_2^j = prediction model for the jth well with groundwater and surface data (Model 2); f_3^j = simulation model for the jth well with groundwater data (Model 3); f_4^j = simulation model for the jth well with groundwater and surface data (Model 4).

RESULTS AND DISCUSSION

This work applied three different approaches to select the training and testing data sets of the GA–SVR method. In the first approach all the available data in the statistical period (84 months total) were employed for training. The purpose of taking this approach is assessing the ability of the GA–SVR to reproduce groundwater-level fluctuations.

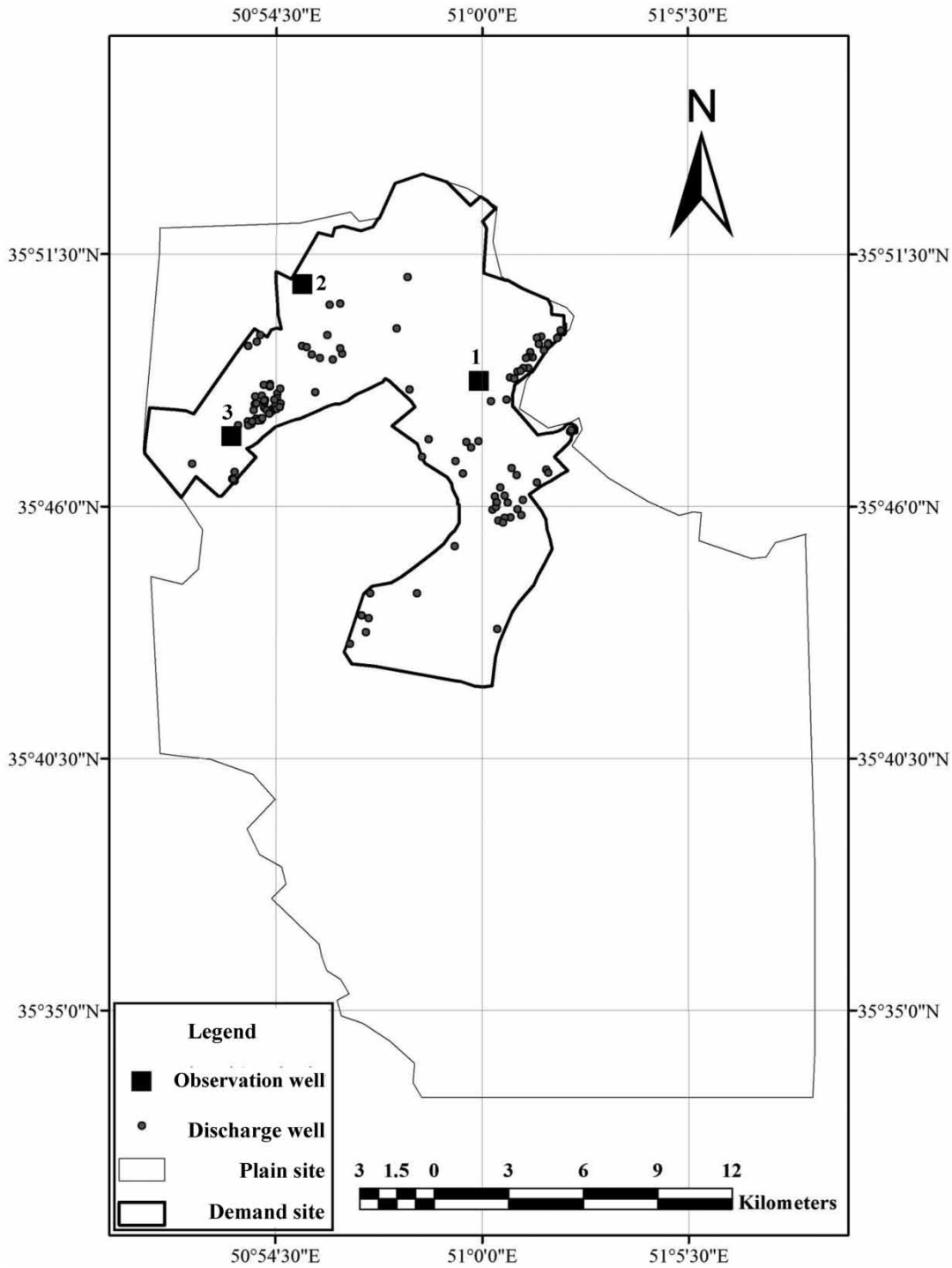

Figure 2 | The study area and well locations in the Karaj plain region.

