Title

Modeling Water-Quality Parameters Using Genetic Algorithm—Least Squares Support Vector Regression and Genetic Programming

Journal

Journal of Environmental Engineering, 143(7)

ISSN

0733-9372

Authors

Bozorg-Haddad, Omid

Soleimani, Shima

Loáiciga, Hugo A

Publication Date

2017-07-01

DOI

10.1061/(asce)ee.1943-7870.0001217

Peer reviewed

Modeling Water-Quality Parameters Using Genetic Algorithm–Least Squares Support Vector Regression and Genetic Programming

Omid Bozorg-Haddad[1]; Shima Soleimani[2]; and Hugo A. Loáiciga, F.ASCE[3]

Abstract: The modeling and monitoring of water-quality parameters is necessary because of the ever increasing use of water resources and contamination caused by sewage disposal. This study employs two data-driven methods for modeling water-quality parameters. The methods are the least-squares support vector regression (LSSVR) and genetic programming (GP). Model inputs to the LSSVR algorithm and GP were determined using principal component analysis (PCA). The coefficients of the LSSVR were selected by sensitivity analysis employing statistical criteria. The results of the sensitivity analysis of the LSSVR showed that its accuracy depends strongly on the values of its coefficients. The value of the Nash-Sutcliffe (NS) statistic was negative for 60% of the combinations of coefficients applied in the sensitivity analysis. That is, using the mean of a time series would produce a more accurate estimate of water-quality parameters than the LSSVR method in 60% of the combinations of parameters tried. The genetic algorithm (GA) was combined with LSSVR to produce the GA-LSSVR algorithm with which to achieve improved accuracy in modeling water-quality parameters. The GA-LSSVR algorithm and the GP method were employed in modeling Na^+, K^+, Mg^{2+}, SO_4^{2-}, Cl^-, pH, electric conductivity (EC), and total dissolved solids (TDS) in the Sefidrood River, Iran. The results indicate that the GA-LSSVR algorithm has better accuracy for modeling water-quality parameters than GP judged by the coefficient of determination (R^2) and the NS criterion. The NS static established, however, that the GA-LSSVR and GP methods have the capacity to model water-quality parameters accurately. **DOI: 10.1061/(ASCE)EE.1943-7870.0001217.** © *2017 American Society of Civil Engineers.*

Author keywords: Genetic algorithm–least squares support vector regression (GA-LSSVR) algorithm; Genetic programming (GP) method; Water quality; Modeling; Sensitivity analysis; Principal component analysis.

Introduction

The monitoring of water-quality parameters in rivers is becoming increasingly important due to the rise in freshwater use. Field monitoring and testing of freshwater is time-consuming and costly (Chapra 2008). Alternatively, there are data-driven methods for determining water-quality characteristics. Genetic programming (GP) is one of the applications of the genetic algorithm (GA) and has been effective in approximating complex functions (Izadifar and Elshorbagy 2010). Aytek and Alp (2008) used GP to model rainfall runoff. Kisi and Shiri (2010) applied GP to forecast short-term and long-term river flow. Izadifar and Elshorbagy (2010) modeled evapotranspiration using GP. The method was applied to model the stage-discharge curve (Azamathulla et al. 2011). Hashmi et al. (2011) implemented GP to downscaling precipitation data. Genetic programming was applied to estimating the scour depth, forecasting the suspended sediment, estimating river water quality parameters, and total dissolved solids (TDS) (Azamathulla 2012; Kisi et al. 2012; Ghavidel and Montaseri 2014; Orouji et al. 2013). Genetic programming is effective for solving large combinatorial problems (Azamathulla and Ghani 2011), yet it is very sensitive to the choice of initial parameters, such as the mutation rate.

