Simulation of Methyl Tertiary Butyl Ether Concentrations in River-Reservoir Systems Using Support Vector Regression

Mahyar Aboutalebi, M.ASCE[1]; Omid Bozorg-Haddad[2]; and Hugo A. Loáiciga, F.ASCE[3]

Abstract: Mathematical and numerical models are used to simulate the transport of pollutants released into a water body. Such simulations can be computationally burdensome, however. One approach to overcome computational burdens associated with the simulation of pollutant transport is to use data-mining tools. The aim of this study is to simulate the concentration of methyl tertiary butyl ether (MTBE) at various locations within a river-reservoir system using the support vector regression (SVR) data-mining tool. The SVR tool is optimized by means of a genetic algorithm (GA). This paper's results indicate that the developed and optimized SVR tool is more accurate than artificial neural networks (ANN) and genetic programming (GP) when judged by the correlation coefficient of regression analysis (R^2). **DOI: 10.1061/(ASCE)IR.1943-4774.0001007.** © *2016 American Society of Civil Engineers.*

Author keywords: Simulation; Data mining; Support vector regression; Genetic algorithm.

Introduction

Many recent publications on water resources have dealt with topics such as reservoir operation (Ahmadi et al. 2014; Bolouri-Yazdeli et al. 2014; Ashofteh et al. 2013a, 2015a), groundwater resources (Bozorg-Haddad et al. 2013; Fallah-Mehdipour et al. 2013b), conjunctive use operation (Fallah-Mehdipour et al. 2013a), design-operation of pumped-storage and hydropower systems (Bozorg-Haddad et al. 2014), flood management (Bozorg-Haddad et al. 2015b), water project management (Orouji et al. 2014; Shokri et al. 2014), hydrology (Ashofteh et al. 2013b), qualitative management of water resources systems, (Orouji et al. 2013; Bozorg-Haddad et al. 2015a), water distribution systems (Seifollahi-Aghmiuni et al. 2013; Soltanjalili et al. 2013; Beygi et al. 2014), agricultural crops (Ashofteh et al. 2015c), sedimentation (Shokri et al. 2013), and algorithmic developments (Ashofteh et al. 2015b). Yet, very few water-resources publications have dealt with the simulations of contaminants [methyl tertiary butyl ether (MTBE) is a case in point] in river-reservoir systems using support vector machine (SVM).

The sudden release of toxic chemicals into water bodies poses unique challenges for emergency response due to their unexpected occurrence and rapid transport in water bodies (Hou et al. 2014). There are water-quality models that can be used to simulate the transport of pollutants released to water bodies, which is one way to respond to sudden pollutant releases using real-time approaches. Yet, those models may be computationally burdensome. One approach to reduce the computational burden associated with the simulation of pollutant transport in water resources systems is by resorting to data-mining tools. Data-mining tools have been used in several studies involving the simulation of hydrologic phenomena. One powerful data-mining tool, however, the support vector machine (SVM), has received increasing use in the simulation of quantity and quality phenomena related to water resources.

Concerning the application of data-mining tools to quantify phenomena, Savic et al. (1999) used genetic programming (GP) to model runoff in the Kirkton basin in Scotland. The latter authors compared the performance of GP with that of artificial neural networks (ANN). Their results showed better accuracy for GP than ANN. Asefa et al. (2006) used SVM to predict the hourly and seasonal inflow in the Sevier River basin, in Utah in the United States. Sivapragasam et al. (2007) evaluated the accuracy of inflow prediction in the operation of the Kovilar and Priyar reservoirs in India to supply agricultural water. They used GP to predict inflows. Behzad et al. (2009) evaluated the performance of ANN and SVM in prediction of the Bakhtiari River runoff in Iran and showed that SVM is more accurate in predicting runoff than ANN and ANN-GP methods. Wang et al. (2009) compared the performance of different methods of monthly inflow prediction in two rivers in China. They applied the autoregressive moving average (ARMA), ANN, adaptive neural-based fuzzy inference system (ANFIS), GP, and SVM. Their results indicated that GP, ANFIS, and SVM had better results (that is, smaller errors) than several other methods. Yoon et al. (2011) applied ANN and SVM models to predict the groundwater level in coastal aquifers in Korea. Their results showed better performance for SVM than ANN. Wei (2012) coupled kernel wavelet function with SVM to predict the water level in a measuring station of the Tanshui River in China. The wavelet SVM performed better in predictions than the SVM coupled with a Gaussian kernel. Maity et al. (2013) applied the SVM and auto-regressive integrated moving average (ARIMA) method to predict monthly river inflow in the Mahanadi River in India. Their results indicated that SVM had better predictive accuracy than ARIMA. More comprehensive reviews and comments on application of SVM are

[1]M.Sc. Graduate, Dept. of Irrigation and Reclamation Engineering, Faculty of Agricultural Engineering and Technology, College of Agriculture and Natural Resources, Univ. of Tehran, Karaj, 1437835693 Tehran, Iran. E-mail: Aboutalebi@ut.ac.ir

[2]Associate Professor, Dept. of Irrigation and Reclamation Engineering, Faculty of Agricultural Engineering and Technology, College of Agriculture and Natural Resources, Univ. of Tehran, Karaj, 3158777871 Tehran, Iran (corresponding author). E-mail: OBHaddad@ut.ac.ir

[3]Professor, Dept. of Geography, Univ. of California, Santa Barbara, CA 93106. E-mail: Hugo.Loaiciga@ucsb.edu

Note. This manuscript was submitted on March 12, 2015; approved on November 19, 2015; published online on March 16, 2016. Discussion period open until August 16, 2016; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Irrigation and Drainage Engineering***, © ASCE, ISSN 0733-9437.

found in Shamshirband et al. (2014), Petkovic et al. (2014), Mohammadi et al. (2015), and Aboutalebi et al. (2015).

Regarding the application of data-mining tools to study water-quality phenomena, Singh et al. (2011) implemented SVM to manage water quality in the city of Lucknow, India, using support vector classification (SVC) and support vector regression (SVR). Their results showed that SVR had better performance than kernel discriminate analysis (KDA), kernel partial least squares (KPLS), linear discriminant analysis (DA), and partial least squares (PLS). Das et al. (2012) evaluated the performance of ANN and SVM in prediction of the hydraulic conductivity coefficient of clay. The results showed that the SVR had 10% better accuracy than the ANN in predicting the hydraulic conductivity. Orouji et al. (2013) used ANFIS and GP to model water-quality parameters at the Astaheh station in Sefid Rood, Iran. They showed that GP had better accuracy than ANFIS.

This literature review revealed that less attention has been paid to the performance of data-mining tools in water quality problems. Moreover, the SVR had consistently better performance in the prediction and simulation of hydrologic phenomena than alternative methods such as numerical models and other data-mining tools such as ANFIS, ANN, and GP. This finding suggests that the SVR may be a good candidate tool for simulating complex processes such as the concentrations of pollutants at various locations within a water resources system with accuracy superior to those of traditional numerical water-quality models. The aim of this study is to evaluate the performance and accuracy of different data-mining tools to predict pollutant concentrations in river-reservoir systems.