The second approach divides the data into training and testing sets. The division in this approach was done chronologically. The first 6 years of the statistical period were selected for training the GA–SVR method and the last year was used to test the method. The aim of this division is assessing the ability of the GA–SVR method in estimating

groundwater level with non-trained data, i.e., testing the predictive accuracy of the trained SVR with a data set not used in its training. In the third approach, similar to the second, the data were divided into two sets: training and testing. The difference introduced in the third approach was the random selection of the data sets. The purpose of applying the third approach is increasing the probability of finding similarity in the correlation structure between the training and testing data sets.

Assessing the ability of the GA–SVR method to reproduce groundwater-level fluctuation

The prediction and simulation models consider all the available data (84 month) in the first approach. The GA–SVR method was trained with all available data to test its ability in reproducing groundwater levels. The GA–SVR calculated parameters and the error criteria are listed in Table 1.

The differences between the observation data and the calculated results from the best prediction and simulation models for each well are illustrated in Figure 3. Among prediction and simulation models, the model with the lowest value of the *RMSE* objective function was the best one. It is seen in Figure 3 and Table 1 that the GA–SVR method is able to fit the selected models to the groundwater-level data quite well. One noteworthy issue in Table 1 is that the prediction model of the Shah-Abbasi well increases its accuracy with the addition of surface data, which is not the case in the other wells.

Table 1 | Parameters and statistical criteria for GA–SVR with the first selection approach

Observation well	Model	C	ε	γ	*RMSE*(m)	R^2
Shah-Abbasi (well number 1)	1	5.272	0.038	7.130	0.045	0.9996
	2	6.100	0.014	0.433	0.014	0.9999
	3	3.761	0.012	2.419	0.052	0.9992
	4	3.785	0.011	2.866	0.019	0.9991
Tarbiat-Moallem (well number 2)	1	5.754	0.028	8.423	0.027	0.9994
	2	2.411	0.096	0.606	0.094	0.9965
	3	3.283	0.010	2.486	0.015	0.9997
	4	2.403	0.010	0.325	0.010	0.9999
Mehr-Shahr (well number 3)	1	3.148	0.026	1.768	0.064	0.9868
	2	2.906	0.090	0.375	0.084	0.9880
	3	2.235	0.010	1.944	0.011	0.9996
	4	2.344	0.042	1.176	0.040	0.9972

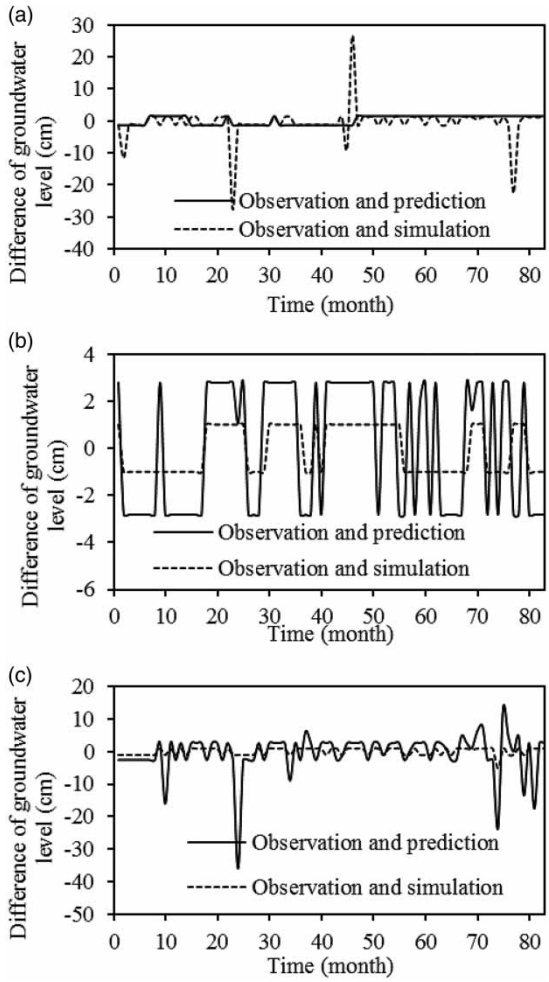

Figure 3 | The results of the best prediction and simulation model obtained by the GA–SVR method with the first selection approach, (a) Shah-Abbasi, (b) Tarbiat-Moallem, and (c) Mehr-Shahr wells.