The use of support vector regression (SVR) in hydrology has increased in recent decades. Raghavendra and Deka (2014) reviewed the application of SVR in hydrology. The least-squares support vector regression (LSSVR) method is a data-driven method, which was developed by Suykens et al. (2002). Tripathi et al. (2006) and Anandhi et al. (2008) employed the LSSVR method to downscale precipitation data. Maity et al. (2010) forecasted the discharge of the Mahanadi River in India by means of the autoregressive integrated moving average (ARIMA) and LSSVR methods. The correlation coefficient between the observed and predicted streamflows was found to be 0.77 and 0.67 for LSSVR and ARIMA, respectively. The comparison of results showed the superior accuracy of the LSSVR method. Bhagwat and Maity (2013) predicted daily flow discharge in the Narmada River and the Mahanadi River using the LSSVR method. The coefficients of the LSSVR method were calculated by trial and error. Their results showed that the minimum and average flow discharges were predicted by the LSSVR appropriately, whereas the accuracy of maximum flow discharge predictions was poor.

The LSSVR method has been used to forecast and model different water-quality parameters (Tan et al. 2012). Yunrong and Liangzhong (2009) studied the prediction of river water quality using the LSSVR model in the Liuxi River, China. They predicted chemical oxygen demand (COD), dissolved oxygen (DO), and

[1]Professor, Dept. of Irrigation and Reclamation Engineering, Faculty of Agricultural Engineering and Technology, College of Agriculture and Natural Resources, Univ. of Tehran, Karaj, 31587-77871 Tehran, Iran (corresponding author). E-mail: OBHaddad@ut.ac.ir

[2]Formerly, M.Sc. Graduate Student, Dept. of Irrigation and Reclamation Engineering, Faculty of Agricultural Engineering and Technology, College of Agriculture and Natural Resources, Univ. of Tehran, Karaj, 31587-77871 Tehran, Iran. E-mail: Shimasoleimani@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 May 20, 2016; approved on December 9, 2016; published online on February 27, 2017. Discussion period open until July 27, 2017; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Environmental Engineering***, © ASCE, ISSN 0733-9372.

other water-quality characteristics by means of a combined algorithm comprising the LSSVR methods and the particle swarm optimization (PSO) algorithm. Singh et al. (2011) utilized the clustering method, or support vector clustering (SVC), to optimize surface water quality monitoring in the city of Lucknow, India. Tan et al. (2012) predicted phosphorus values in China with the LSSVR method. They compared the efficiency of the LSSVR method with neural networks of the radial basis function (RBF) and back-propagation (BP). Liu et al. (2013) addressed water-quality parameters prediction in aquaculture employing the GP and real-value genetic algorithm support vector regression (RGA-SVR). They used the GA to modify the coefficients of the SVR method. Their comparison of results showed the superiority of the RGA-SVR algorithm over other methods.

Several studies have dealt with the selection of input variables to a model and determining the model structure. Yoon et al. (2011) applied cross correlation to determine the inputs to the support vector machine (SVM) model to predict groundwater level. Noori et al. (2010) employed principal component analysis (PCA) to select model structure and inputs that have the greatest effect on the output of an artificial neural network (ANN) model to estimate daily monoacid concentration. Noori et al. (2011) determined ANN model inputs by applying the gamma test to estimate solid waste characteristics. Fallah-Mehdipour et al. (2014) employed default (user-defined) models to predict groundwater using GP method. Ghavidel and Montaseri (2014) implemented stepwise regression to determine model structure to estimate TDS applying the ANN model.