Methodology

This section briefly summarizes the theoretical underpinnings of the ANN, GP, GA, and SVM methods.

Data Mining

Data mining is the orderly search for and use of worthy data embedded in large data sets using computers. The two main goals of data mining are classification and prediction. The most important tools in data mining are ANN, GP, and SVM, which are mostly used in prediction problems.

Artificial Neural Network

An ANN is a data-processing system that emulates the animal brain. It is used for pattern recognition, classification, and prediction problems. ANN has the capability of mapping and intelligence learning of nonlinear functions through a training process. The ability of ANN to map a set of input data (independent) to output data (dependent) with an acceptable margin of error has rendered it as useful tool for modeling. ANN has been used extensively in the field of water resources and environmental engineering. Like many other data-driven models, ANN enjoys the capacity for adaptive learning.

An ANN is composed of five main components: input data, intermediate layers, neurons (parallel processors), the transfer function, and the output data. Fig. 1 shows nonlinear mapping of an input data vector to an output data vector with ANN.

As seen in Fig. 1, data inputs are sent to an ANN. During the optimization problem (with the goal of minimizing error simulation) output values, weight, and bias are calculated and the results are processed by a nonlinear function. Finally, simulation is performed. The Levenverg-Marquardt (LM) algorithm (Marquardt 1963) is widely used for training ANNs. The LM algorithm is commonly used for solving optimization problems, such as minimizing the squared error in curve-fitting applications, and specifically in calculating the optimizing weight of an ANN.

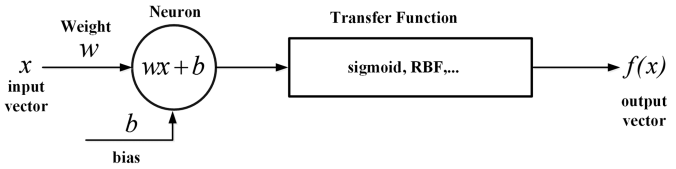

Fig. 1. Schematic components in a single layer ANN

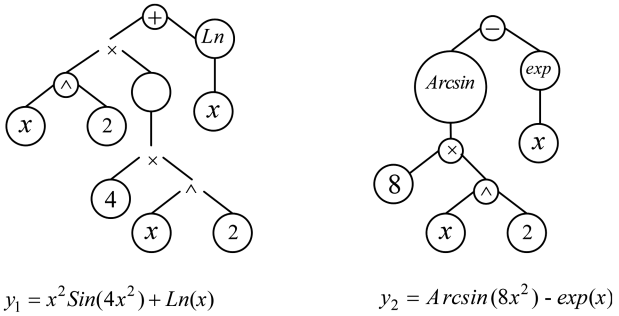

Fig. 2. Showing the mathematical equations used in GP

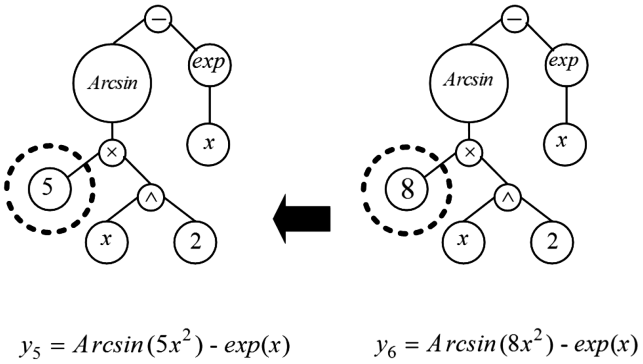

Fig. 3. Showing the various stages of crossover in GP

Genetic Programming

GP is a variety of genetic algorithm (GA), and it is a relatively new evolutionary method (Koza 1992, 1994; Banzaf et al. 1998; Khu et al. 2001). The GP is inspired by Darwin's theory of evolution. In the GA, decision variables (i.e., genes) are entered into the search process implemented to solve optimization problems. GP introduces a series of variables and functions used in the search process. These series are known as the connection series (T) and functions series (F). For example, the series $T = \{x, 1, 2, -1, -2, \ldots\}$ and $F = \{\div, \times, +, -, \exp, \sin, \cos, \log, \ldots\}$ can be chosen. Then chromosomes are generated by selecting a random initial solution set from connection and functions series. Fig. 2 shows an example of two chromosomes in GP.

Next, a corresponding objective function for each chromosome is calculated. A constraint is applied to each objective function in the form of a penalty function. In the next stage, the genetic operators (crossover and mutation) are applied. Figs. 3 and 4 illustrate how these operators are used.

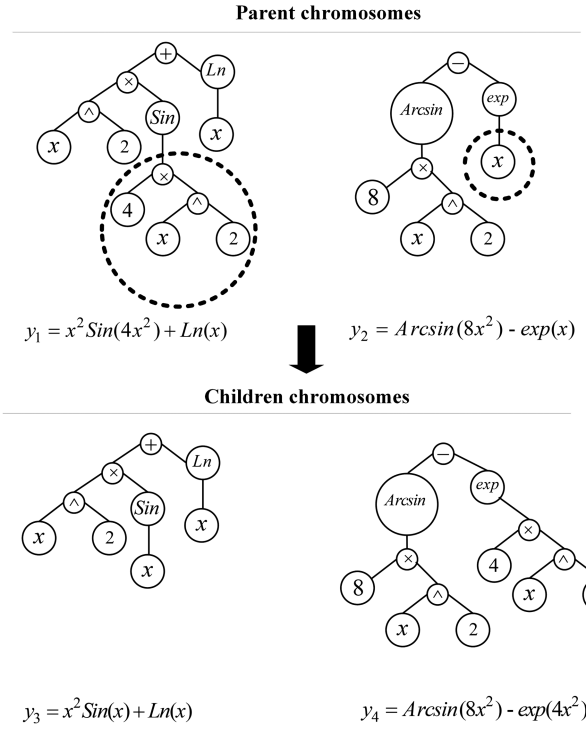

Fig. 4. Showing the various stages of mutation in GP

As shown in Fig. 3, by cutting and crossover on the parent chromosomes, new children are made and so a new generation of chromosomes is formed.

Fig. 4 depicts genes containing the number 8 changing to number 5 through mutation. Then an iterative development process is carried out on the children. Finally, nearly-optimal objectives functions are reached once they remain almost constant after repeated optimization iterations.

Support Vector Machine

SVM is a data-driven model that can classify or predict data after following a learning or training process. The SVM was introduced by Vapnik et al. (1995). Vapnik et al. (1998) extended the SVM as a forecasting tool in various branches of science and engineering. In the following, the regression form of SVM (SVR) is described.