The models that consider the surface variables (precipitation and evaporation) in prediction and simulation of the Shah-Abbasi well's groundwater level are 0.03 and 0.04 m more accurate than the models that only consider the subsurface factors. The reason may be found in the data correlation structure. The correlation structure of the observation wells' groundwater levels is listed in Table 2.

It is seen in Table 2 that the groundwater level of the Shah-Abbasi well is correlated with the groundwater level of the Shah-Abbasi well at time step $t-1$ (the previous time step), with the groundwater level of the Mehr-Shahr well at time steps of t and $t-1$, and with the evaporation at time step t at the 1% significant level. The groundwater

Table 2 | Correlation structure of data with the first selection approach

Variables	h_t^1	h_t^2	h_t^3
h_t^1	1.00	−0.05	0.63**
h_t^2	−0.05	1.00	0.28*
h_t^3	0.63**	0.28*	1.00
h_{t-1}^1	0.93**	−0.11	0.57**
h_{t-1}^2	−0.03	0.89**	0.20
h_{t-1}^3	0.64**	0.28**	0.93**
EV_t	0.34**	0.06	0.13
EV_{t-1}	0.21*	−0.06	−0.02
P_t	−0.22*	−0.01	−0.18
P_{t-1}	−0.12	0.03	−0.08

** and * mean 1% and 5% significant levels, respectively.

Table 3 | Parameters of the GA–SVR algorithm with the second selection approach

Observation well	Model	C	ε	γ
Shah-Abbasi (well number 1)	1	6.481	0.060	5.318
	2	10.200	0.100	10.100
	3	3.167	0.061	1.644
	4	100.200	0.010	10.848
Tarbiat-Moallem (well number 2)	1	2.643	0.035	3.363
	2	1.622	0.014	0.387
	3	2.761	0.009	2.632
	4	94.500	0.010	1.850
Mehr-Shahr (well number 3)	1	2.868	0.080	2.845
	2	20.031	0.010	5.953
	3	1.790	0.013	0.730
	4	15.410	0.018	2.29

level of the same well is correlated at the 5% significant level with evaporation at time step $t-1$ and precipitation at time step t. Thus, the prediction accuracy of the groundwater level of the Shah-Abbasi well is improved by considering surface factors. The other two wells models do not exhibit significant correlation with surface factors.

Chronological selection of the training and testing data

The GA–SVR method was trained with the training data set and the parameters (C, γ, and ε) were determined. The predictive skill of the GA–SVR method was tested using the optimized parameters and the testing data. The optimized parameter values are listed in Table 3.

The calculated error criteria are listed in Table 4. For the purpose of comparison, the results of Fallah-Mehdipour *et al.* (2013) are shown in Table 4. The contents of Table 4 show that the first and fourth models for the Shah-Abbasi well fit the data quite well while the results of Models 2 and 3 are not acceptable because of the correlation structure of the data. The correlation structure of the training and testing data is shown in Table 5. It is seen in Table 5 that the correlation structures of the training and testing data differ. The SVR is a data-based method. Therefore, the SVR's predictive accuracy drops dramatically if the structure of the testing data is

Table 4 | Statistical criteria of the GA–SVR algorithm and GP method with the second selection approach

		GA–SVR				GP			
		Training		Testing		Training		Testing	
Observation well	Model	*RMSE* (m)	R^2	*RMSE* (m)	R^2	*RMSE* (m)	R^2	*RMSE* (m)	R^2
Shah-Abbasi (well number 1)	1	0.060	0.9994	0.171	0.8601	0.71	0.86	0.18	0.89
	2	0.098	0.9993	0.691	0.3317	0.62	0.89	0.15	0.90
	3	0.163	0.9934	0.704	0.7749	0.73	0.85	0.18	0.88
	4	0.010	0.9999	0.097	0.9367	0.6	0.90	0.15	0.91
Tarbiat-Moallem (well number 2)	1	0.117	0.9858	0.533	0.7213	0.4	0.83	0.44	0.73
	2	0.025	0.9996	0.427	0.8926	0.4	0.83	0.41	0.78
	3	0.015	0.9998	0.622	0.6073	0.43	0.80	0.44	0.73
	4	0.010	0.9999	0.399	0.8272	0.42	0.81	0.43	0.77
Mehr-Shahr (well number 3)	1	0.073	0.9884	0.130	0.5332	0.21	0.87	0.16	0.48
	2	0.010	0.9998	0.263	0.4559	0.19	0.89	0.15	0.55
	3	0.053	0.9922	0.165	0.7093	0.21	0.87	0.16	0.48
	4	0.018	0.9995	0.166	0.5957	0.19	0.89	0.15	0.54