A review of the literature demonstrates that data-driven methods are applicable in many fields of hydrology and water resources management, but were not widely applied in recent investigations (Ahmadi et al. 2015; Akbari-Alashti et al. 2014; Beygi et al. 2014; Bozorg-Haddad et al. 2013, 2015b, a; Farhangi et al. 2012; Fallah-Mehdipour et al. 2013a, b, c; Jahandideh-Tehrani et al. 2015; Orouji et al. 2014a, b). The LSSVR method is a relatively new method among data-driven ones whose capability has been proven to predict and model phenomena in various fields. Previous work has shown that the LSSVR's most significant disadvantage is its high sensitivity to the trade-off parameter between the error margin (γ) and the width of the Gaussian basis function (σ). This work applies an optimization algorithm (GA) with the LSSVR to adjust and optimize the LSSVR coefficients. The GA-LSSVR algorithm is applied to model various water-quality parameters in the Sefidrood River, Iran. To this date, several authors have modeled water-quality parameters such as biological oxygen demand (BOD), COD, and DO (Yunrong and Liangzhong 2009; Singh et al. 2011). This work targets several other key parameters that include Na^+, K^+, Mg^{2+}, SO_4^{2-}, Cl^-, pH, electric conductivity (EC), and TDS. Moreover, the results of the proposed GA-LSSVR algorithm will be compared with those obtained with the GP method in modeling water-quality parameters.

LSSVR Method

The LSSVR method was developed by Suykens et al. (2002), who modified the original method introduced by Vapnik (1995). The LSSVR method assumes that the relation between the input and output data is nonlinear, although by means of a mapping function called Feature Space the relation between the input and output data is made linear. A linear between inputs and outputs in the feature space is expressed by Eq. (1) (Vapnik 1995)

$$y(x) = \boldsymbol{\omega}' \cdot \varphi(x) + b \tag{1}$$

where x and $y(x)$ = input and output of the observed training data set, respectively; $\boldsymbol{\omega}'$ = transposed form of the weighting vector $\boldsymbol{\omega}$; $\phi(x)$ = nonlinear vectorial function that maps the data from the domain space to the range space; and b = bias. The parameters b and ω are obtained by solving the following optimization equations:

$$\min J(\omega, e) = \frac{1}{2}\boldsymbol{\omega}'\boldsymbol{\omega} + \frac{1}{2}\gamma\sum_{t=1}^{T} e_t^2$$

Subject to

$$y_t = \boldsymbol{\omega}'\varphi(x_t) + b + e_t \qquad t = 1, 2, \ldots, T \tag{2}$$

in which $J(\omega, e)$ = loss function; e_t = error at each time step t; γ = adjustable coefficient; and t = time step counter that varies from 1 to T. The Lagrange function of Eqs. (1) and (2) is defined as follows:

$$L(\boldsymbol{\omega}, b, e_t, l_t) = J(\boldsymbol{\omega}, e_t) - \sum_{t=1}^{T} l_t[\boldsymbol{\omega}'\varphi(x_t) + b + e_t - y_t] \qquad t = 1, 2, \ldots, T \tag{3}$$

where $L(\omega, b, e_t, l_t)$ = Lagrange function; and l_t = Lagrange coefficients at the time step t, which are obtained by taking the partial derivatives of the Lagrangian L

$$\frac{\partial L}{\partial \omega} = 0 \rightarrow \boldsymbol{\omega} = \sum_{t=1}^{T} l_t\varphi(x_t)$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{t=1}^{T} l_t = 0$$

$$\frac{\partial L}{\partial e_t} = 0 \rightarrow l_t = \gamma e_t \qquad t = 1, 2, \ldots, T$$

$$\frac{\partial L}{\partial l_t} = 0 \rightarrow \boldsymbol{\omega}'\varphi(x_t) + b + e_t - y_t = 0 \qquad t = 1, 2, \ldots, T \tag{4}$$

where x_t = tth data of input vector $\mathbf{x}$. Mercer's condition introduces a kernel function $K(\mathbf{x}, x_t)$ as follows:

$$K(\mathbf{x}, x_t) = \varphi(\mathbf{x})'\varphi(x_t) \qquad t = 1, 2, \ldots, T \tag{5}$$

As a result, the output y estimated (computed) by the LSSVR method is obtained with Eq. (6)

$$\hat{y} = \sum_{t=1}^{T} l_t K(\mathbf{x}, x_t) + b \tag{6}$$

where $\hat{y}$ = estimated value of y by the LSSVR method. The kernel function K can be selected among the linear, polynomial, and sigmoid functions as well as RBF and multi layer perceptron (MLP). The RBF kernel was applied in this study

$$K(\mathbf{x}, x_t) = e^{(-\|\mathbf{x}-x_t\|^2/\sigma^2)} \tag{7}$$

where σ = coefficient of the RBF. The LSSVR method is portrayed graphically in Fig. 1.