Support Vector Regression

The SVR, like other data-driven models, must undergo a training (or calibration) process. In other words, after achieving weights and bias based on a training data set (input and output data), the SVR enters a testing process whereby it must approximate observed values. Vapnik et al. (1998) defined two functions used by SVR. The first function in the training process calculates the errors of predictions. The second (linear) function calculates the output values based on the values of the input data, weights, and bias. The first function, called error function of epsilon insensitive (e-intensive) function of SVR, is as follows:

$$|y - f(\mathbf{x})| = \begin{cases} 0 & \text{if } |y - f(\mathbf{x})| \le \kappa \\ |y - f(\mathbf{x})| - \kappa = \xi & \text{otherwise} \end{cases} \quad (1)$$

in which $\boldsymbol{x}$ = vector of input variables; y = value of observed output; $f(\boldsymbol{x})$ = value of calculated output by SVR; κ = sensitivity of prediction error $|y - f(\boldsymbol{x})|$ (this is an SVR parameter);

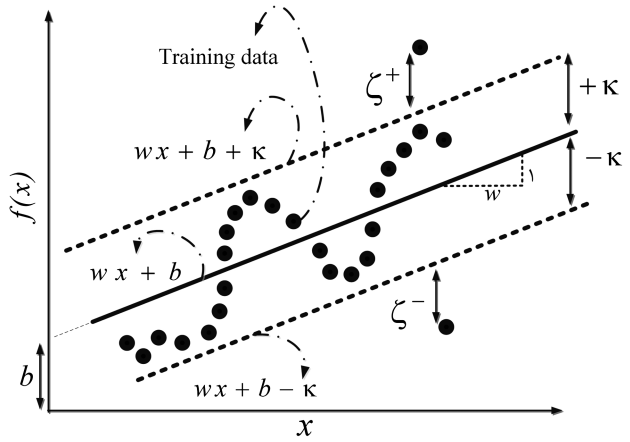

Fig. 5. Geometry of the e-insensitive function

ξ = penalty for the values that are out of range $(-\kappa, +\kappa)$; and $|\ldots|$ = absolute sign. Eq. (1) is illustrated in Fig. 5.

It is seen in Fig. 5 that the main feature of this function is that the e-intensive function does not consider a penalty for the values that are in the range of $(-\kappa, +\kappa)$. The values that are assigned penalties are outside of the range of $(-\kappa, +\kappa)$, which receive a penalty equal to ξ.

The second function (the computational function) of SVR is given by Eq. (2)

$$f(\mathbf{x}) = \mathbf{w}^T \cdot \mathbf{x} + b \quad (2)$$

in which $\boldsymbol{w}$ = value of a vector of weights of the vector of variable $\boldsymbol{x}$; b = bias value of $\mathbf{w}^T \cdot \mathbf{x}$ with respect to the observed value; and T = transpose symbol.

SVR calculates w and b by solving an optimization problem. The objective of the optimization problem is to minimize the e-intensive function and vector w. In addition to minimizing these two objectives, the calculated responses are located with respect to the range $(-\kappa, +\kappa)$ and are appended as a constraints to the optimization model. The SVR optimization model with constraints is defined as follows:

$$\text{Minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i^- + \xi_i^+) \quad (3)$$

$$\text{Subject to } (\mathbf{w}^T \cdot \mathbf{x} + b) - y_i < \kappa + \xi_i^+, \quad i = 1, 2, \ldots, m$$
$$y_i - (\mathbf{w}^T \cdot \mathbf{x} + b) \le \kappa + \xi_i^-, \quad i = 1, 2, \ldots, m \quad \xi_i^+, \xi_i^- \ge 0 \quad (4)$$

in which C = penalty coefficient; m = number of training data; ξ_i^- and ξ_i^+ = deviations located respectively above and below the range of $(-\kappa, +\kappa)$; and y_i = ith output value.

To solve the optimization problem [Eq. (3)] the Lagrange objective function (L) is formed as follows:

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i^+ + \xi_i^-) - \sum_{i=1}^{m}(\eta_i\xi_i^+ + \eta_i^*\xi_i^-)$$
$$- \sum_{i=1}^{m}\alpha_i[\kappa + \xi_i^+ + y_i - (f(\mathbf{x})] - \sum_{i=1}^{m}\alpha_i^*[\kappa + \xi_i^- + y_i - (f(\mathbf{x})] \quad (5)$$

$$\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \ge 0 \quad (6)$$

in which $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$ = Lagrange coefficients corresponding to the training data.

The partial derivatives with respect to variables b and w are set equal to zero

$$\partial L/\partial b = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) \tag{7}$$

$$\partial L/\partial w = w - \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)x_i = 0 \tag{8}$$

After achieving the α_i and α_i^*, the value of w is calculated as follows:

$$w = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)x_i \tag{9}$$

Replacing Eq. (9) into Eq. (2) and solving the ensuing optimization problem yields e Eqs. (10)–(13) with the final solution as follows:

$$\alpha[\kappa + \xi_i^- + y_i - (\mathbf{w}^T \cdot x_i + b)] = 0 \quad i = 1, 2, \ldots, m \tag{10}$$

$$\alpha_i^*[\kappa + \xi_i^+ + y_i - (\mathbf{w}^T \cdot x_i + b)] = 0 \quad i = 1, 2, \ldots, m \tag{11}$$

$$(C - \alpha_i)\xi_-^i = 0 \quad i = 1, 2, \ldots, m \tag{12}$$

$$(C - \alpha_i^*)\xi_i^+ = 0 \quad i = 1, 2, \ldots, m \tag{13}$$

The values of α_i, α_i^*, ξ_i^+, and ξ_i^- are obtained from Eqs. (10)–(13). The values b_i are evaluated with Eq. (14)

$$b_i = y_i - (\mathbf{w}^T \cdot x_i) + \kappa \quad \forall\ i = 1, 2, \ldots, m \tag{14}$$

After calculating the values of w and b_i, replacing them in Eq. (2), yields the output value of SVR. C and κ are the SVR parameters. They are determined as explained in the next subsection.

Nonlinear Support Vector Regression

Vapnik et al. (1995) considered linear regression functions. When the linear functions do not fit the training data, one can use transfer or kernel functions. The SVR kernel function is called the transition function. The kernel functions are processing functions, which transform nonlinear data into semilinear or linear data. For example, applying the logarithmic function on data with exponential distribution renders them linear. The introduction of the kernel function in the SVR transforms Eq. (2) into Eq. (15)

$$f(x) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)K(x_i, x) + b \tag{15}$$

in which K = kernel function. Dibike et al. (2001) used different kernel functions to model rainfall-runoff by means of SVR. They showed that the radial basis function (RBF) kernel has better performance than other functions. Han and Clacki (2004) also concluded that the RBF results in better performance of the regression process. The RBF equation is as follows:

$$K(x, x_i) = \exp\left(-\frac{|x - x_i|^2}{2\gamma^2}\right) \quad i = 1, 2, \ldots, m \tag{16}$$

in which γ = RBF parameter. As mentioned earlier, the use of kernel functions transforms the nonlinear behavior of data to semilinear or linear characteristics.