Table 5 | Correlation structure of data with the second selection approach

Variables	h_t^1	h_t^2	h_t^3
(a) Training			
h_t^1	1.00	0.01	0.59**
h_t^2	0.01	1.00	0.35**
h_t^3	0.59**	0.35**	1.00
h_{t-1}^1	0.92**	−0.04	0.52**
h_{t-1}^2	0.01	0.89**	0.25*
h_{t-1}^3	0.60**	0.36**	0.93**
EV_t	0.40**	0.14	0.17
EV_{t-1}	0.25*	0.03	−0.01
P_t	−0.23*	−0.06	−0.20
P_{t-1}	−0.12	−0.02	−0.08
(b) Testing			
h_t^1	1.00	0.84**	0.50
h_t^2	0.84**	1.00	0.61*
h_t^3	0.50	0.61*	1.00
h_{t-1}^1	0.93**	0.64*	0.40
h_{t-1}^2	0.74**	0.85**	0.38
h_{t-1}^3	0.72**	0.75**	0.69*
EV_t	−0.63*	−0.57*	−0.55
EV_{t-1}	−0.53	−0.74**	−0.49
P_t	0.38	0.54	0.48
P_{t-1}	0.38	0.64*	0.16

** and * mean 1% and 5% significant levels, respectively.

different to that of the data with which the GA–SVR method is trained. This problem was observed with Models 2 and 3.

According to Table 5 the training data set of the Shah-Abbasi well's groundwater level at time step t is correlated with the groundwater level of the Shah-Abbasi well at time step $t-1$, with the groundwater level of Mehr-Shahr well at time steps t and $t-1$, and with precipitation at time step t. The correlation structure of the testing data set of the Shah-Abbasi well reveals a correlation between the groundwater level of the Shah-Abbasi well at time step t with the groundwater level of Shah-Abbasi well at time step $t-1$, with the groundwater level of the Mehr-Shahr well at time step $t-1$, with the groundwater level of the Tarbiat-Moallem well at time step $t-1$, and with evaporation at time step t. The correlation structure in this case agrees with Model 1, which reveals a correlation between the groundwater level of the Shah-Abbasi well at time step t with the groundwater level of the Shah-Abbasi and the two other wells at time step $t-1$. Notice, however, that Model 3 reveals a correlation between the groundwater level of the Shah-Abbasi at time step t with the groundwater level of the other two wells at time steps t and $t-1$, and with the groundwater level of the Shah-Abbasi well at time step $t-1$. The SVR method with inputs and outputs of Model 3 shows a strong correlation between the groundwater level of the Shah-Abbasi well at time step t and the groundwater level of the Mehr-Shahr well at time step t during the training procedure, while this strong correlation does not exist in the testing data set. Such contrasts in the correlation structure hinder the ability of the SVR method to fit data and increase the predictive error of the models. The same kind of analysis can be performed for the other two wells, which highlights the impact of the correlation structure on the GA–SVR method's predictive accuracy.

The statistical period's groundwater-level data used for testing has a correlation structure different to that of the training data for the Tarbiat-Moallem well. Because of this, the accuracy of the GA–SVR method and GP for this well is lower than those of the other two observation wells. The average error of the simulation and prediction models for this well is about 0.4 m, while this amount is about 0.1 m for the other two wells.

The calculated results shown in Table 4 indicate that the GA–SVR method has superior predictive accuracy compared to GP. Applying the coupled GA–SVR method improved its predictive and simulation accuracies respectively by about 5% and 35% in comparison with the GP model for the Shah-Abbasi well. The predictive accuracy of the prediction models obtained with the GP algorithm for the Tarbiat-Moallem well was 4% better than the GA–SVR method. The accuracy of the GA–SVR method is about 7% higher than the GP's for the simulation models. The application of the GA–SVR method in the prediction models for the Mehr-Shahr well resulted in 18% better accuracy than that achieved with GP; however, the simulation model with GP exhibited a 9% improvement in comparison to the GA–SVR method.