The LSSVR method does not propose any mechanism for selecting its coefficients σ and γ. Nevertheless, the LSSVR method precision is highly dependent on the selection of these coefficients (Liu et al. 2013).

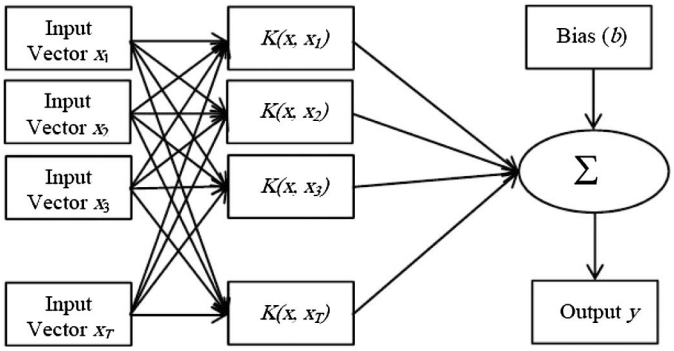

Fig. 1. Diagram of the LSSVR method

Optimal Selection of the LSSVR Coefficients Using the GA

Evolutionary optimization algorithms find solutions in the neighborhood of global optima, and have been used in many optimization problems related to water resources (Bozorg-Haddad et al. 2011, 2014a, b; Su et al. 2013). The GA is a heuristic search method, inspired by the evolution of organisms in nature commonly applied to solve optimization problems (Chiu and Chen 2009). An optimization problem is defined as a community of genotypes in the GA (Soleimani et al. 2016). According to the law of survival of the fittest, the best individuals are selected for reproduction due to their ability to survive and adapt to their environments. Those selected individuals produce a new generation by crossover and mutation processes, unleashing an evolutionary process across generations producing individuals that either adapt to their environments or go extinct. The population of best individuals obtained through many generations approaches the near-optimal solution of an optimization problem. Fig. 2 is a flowchart of the GA-LSSVR algorithm.

The GA was applied in this study with 20 populations, one elitism, 0.8 crossover rate, 0.02 mutation rate, and 100 iterations, for finding the coefficients of the LSSVR method (Deb and Deb 2014). The objective function used in the study is that written in Eq. (8). For comparison purposes, the root-mean-square error (RMSE), the R^2, and the Nash-Sutcliffe (NS) statistic were calculated also. The NS statistic is a goodness-of-fit index that ranges from $-\infty$ to 1, whereby a NS statistic equal to 1 indicates that the calculated data have a perfect match with the observed data; a value equal to 0 means that the calculated data are no better than using the average of the data as the predictor; and a value less than 0 implies that the calculated values are of less quality than using the average as the predictor. The RMSE, R^2, and NS criteria are respectively given by Eqs. (8)–(10)

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \tag{8}$$

$$R^2 = \frac{\sum_{t=1}^{T}(y_t - \bar{y})^2(\hat{y}_t - \hat{\bar{y}})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2 \times \sum_{t=1}^{T}(\hat{y}_t - \hat{\bar{y}})^2} \tag{9}$$