The SVR does not use all the data in its calculations. In fact, during the optimization phase, if the Lagrange multipliers equal zero, then those data are removed from the calculations. But if the Lagrange multipliers of data do not equal zero, then those data are called support vector (SV) and their weights and biases are calculated.

Determination of the SVR Parameters

The parameters of the SVR are C and κ, and the RBF kernel function's parameter is γ. Since the performance of SVR largely depends on its parameters, their optimal choice is paramount to the successful implementation of the algorithm (Samsudin et al. 2011). In the proposed tool refers to SVR-GA, the parameters are considered decision variables and the objective function maximizes the accuracy of the data-mining tool. Therefore, metaheuristic algorithms such as the GA find their parameters as part of the optimization process.

Criteria for Evaluating Results of ANN, GP, and SVR-GA

Eqs. (17) and (18) are used in this study to assess the performance of the ANN, GP, and SVR-GA (RMSE stands for root-mean square; R^2 refers to regression coefficient)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Hyd_{\text{obs}} - Hyd_{\text{sim}})^2}{n}} \tag{17}$$

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{n}(Hyd_{\text{obs}} - Hyd_{\text{sim}})^2}{\sum_{i=1}^{n}(Hyd_{\text{obs}} - \bar{H}yd_{\text{obs}})^2}\right] \tag{18}$$

in which n = number of observed data; Hyd_{sim} = simulated MTBE concentration in the outlet valve of the reservoir by the data-mining tool; Hyd_{obs} = simulated concentration in the outlet valve of the reservoir by CE-QUAL-W2 model; and $\bar{H}yd_{\text{obs}}$ = average of value of simulated concentration in the outlet valve of a reservoir.

Genetic Algorithm

GAs were introduced by Holland (1975). A GA is a search algorithm inspired by natural biological process of natural selection. This method is based on Darwin's theory of evolution. The GA begins with a set of initial random solutions called populations. Each population consists of a set of chromosomes, and each chromosome is a set of genes that are the decision variables of the problem. The number of populations affects the performance of the GA. If the number of initial random populations is too low, it may fail to search the entire solution space, in which case it may not converge to the optimal answer. If the number is too high, the convergence to the optimal solution could be onerously slow. The selection process is based on the merit of the objective functions corresponding to each chromosome in each generation. One can use techniques such as roulette wheel and competitive selection to select the chromosomes that are transmitted to the next generation. In the roulette wheel, depending on the fitness function value of each chromosome, a value of the level of the wheel is assigned. One of the chromosomes is selected randomly for the next generation. Chromosomes with the highest level of competence gain a higher level, which increase their probability of being selected for