Table 4 shows significant differences in the *RMSE* between the various models for each well. For the purpose of illustration, the accuracy of Models 1 and 2 for the

Shah-Abbasi well was 0.171 and 0.691, respectively. It is obvious that Model 1 achieved the better conformance with the correlation structure of the test data in this well according to Equations (7) and (8) and the correlation structure of the training and testing data presented in Table 5. The evaporation at the $t-1$ time step was included in Model 2. The training data for the groundwater level of the Shah-Abbasi well shows significant correlation with evaporation in the $t-1$ time step; however, this correlation is not seen in the testing data. In summary, the GA–SVR method is trained in a manner that specifies weights to link the evaporation at $t-1$ time step to the groundwater level in the Shah-Abbasi well, but these weights introduce errors in the GA–SVR method with Model 2 because of the dissimilar correlation structure of testing data. Similar analyses can be performed for the other models and wells. This means that the accuracy of the GA–SVR method in prediction and simulation of the groundwater level is higher than the GP's, but it is also more sensitive to the correlation structure of the data in comparison to GP. The results of GA–SVR method are presented in Figure 4.

Random selection of the training and testing data

The previous results established that the selection of the training and testing data set affects the GA–SVR method's performance. The recorded monthly time series data used in this work constitute an extensive hydrologic time series. The hydrologic and climatic indicators have a specific pattern of temporal fluctuations. The second and third approaches are the principal methods to select the training and testing data. The second approach selected the data chronologically. One of the advantages of that approach is capturing seasonal fluctuations of the time series. It is essential to notice that selecting the training and testing data sequentially does not mean that the correlation structure between these two data sets is well preserved. The third approach selects the training and testing data randomly. The specific time patterns in the original data series might not be preserved in this approach. However, the random selection preserves the correlation structure between the training and testing data better than the chronological selection of the same data sets. The dissimilarity of the training and testing data correlation structures was observed in the

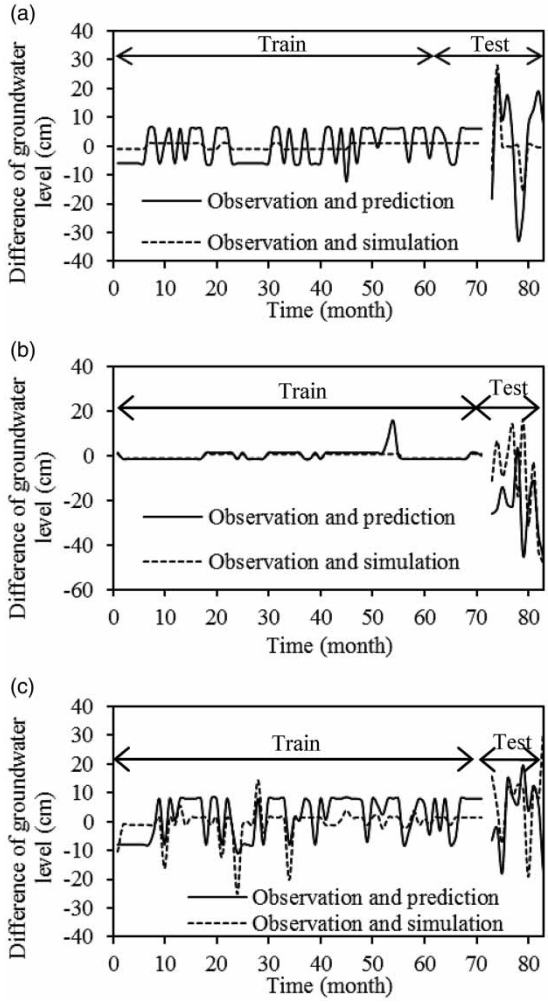

Figure 4 | The results of best prediction and simulation models obtained by the GA–SVR method with the second selection approach, (a) Shah-Abbasi, (b) Tarbiat-Moallem, and (c) Mehr-Shahr wells.

second selection approach in this study. The third approach was implemented to evaluate the effect of the training and test data selection on prediction accuracy. The correlation structure of the randomly selected training and testing data sets is listed in Table 6.

Comparison of the results listed in Table 6 with those of Table 2 indicates that the correlation structure between the training and testing data sets conforms well with the correlation structure of the total data set, something not observed with the second approach wherein the two data sets (training and testing) where chosen chronologically. It is also evident in Table 6 that there is similarity in the correlation structure between the training and testing data sets.