$$\text{NS} = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \tag{10}$$

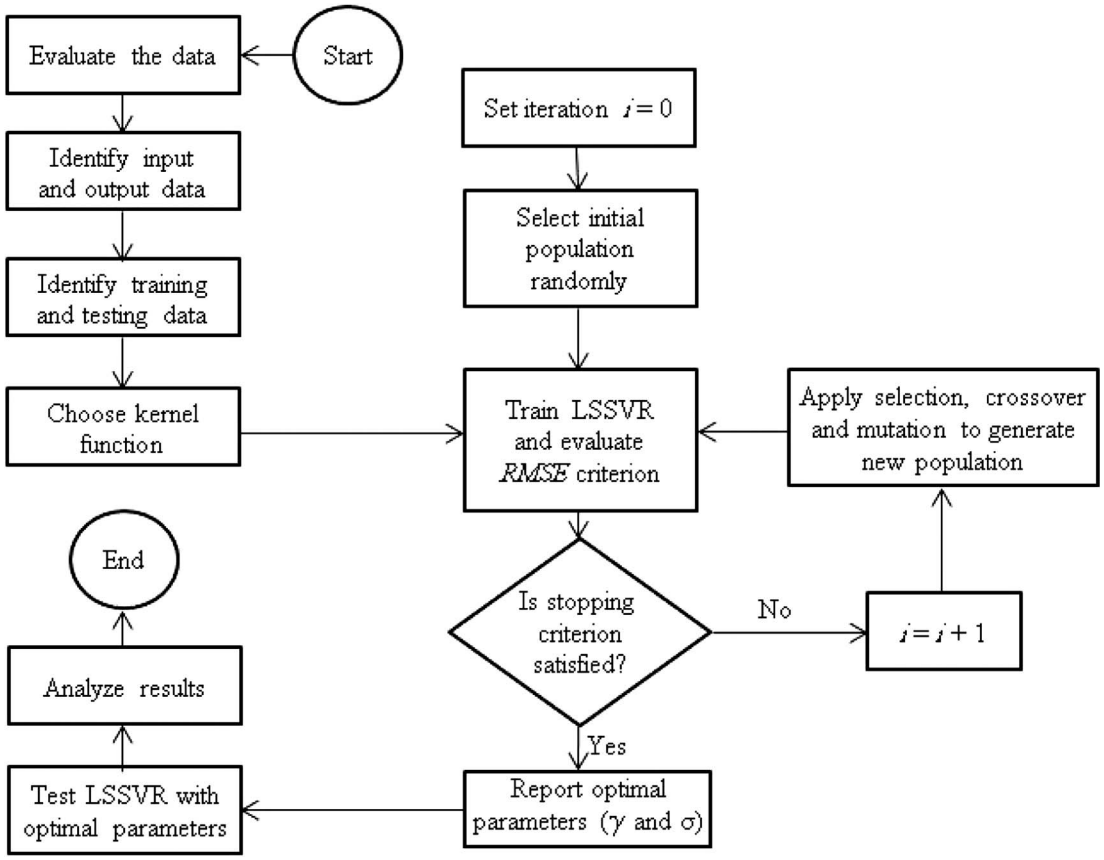

Fig. 2. Flowchart of the GA-LSSVR algorithm

J. Environ. Eng., 2017, 143(7): 04017021

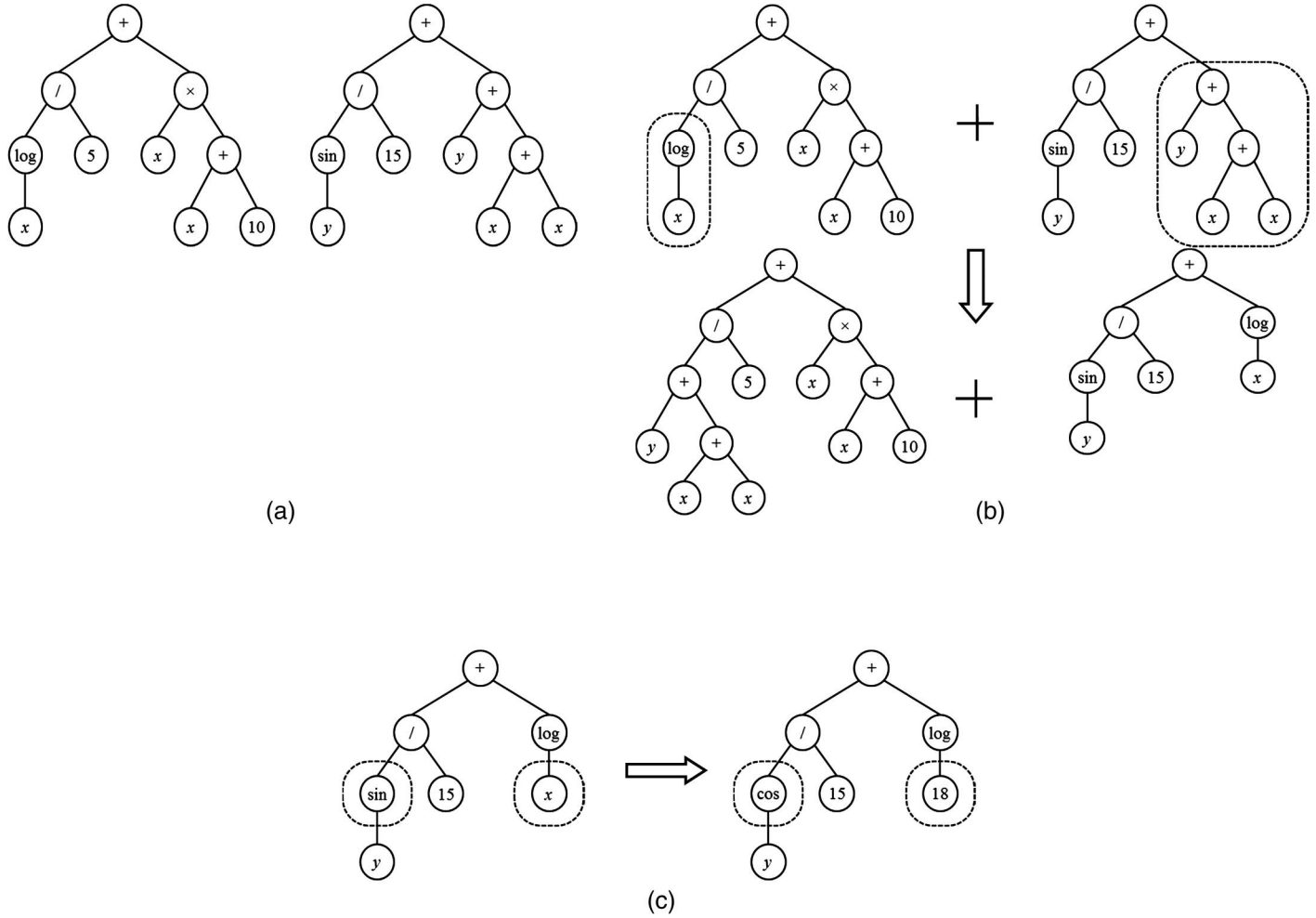

Fig. 3. Genetic programming construction: (a) two samples of tree-structured solutions; (b) crossover operator; (c) mutation operator

where y_t and $\hat{y}_t$ = observed and calculated data at time step t, respectively; and $\bar{y}$ and $\hat{\bar{y}}$ = average of observed and computed data, respectively. The diagnostic variables in Eqs. (8)–(10) were applied by Wang et al. (2009), Rajaee et al. (2009), and Orouji et al. (2013), among others.

Genetic Programming

Genetic programming is a variation of the GA introduced by Koza (1990). It is based on Darwin's theory of evolution (Aytek and Alp 2008). Genetic programming is a branching algorithm in which each branch issues from a set of input variables and main operators (functions) (Orouji et al. 2013). From the input variables and objective functions the GP creates a number of solutions in a tree-like structure and develops the appropriate solution to an optimization problem by comparing the results of consecutive iterations. Generally, GP applies four steps to find an optimal solution. These steps are (Miller and Thomson 2000):

1. Generate an initial population of random decision variables, functions, and constant numbers as the terminal set;
2. Evaluate each decision tree in the population by applying fitness functions to assess how well a decision tree solves an optimization problem;
3. Generate a new population using genetic operators such as copying the best existing tree, crossover, mutation, and reproduction; and
4. The best decision tree that appeared at the time of reaching a termination criterion is specified as the GP result.