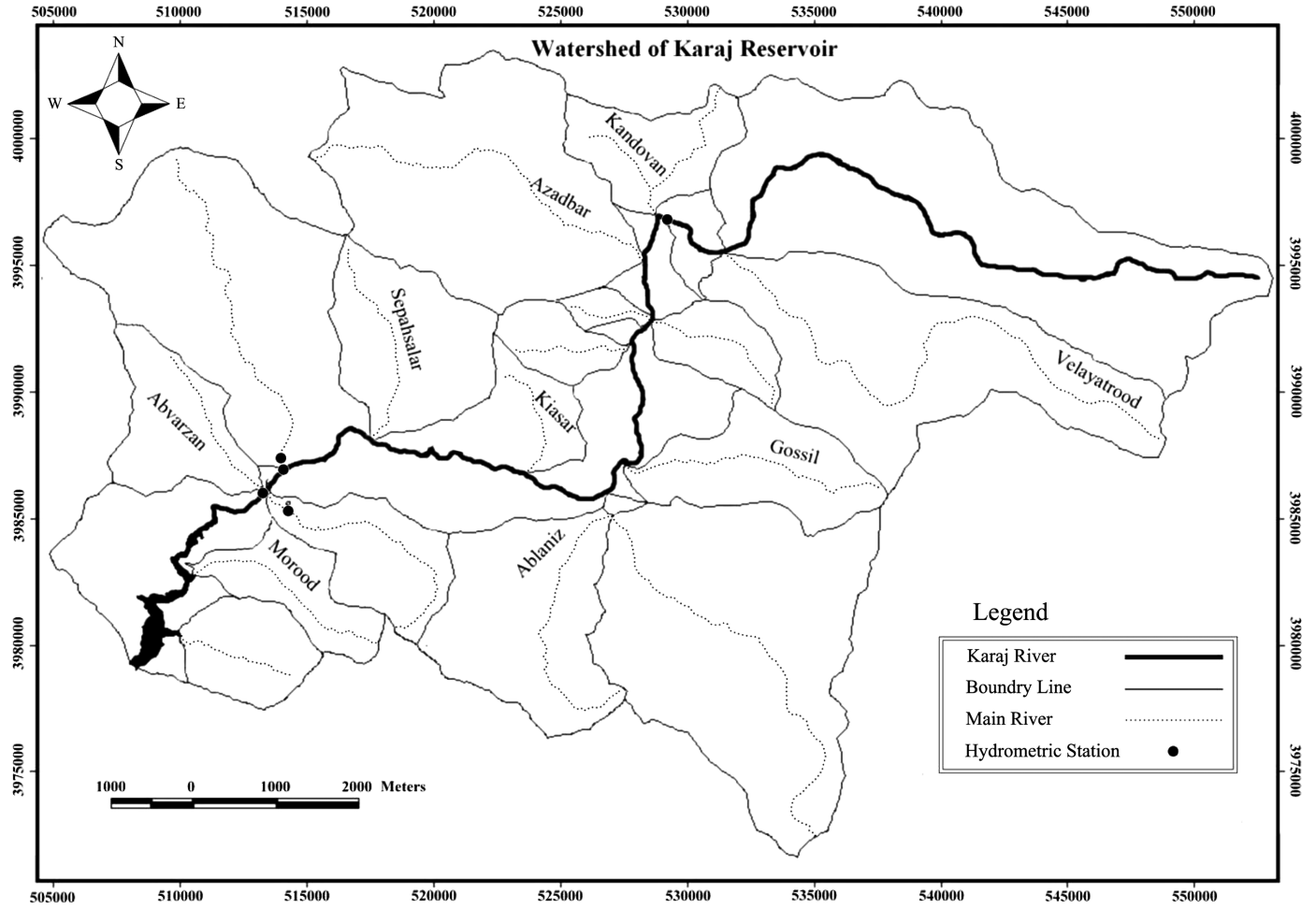

Fig. 6. Location map of Karaj reservoir within the Karaj watershed

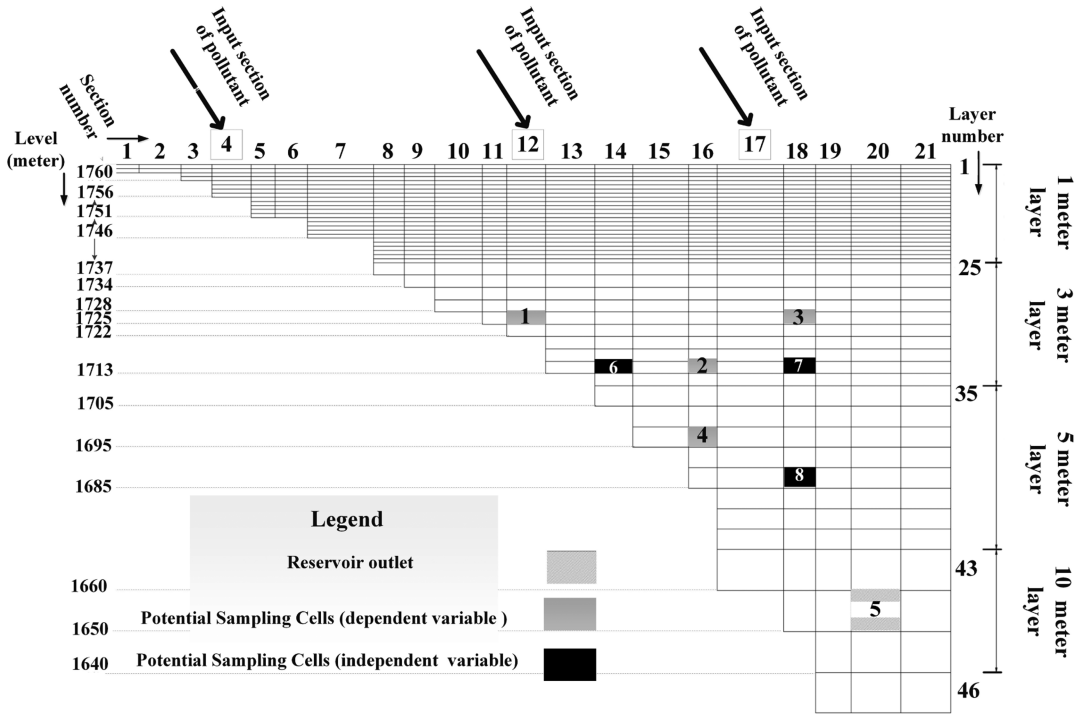

Fig. 7. Defining the discrete form of the Karaj river-reservoir system

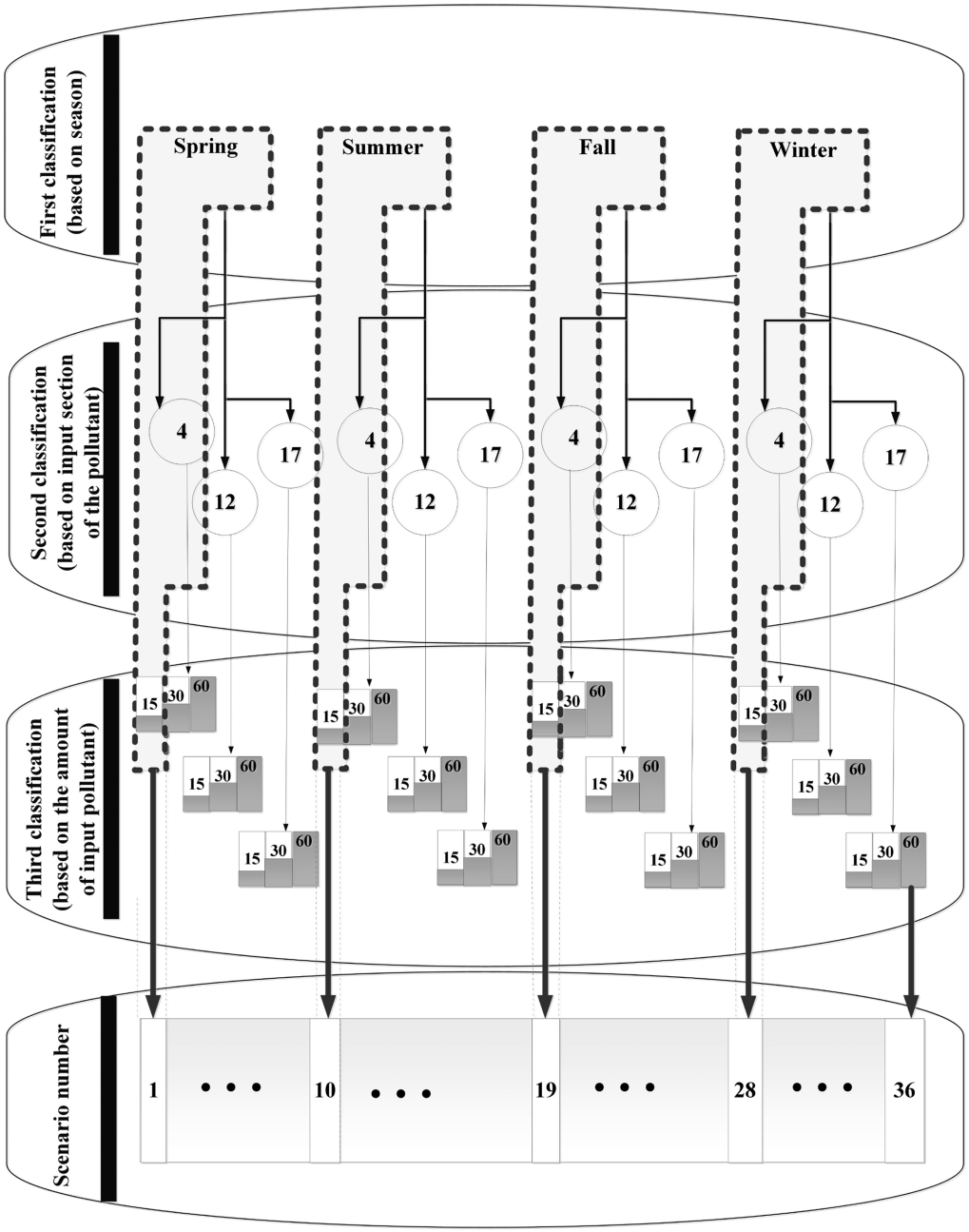

Fig. 8. Thirty-six scenarios to create a database

the next generation. In the competitive method, after the classification of chromosomes, subsidiaries are randomly created, and from each subsidiary the chromosomes that have the best objective function are selected. Thus, the selection criteria of chromosomes are based on their merits. To create the next generation, whose members are called children, one can proceed in two ways:

1. The combination of two chromosomes, using the crossover operator; and
2. Correction of chromosomes, using the mutation operator.

One of the factors affecting the performance of the GA is the rate of crossover. This rate is equal to the number of offspring produced by each generation divided by the number of people in the present generation. The larger the rate of crossover, the larger the search space. But if this ratio is too large, the search becomes prohibitively slow.