Table 6 | Correlation structure of data with the third selection approach

Variables	h_t^1	h_t^2	h_t^3
(a) Training			
h_t^1	1.00	−0.05	0.62**
h_t^2	−0.05	1.00	0.27*
h_t^3	0.62**	0.27*	1.00
h_{t-1}^1	0.93**	−0.13	0.53**
h_{t-1}^2	−0.04	0.89**	0.18
h_{t-1}^3	0.63**	0.28*	0.92**
EV_t	0.24*	0.09	0.07
EV_{t-1}	0.15	−0.07	−0.08
P_t	−0.18	0.01	−0.21
P_{t-1}	−0.03	0.05	0.03
(b) Testing			
h_t^1	1.00	−0.12	0.64*
h_t^2	−0.12	1.00	0.34
h_t^3	0.64*	0.34	1.00
h_{t-1}^1	0.94**	−0.06	0.69*
h_{t-1}^2	−0.03	0.79**	0.34
h_{t-1}^3	0.67*	0.35	0.97**
EV_t	0.81**	−0.16	0.45
EV_{t-1}	0.51	−0.00	0.21
P_t	−0.51	−0.22	−0.15
P_{t-1}	−0.39	−0.015	−0.44

** and * mean 1% and 5%, significant levels, respectively.

Recall from Table 5 that there is no clear similarity between the training and testing data in the second approach, and that differences in correlation structure were observed between the first and second approaches. These results suggest that the third approach, i.e., random selection of the training and testing data sets, is more appropriate than the second approach of chronological selection of the data sets.

The GA–SVR method was run to evaluate the third approach. The calculated results are presented in Table 7. Comparing Tables 4 and 7 highlights the effects of random and chronological selection of the training and testing data sets. For the purpose of illustration, the error of Model 2 for the Shah-Abbasi well was 0.691 with the chronological (second) approach. This error's value equaled 0.191 with the randomized (third) approach, which amounts to a 72% improvement in predictive accuracy. The similar *RMSE* values for the training and testing data sets of the various

Table 7 | Statistical criterion of the GA–SVR method with the third selection approach

		Training		Testing	
Observation well	Model	*RMSE* (m)	R^2	*RMSE* (m)	R^2
Shah-Abbasi (well number 1)	1	0.053	0.9992	0.181	0.9950
	2	0.057	0.9991	0.191	0.9926
	3	0.042	0.9996	0.184	0.9930
	4	0.010	0.9999	0.188	0.9943
Tarbiat-Moallem (well number 2)	1	0.122	0.9853	0.285	0.9279
	2	0.077	0.9952	0.251	0.9081
	3	0.041	0.9983	0.258	0.8876
	4	0.039	0.9992	0.236	0.9044
Mehr-Shahr (well number 3)	1	0.038	0.9963	0.141	0.9663
	2	0.048	0.9919	0.140	0.9729
	3	0.029	0.9976	0.131	0.9834
	4	0.029	0.9984	0.144	0.9683

models and their low errors indicate that the randomized selection of the training and testing data set overcomes the overtraining problem of the GA–SVR method. For the purpose of illustration, notice that the predictive accuracy of Model 4 with randomized selection of the training data was 40% better than that achieved with chronological selection of the training and testing data sets. The R^2 values ranged between 0.993–0.995, 0.888–0.9279, and 0.9663–0.9834 for wells 1, 2, and 3, respectively, in testing groundwater levels, which are significantly larger than those obtained with the GP model.

CONCLUSION

This paper introduced the coupled GA–SVR method to the prediction and simulation of groundwater levels. The GA-SVR was tested with a data set from the Karaj aquifer in Iran. Selection of the training and testing data set affects the GA–SVR method's performance. Therefore, three approaches for selecting the calibration and testing data sets were considered: (1) all of the available data were employed for training; (2) divide the data into training and testing sets chronologically; and (3) divide the data into training and testing sets randomly. This paper's results have established that the GA–SVR method exhibits accurate prediction skill when its parameters are optimized and the training and testing data sets are chosen randomly. In addition, the results demonstrate the GA–SVR algorithm's

accuracy in predicting and simulating groundwater levels with the three cited approaches, and establishes that the predictive accuracy of randomized selection of the training data was 40% better than that achieved with chronological selection of the training and testing data sets. Moreover, the R^2 values ranged between 0.993–0.995, 0.888–0.928, and 0.966–0.983 for wells 1, 2, and 3, respectively, in testing the groundwater level with the randomized approach, which implies a significant improvement of GA–SVR results compared to those obtained with the GP.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGEMENT

The authors thank Iran's National Science Foundation for its financial support of this research.