In addition, mutation is another operator that is used to create diversity of chromosomes in next populations. This operation is similar to biological process and replaces one or several genes and creates new genes for new populations. The larger the rate of mutation, the larger diversity in the populations, and vice versa.

Case Study

The Karaj dam is a source of drinking water for the cities of Tehran, Iran, and provides water needed for agriculture, and provides flood control and hydroelectric power generation. Its construction began in 1957 and was completed in 1961. Fig. 6 has a location map of the Karaj dam. This case study considers the sudden release of MTBE

in the amounts of 14,000 30,000 and 60,000 L at three locations [beginning (Section 4), middle (Section 12), and end (Section 17)] of the Karaj river-reservoir system, Iran, at the beginning of summer, autumn, winter, and spring, giving rise to $3(\text{sections}) \times 3$ $(\text{amounts of release}) \times 4(\text{seasons}) = 36$ pollution scenarios. The three methods (ANN, GP, and SVR-GA) are used to assess the MTBE contamination at selected cells for each pollution scenario, and their results are compared at selected monitoring cells.

Preparation of Simulation Model of Contaminant Transport in the Reservoir

The CE-QUAL-W2 water-quality model (Shokri et al. 2014) was used to generate MTBE concentrations in the Karaj river-reservoir system for each of the 36 pollution scenarios. The geometry of the reservoir must be defined as a discrete network in the CE-QUAL-W2 model. The geometry data was extracted from a 1:5,000-scale topographic map. This file contains information such as the number of segments, length, width, depth, and positions of reservoir features. The river-reservoir system contains 21 segments (including 2 boundary segments), and 46 layers (including 2 boundary layers) with thicknesses of 1, 3, 5, and 10 m as shown in Fig. 7. The concentration of MTBE was simulated with CE-QUAL-W2 for a period 1,385 days, which is the length of the period for which hydrometric and meteorological data were available. Fig. 7 shows that Cells 1 through 8 are places with concentration data provided by the CE-QUAL-W2 model. Cells 1 through 5 were considered the dependent or output variables and Cells 6–8 considered the independent or input variables for the data driven tools. Cell 5 is the location of the outlet valve of the reservoir.

Simulating Water Quality in the River-Reservoir System with a Data-Mining Tool

The first database for the training and testing phases of the data-mining tool was produced. The coefficients of the data-mining tools were determined based on the training phase. Finally, databases of MTBE concentrations were created with the CE-QUAL-W2 model for different modes of occurrence of pollutants (36 scenarios). In other words, the CE-QUAL-W2 model runs in each of the 36 scenarios and the MTBE concentrations were calculated in the eight cells (numbered 1 through 8) shown in Fig. 7.

Fig. 8 presents the details of the 36 pollution scenarios. For example, in the spring, there are three sections where pollution occurs (Sections 4, 12, and 17). In each of these sections, the amount of spilled MTBE can be 15,000, 30,000, or 60,000 L, originating three scenarios for each section, and nine scenarios for the three sections in the spring. This same number of scenarios is repeated in the Summer, autumn, and winter, for a total of 36 scenarios. After creating the database, a portion of the database is used as training data and the other as testing data. In order for all the data values to have the same chance of being selected in each phase (training or testing), they must be chosen randomly. In this study, training and testing data were selected in this manner. The values of the GA parameters (percentage of crossover, percentage of mutations, number of iterations, and number of initial populations), were set at 70 and 15%, and 50 and 1,000, respectively. The RBF function is used with the SVR-GA and 75% of the randomly selected data are considered as training data, and the remainder data are considered as testing data. The *MATLAB* software was used for implementing of the ANN, GP, and SVR-GA. For the implementation of the ANN, a three-layer MLP network was used. The implementation of the GP involved trigonometric, exponential, logarithmic, and polynomial functions. The number of neurons of the ANN was found by trial and error.

Table 1. Evaluation Criteria for ANN, GP, and SVR-GA

Cell number	Tool	RMSE			R^2		
		Training	Testing	Total	Training	Testing	Total
1	ANN	3.25	4.34	3.98	0.94	0.92	0.93
	GP	5.75	6.96	6.16	0.87	0.85	0.86
	SVR-GA	2.80	3.35	3.10	0.96	0.94	0.95
2	ANN	3.75	4.89	4.20	0.92	0.90	0.91
	GP	5.98	7.30	6.40	0.86	0.84	0.94
	SVR-GA	3.12	3.57	3.21	0.95	0.92	0.93
3	ANN	3.10	4.12	3.47	0.95	0.93	0.94
	GP	5.24	6.18	5.50	0.89	0.87	0.88
	SVR-GA	2.57	3.05	2.72	0.97	0.95	0.96
4	ANN	3.90	5.10	4.30	0.92	0.89	0.91
	GP	6.94	7.80	7.21	0.85	0.82	0.84
	SVR-GA	3.40	4.30	3.73	0.95	0.92	0.93
5	ANN	2.90	3.25	3.03	0.97	0.95	0.96
	GP	4.96	6.10	5.30	0.90	0.88	0.89
	SVR-GA	2.19	2.83	2.45	0.98	0.96	0.97

Results and Discussion

The results with the data mining tools are presented in Table 1.

It is seen in Table 1 that in all five cells (Cells 1, 2, 3, 4, and 5) of the Karaj river-reservoir system shown in Fig. 7, the coupled SVR-GA dominates the ANN and GP from the viewpoint of superior performance. Also, the ANN had better performance compared to GP. One important finding is that the concentrations obtained with ANN, GP, and SVR-GA at Cell 5 were better than those for the other cells. In other words, the MTBE concentrations at the release gate (on Cell 5) of reservoir were smaller than those at the other four testing cells. Table 1 indicates that the accuracies of ANN, GP, and SVR-GA were 93, 87, and 96% respectively, based on the value of R^2.

Fig. 9 presents a graphical evaluation of the accuracy of MTBE concentrations simulated with ANN, GP, and SVR-GA with respect to the observed concentrations on Cells 1 through 5. According to Fig. 9 all three data-mining tools simulated the pollutant MTBE concentration at the five testing cells (Cells 1 through 5) with an acceptable accuracy. The minimum accuracy equals 82% in Table 1 for R^2 corresponding to GP in testing process. Table 2 shows the optimized values of the SVR-GA tool, and the number of hidden layers of neurons of the ANN, which were derived by trial and error.