REFERENCES

Adamowski, J. & Chan, H. F. 2011 A wavelet neural network conjunction model for groundwater level forecasting. *Journal of Hydrology* **407** (1), 28–40.

Affandi, A. K. & Watanabe, K. 2007 Daily groundwater level fluctuation forecasting using soft computing technique. *Nature and Science* **5** (2), 1–10.

Banerjee, P., Prasad, R. K. & Singh, V. S. 2009 Forecasting of groundwater level in hard rock region using artificial neural network. *Environmental Geology* **58** (6), 1239–1246.

Behzad, M., Asghari, K. & Coppola, E. A. 2010 Comparative study of SVMs and ANNs in aquifer water level prediction. *Journal of Computing Civil Engineering* **24** (5), 408–413.

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

Bozorg-Haddad, O., Adams, B. J. & Mariño, M. A. 2008 Optimum rehabilitation strategy of water distribution systems using the HBMO algorithm. *Journal of Water Supply: Research and Technology – AQUA* **57** (5), 327–350.

Bozorg-Haddad, O., Afshar, A. & Mariño, M. A. 2011 Multireservoir optimisation in discrete and continuous domains. *Proceedings of the Institution of Civil Engineers: Water Management* **164** (2), 57–72.

Bozorg-Haddad, O., Fallah-Mehdipour, E., Mirzaei-Nodoushan, F. & Mariño, M. A. 2014a Discussion of 'A GA-based support vector machine model for the prediction of monthly reservoir storage'. *Journal of Hydrologic Engineering* **20** (2), 07014009.

Bozorg-Haddad, O., Moravej, M. & Loáiciga, H. 2014b Application of the water cycle algorithm to the optimal operation of reservoir systems. *Journal of Irrigation and Drainage Engineering*. doi:10.1061/(ASCE)IR.1943-4774.0000832,04014064.

Chiu, D. Y. & Chen, P. J. 2009 Dynamically exploring internal mechanism of stock market by fuzzy-based support vector machines with high dimension input space and genetic algorithm. *Expert System with Applications* **36** (2), 1240–1248.

Coppola, E., Poulton, M., Charles, E., Dustman, J. & Szidarovszky, F. 2003 Application of artificial neural networks to complex groundwater management problem. *Natural Resources Research* **12**, 303–320.

Emamgholizadeh, S., Moslemi, K. & Karami, G. 2014 Prediction the groundwater level of bastam plain (Iran) by artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS). *Water Resources Management* 1–14. doi:10.1007/s11269-014-0810-0.

Fallah-Mehdipour, E., Bozorg-Haddad, O. & Mariño, M. A. 2013 Prediction and simulation of monthly groundwater levels by genetic programming. *Journal of Hydro-Environment Research* **7** (4), 253–260.

Fallah-Mehdipour, E., Bozorg-Haddad, O. & Mariño, M. A. 2014 Genetic programming in groundwater modeling. *Journal of Hydrologic Engineering*. doi:10.1061/(ASCE)HE.1943-5584.0000987.

Gong, Y., Zhang, Y., Lan, S. & Wang, H. 2016 A comparative study of artificial neural networks, support vector machines and adaptive neuro fuzzy inference system for forecasting groundwater levels near Lake Okeechobee, Florida. *Water Resources Management* **30** (1), 375–391.

Hsu, K. L., Gupta, H. V., Gao, X., Sorooshian, S. & Imam, B. 2002 Self-organizing linear output map (SOLO): an artificial neural network suitable for hydrologic modeling and analysis. *Water Resources Research* **38** (12), 1–17.

Jha, M. K. & Sahoo, S. 2014 Efficacy of neural network and genetic algorithm techniques in simulating spatio-temporal fluctuations of groundwater. *Hydrological Processes*. doi:10.1002/hyp.10166.

Jin, L., Chang, J. X. & Zhang, W. G. 2009 Groundwater level dynamic prediction based on chaos optimization and support vector machine. In: *Proceedings of the 3rd International Conference on Genetic and Evolutionary Computing*. IEEE Transaction, Zhengzhou, China. http://dx.doi.org/10.1109/WGEC.2009.25.

Li, C. H., Lin, C. T., Kuo, B. C. & Chu, H. S. 2010 An automatic method for selecting the parameter of the RBF kernel functions to support vector machines. In: *Geoscience Remote Sensing Symposium (IGARSS)*. IEEE Int., IEEE, Piscataway, NJ, pp. 836–839.

Maheswaran, R. & Khosa, R. 2013 Long term forecasting of groundwater levels with evidence of non-stationary and nonlinear characteristics. *Computers & Geosciences* **52**, 422–436.

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. *Environmental Modelling and Software* **15** (1), 101–124.

Mohanty, S., Jha, M. K., Kumar, A. & Sudheer, K. P. 2010 Artificial neural network modeling for groundwater level forecasting in a River Island of Eastern India. *Water Resources Management* **24** (9), 1845–1865.

Mohanty, S., Jha, M. K., Kumar, A. & Panda, D. K. 2013 Comparative evaluation of numerical model and artificial neural network for simulating groundwater flow in Kathajodi–Surua Inter-basin of Odisha, India. *Journal of Hydrology* **495**, 38–51.

Moosavi, V., Vafakhah, M., Shirmohammadi, B. & Behnia, N. 2013 A wavelet-ANFIS hybrid model for groundwater level forecasting for different prediction periods. *Water Resources Management* **27** (5), 1301–1321.

Noory, H., Liaghat, A. M., Parsinejad, M. & Bozorg-Haddad, O. 2012 Optimizing irrigation water allocation and multicrop planning using discrete PSO algorithm. *Journal of Irrigation and Drainage Engineering* **138** (5), 437–444.

Nourani, V., Mogaddam, A. A. & Nadiri, A. O. 2008 An ANN-based model for spatiotemporal groundwater level forecasting. *Hydrological Processes* **22** (26), 5054–5066.

Rao, S. V. N., Thandaveswara, B. S., Bhallamudi, M. S. & Srinivasulu, V. 2003 Optimal groundwater management in deltaic regions using simulated annealing and neural networks. *Water Resources Management* **17**, 409–428.

Sabbaghpour, S., Naghashzadehgan, M., Javaherdeh, K. & Bozorg-Haddad, O. 2012 HBMO algorithm for calibrating water distribution network of Langarud city. *Water Science and Technology* **65** (9), 1564–1569.

Shiri, J. & Kişi, Ö. 2011 Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. *Computers & Geosciences* **37** (10), 1692–1701.

Shiri, J., Kisi, O., Yoon, H., Lee, K. K. & Hossein Nazemi, A. 2013 Predicting groundwater level fluctuations with meteorological effect implications – a comparative study among soft computing techniques. *Computers & Geosciences* **56**, 32–44.

Su, J., Wang, X., Liang, Y. & Chen, B. 2013 A GA-based support vector machine model for the prediction of monthly reservoir storage. *Journal of Hydrologic Engineering*. doi:10.1061/(ASCE)HE.1943-5584.0000915.

Sujay Raghavendra, N. & Deka, P. C. 2014 Support vector machine applications in the field of hydrology: a review. *Applied Soft Computing* **19**, 372–386. http://dx.doi.org/10.1016/j.asoc.2014.02.002.

Trichakis, I. C., Nikolos, I. K. & Karatzas, G. P. 2011 Artificial neural network (ANN) based modeling for karstic groundwater level simulation. *Water Resources Management* **25** (4), 1143–1152.

Tsanis, I., Coulibaly, P. & Daliakopoulos, I. 2008 Improving groundwater level forecasting with a feed forward neural network and linearly regressed projected precipitation. *Journal of Hydroinformatics* **10** (4), 317–330.

Vapnik, V. N. 1995 *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, USA.

Wu, C. L. & Chau, K. W. 2011 Rainfall–runoff modeling using artificial neural network coupled with singular spectrum analysis. *Journal of Hydrology* **399** (3–4), 394–409.

Yoon, H., Jun, S. C., Hyun, Y., Bae, G. O. & Lee, K. K. 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.

Zaheer, I. & Bai, C. G. 2003 Application of artificial neural network for water quality management. *Lowland Technology International* **5** (2), 10–15.

Zhou, T., Wang, F. & Yang, Z. 2017 Comparative analysis of ANN and SVM models combined with wavelet preprocess for groundwater depth prediction. *Water* **9** (10), 781.

First received 16 August 2019; accepted in revised form 16 December 2019. Available online 31 December 2019