Concluding Remarks

The aim of this study was to simulate MTBE concentrations at different locations of the Karaj river-reservoir system using a coupled SVR-GA tool, and to evaluate the performance of this tool compared with the ANN and GP methods. First, a database was prepared using CE-QUAL-W2 model. This data base contains the MTBE concentrations at different locations of the Karaj river-reservoir system corresponding to 36 scenarios pollutant release into reservoir system. Thereafter, by considering the MTBE concentrations at three locations of pollutant entry into the reservoir, the MTBE concentrations at five other points were simulated. The

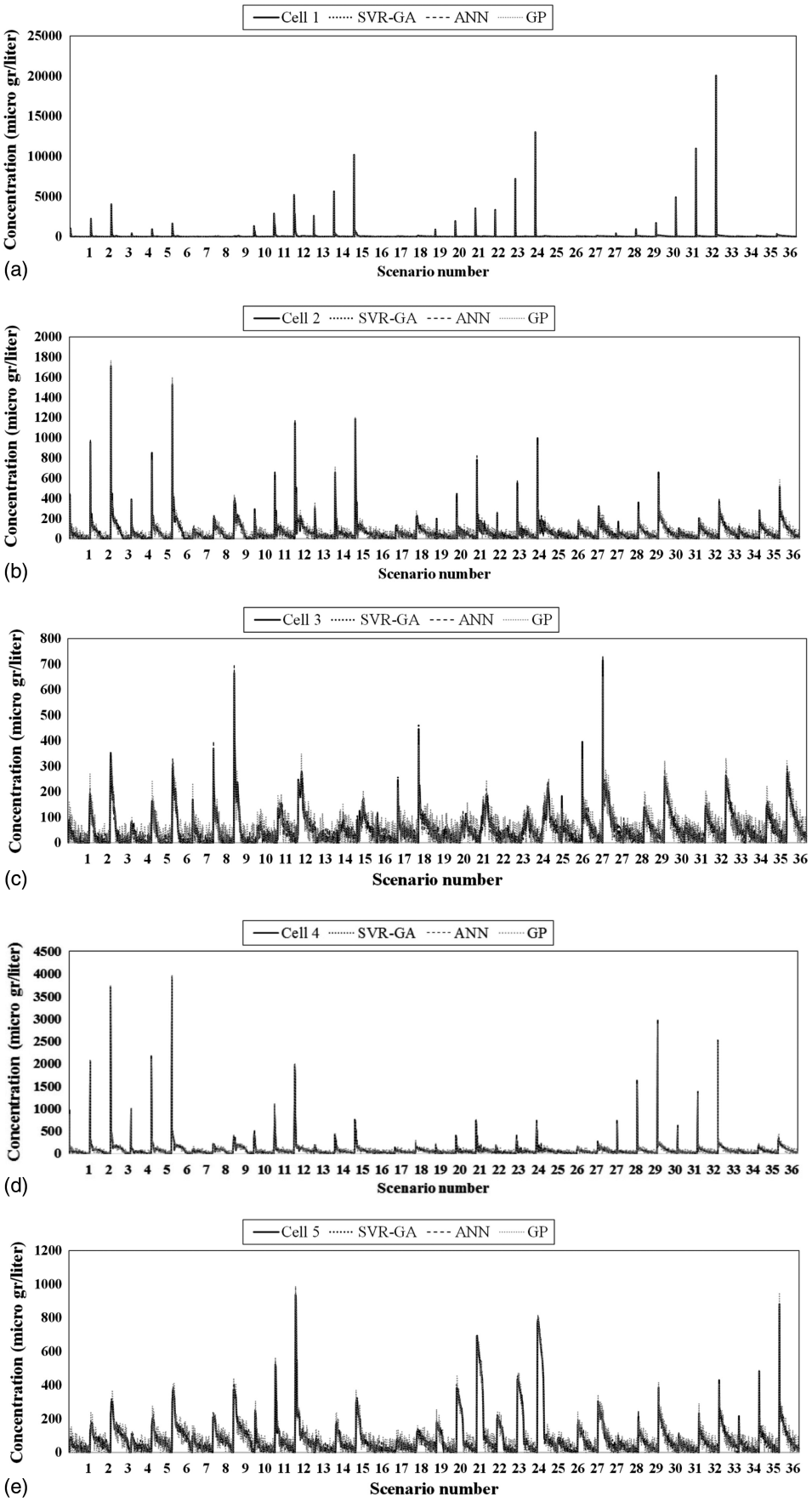

Fig. 9. Simulated MTBE concentrations with the ANN, GP, and SVR-GA and the observed concentrations at (a) testing Cell 1; (b) testing Cell 2; (c) testing Cell 3; (d) testing Cell 4; (e) testing Cell 5

Table 2. Parameters of the ANN and SVR

	SVR			ANN
Cell number	κ	γ	C	Number of neurons
1	0.001	28	58	25
2	0.004	39	75	54
3	0.003	53	83	37
4	0.001	19	92	42
5	0.002	36	63	39

results indicated that the SVR-GA had 3 and 9% better predictive accuracy than the ANN and GP, respectively, based on the value of R^2.

References

Aboutalebi, M., Bozorg-Haddad, O., and Loaiciga, H. A. (2015). "Optimal monthly reservoir operation rules for hydropower generation derived with SVR-NSGAII." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000553, 04015029.

Ahmadi, M., Bozorg-Haddad, O., and Mariño, M. A. (2014). "Extraction of flexible multi-objective real-time reservoir operation rules." *Water Resour. Manage.*, 28(1), 131–147.

Asefa, T., Kemblowski, M., McKee, M., and Khalil, A. (2006). "Multi-time scale stream flow predictions: The support vector machines approach." *J. Hydrol.*, 318(1–4), 7–16.

Ashofteh, P. S., Bozorg-Haddad, O., and Loáiciga, H. A. (2015a). "Evaluation of climatic-change impacts on multi-objective reservoir operation with multiobjective genetic programming." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000540, 04015030.

Ashofteh, P.-S., Bozorg-Haddad, O., Akbari-Alashti, H., and Mariño, M. A. (2015b). "Determination of irrigation allocation policy under climate change by genetic programming." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000807, 04014059.

Ashofteh, P.-S., Bozorg-Haddad, O., and Mariño, M. A. (2013a). "Climate change impact on reservoir performance indices in agricultural water supply." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000496, 85–97.

Ashofteh, P.-S., Bozorg-Haddad, O., and Mariño, M. A. (2013b). "Scenario assessment of streamflow simulation and its transition probability in future periods under climate change." *Water Resour. Manage.*, 27(1), 255–274.

Ashofteh, P.-S., Bozorg-Haddad, O., and Mariño, M. A. (2015c). "Risk analysis of water demand for agricultural crops under climate change." *J. Hydrol. Eng.*, 10.1061/(ASCE)HE.1943-5584.0001053, 04014060.

Banzhaf, W., Nordin, P., Keller, R., and Francone, F. D. (1998). *Genetic programming: An introduction*, Morgan Kaufmann, San Francisco.

Behzad, M., Asghari, K., Eazi, M., and Palhang, M. (2009). "Generalization performance of support vector machines and neural networks in runoff modeling." *Exp. Syst. Appl.*, 36(4), 7624–7629.

Beygi, S., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A., (2014). "Bargaining models for optimal design of water distribution networks." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000324, 92–99.

Bolouri-Yazdeli, Y., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A. (2014). "Evaluation of real-time operation rules in reservoir systems operation." *Water Resour. Manage.*, 28(3), 715–729.

Bozorg-Haddad, O., Ashofteh, P.-S., Ali-Hamzeh, M., and Mariño, M. A. (2015a). "Investigation of reservoir qualitative behavior resulting from biological pollutant sudden entry." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000865, 04015003.

Bozorg-Haddad, O., Ashofteh, P.-S., and Mariño, M. A. (2015b). "Levee's layout and design optimization in protection of flood areas." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000864, 04015004.

Bozorg-Haddad, O., Ashofteh, P.-S., Rasoulzadeh-Gharibdousti, S., and Mariño, M. A. (2014). "Optimization model for design-operation of pumped-storage and hydropower systems." *J. Energy Eng.*, 10.1061/(ASCE)EY.1943-7897.0000169, 04013016.

Bozorg-Haddad, O., Rezapour Tabari, M. M., Fallah-Mehdipour, E., and Mariño, M. A. (2013). "Groundwater model calibration by meta-heuristic algorithms." *Water Resour. Manage.*, 27(7), 2515–2529.

Das, S., Samui, P., and Sabat, A. (2012). "Prediction of field hydraulic conductivity of clay liners using an artificial neural network and support vector machine." *J. Geomech.*, 10.1061/(ASCE)GM.1943-5622.0000129, 606–611.

Dibike, Y. B., Velickov, S., Solomatine, D. P., and Abbott, M. B. (2001). "Model induction with support vector machines: Introduction and application." *J. Comput. Civ. Eng.*, 10.1061/(ASCE)0887-3801(2001)15:3(208), 208–216.

Fallah-Mehdipour, E., Bozorg-Haddad, O., and Mariño, M. A. (2013a). "Extraction of optimal operation rules in aquifer-dam system: A genetic programming approach." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000628, 872–879.

Fallah-Mehdipour, E., Bozorg-Haddad, O., and Mariño, M. A. (2013b). "Prediction and simulation of monthly groundwater levels by genetic programming." *J. Hydro-Environ. Res.*, 7(4), 253–260.

Han, D., and Cluckie, I. (2004). "Support vector machines identification for runoff modeling." *Proc., 6th Int. Conf. on Hydroinformatics*, World Scientific Publishing, Singapore.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*, University of Michigan Press, Ann Arbor, MI.

Hou, D., Ge, X., Huang, P., Zhang, G., and Loáiciga, H. A. (2014). "A real-time dynamic early warning, model based on uncertainty analysis and risk assessment for sudden water-pollution accidents." *Environ. Sci. Pollut. Res.*, 21(14), 8878–8892.

Khu, S. T., Liong, S. Y., Babovic, V., Madsen, H., and Muttil, N. (2001). "Genetic programming and its application in real-time runoff forecasting." *J. Am. Water Resour. Assoc.*, 37(2), 439–451.

Koza, J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection*, MIT Press, Cambridge, MA.

Koza, J. R. (1994). *Genetic programming II: Automatic discovery of reusable programs*, MIT Press, Cambridge, MA.

Maity, R., Bhagwat, P., and Bhatnagar, A. (2013). "Potential of support vector regression for prediction of monthly streamflow using endogenous property." *Hydrol. Process.*, 24(7), 917–923.

Marquardt, D. (1963). "An algorithm for least-squares estimation of nonlinear parameters." *J. Appl. Math.*, 11(3), 431–441.

MATLAB 7.10.0 [Computer software]. MathWorks, Natick, MA.

Mohammadi, K., Shamshirband, S., Tong, C., Arif, M., Petkovi, D., and Ch, S. (2015). "A new hybrid support vector machine-wavelet transform approach for estimation of horizontal global solar radiation." *Energy Convers. Manage.*, 92, 162–171.

Orouji, H., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A. (2013). "Modeling of water quality parameters using data-driven models." *J. Environ. Eng.*, 10.1061/(ASCE)EE.1943-7870.0000706, 947–957.

Orouji, H., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A. (2014). "Extraction of decision alternatives in project management: Application of hybrid PSO-SFLA." *J. Manage. Eng.*, 10.1061/(ASCE)ME.1943-5479.0000186, 50–59.

Petkovi, D., et al. (2014). "Evaluation of modulation transfer function of optical lens system by support vector regression methodologies—A comparative study." *Infrared Phys. Technol.*, 65, 94–102.

Samsudin, R., Saad, P., and Shabri, A. (2011). "River flow time series using least squares support vector machines." *Hydrol. Earth Syst.*, 15(6), 1835–1852.

Savic, D. A., Walters, G. A., and Davidson, J. W. (1999). "A genetic programming approach to rainfall-runoff modeling." *Water Resour. Manage.*, 13(3), 219–231.

Seifollahi-Aghmiuni, S., Bozorg-Haddad, O., and Mariño, M. A. (2013). "Water distribution network risk analysis under simultaneous consumption and roughness uncertainties." *Water Resour. Manage.*, 27(7), 2595–2610.

Shamshirband, S., et al. (2014). "Support vector regression methodology for wind turbine reaction torque prediction with power-split hydrostatic continuous variable transmission." *Energy*, 67, 623–630.

Shokri, A., Bozorg-Haddad, O., and Mariño, M. A. (2013). "Reservoir operation for simultaneously meeting water demand and sediment flushing: A stochastic dynamic programming approach with two uncertainties." *J. Water Resour. Plann. Manage.*, 139(3), 277–289.

Shokri, A., Bozorg-Haddad, O., and Mariño, M. A. (2014). "Multi-objective quantity-quality reservoir operation in sudden pollution." *Water Resour. Manage.*, 28(2), 567–586.

Singh, K. P., Basant, N., and Gupta, S. (2011). "Support vector machines in water quality management." *Anal. Chim. Acta*, 703(2), 152–162.

Sivapragasam, C., Vasudevan, G., and Vincent, P. (2007). "Effect of inflow forecast accuracy and operating time horizon in optimizing irrigation release." *Water Resour. Manage.*, 21(6), 933–945.

Soltanjalili, M., Bozorg-Haddad, O., and Mariño, M. A. (2013). "Operating water distribution networks during water shortage conditions using hedging and intermittent water supply concepts." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000315, 644–659.

Vapnik, V. (1995). *The nature of statistical learning theory*, Springer, New York.

Vapnik, V. (1998). *Statistical learning theory*, Wiley, New York.

Wang, W. C., Chau, K. W., Cheng, C. T., and Qiu, L. (2009). "A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series." *J. Hydrol.*, 374(3–4), 294–306.

Wei, C. (2012). "Wavelet kernel support vector machines forecasting techniques: Case study on water-level predictions during typhoons." *Exp. Syst. Appl.*, 39(5), 5189–5199.

Yoon, H., Jun, S., Hyun, Y., Bae, G., and Lee, K. (2011). "A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer." *J. Hydrol.*, 396(1-2), 128–138.