A random set of decision trees is generated in the first iteration. Each tree employs a number of nodes and branches for each relation. All terminal and functional members are placed in nodes and relate to each other by branches. In this method, each tree expresses a simple or complicated mathematical relation. Two samples of tree-structured solutions in GP are shown in Figs. 3(a and b). In these structures, $\{x, 5, 10\}$ and $\{x, y, 15\}$ are the terminal sets and $\{\log, +, /\}$ and $\{\sin, +, /\}$ are the functional sets. Next, the corresponding objective function of each tree is calculated. The objective function value measures the fitness of each tree. Trees are chosen by selection operators according to their fitness values. Crossover is a genetic operator in which two trees are randomly selected to create new, fitter, trees that replace the parent trees. Finally, the other genetic operator, i.e., mutation, randomly replaces the initial function variables with random values. The crossover and mutation operators in the GP process are presented in Figs. 3(b and c), respectively. The outputs from these operators are the generated trees and become the GP inputs (the initial trees) for the next iteration. The GP searching process is continued until finding a solution.

Principal Component Analysis

Principal component analysis selects a subset of variables among a large set of regressor variables such that the subset of regressors

J. Environ. Eng., 2017, 143(7): 04017021

explains most of the variability of the dependent variable in a multiple regression problem (Noori et al. 2011). The main purpose of PCA is to reduce the number of predictor variables and to detect structure in the statistical relations that might exist between variables. The PCA reduces the complexity of variables and provides a clearer understanding of variables when the analyst is faced with a large number of data (Camdevyren et al. 2005).

In PCA, linear combinations of p initial variables $(x_1, x_2, \ldots, x_p)$ are created to produce p principal components $(\mathrm{PC}_1, \mathrm{PC}_2, \ldots, \mathrm{PC}_p)$. Each principal component is expressed by Eq. (11) (Johnson and Wichern 1982)

$$\begin{aligned} \mathrm{PC}_1 &= w_{11}x_1 + w_{12}x_2 + \cdots + w_{1p}x_p \\ \mathrm{PC}_2 &= w_{21}x_1 + w_{22}x_2 + \cdots + w_{2p}x_p \\ &\vdots \\ \mathrm{PC}_p &= w_{p1}x_1 + w_{p2}x_2 + \cdots + w_{pp}x_p \end{aligned} \tag{11}$$

in which $\mathrm{PC}_i = i$th principal component; w_{ij} = coefficient of the ith principal component and the jth initial variable; and $x_i = i$th initial variable. In PCA the w_{ij} coefficient is estimated in such a way that the first principal component (PC_1) measures the largest possible variance, and the second principal component (PC_2) measures the largest possible variance not accounted for by the first principle component. The PCA process is continued until the last principle component (PC_p) completes the entire variance. The PCA coefficients satisfy the following relations:

$$i = 1, \ldots, p w_{i1}^2 + w_{i2}^2 + \cdots + w_{ip}^2 = 1 \tag{12}$$

$$\forall i \neq j \; w_{i1}w_{j1} + w_{i2}w_{j2} + \cdots + w_{ip}w_{jp} = 0 \tag{13}$$

The calculation of the w_{ij} coefficients is explained by Tabachnick and Fidell (2001), Ouyang (2005), and Noori et al. (2008), among others.

Study Area

The water-quality data from Astane station located on the Sefidrood River in northern Iran served for modeling of water-quality parameters. The length of the river and the drainage area of the Sefidrood River basin are approximately 670 km and 13,450 km^2, respectively. This river discharges into the Caspian Sea, by the City of Rasht, Iran. Water-quality parameters considered in this paper are Na^+, K^+, Mg^{2+}, SO_4^{2-}, Cl^-, pH, EC, and TDS. The location of the study area, Astane station, and the Sefidrood River are shown in Fig. 4.

The training (calibration) and testing data sets represented 70 and 30% of the total existing water-quality data. The training data set includes the time period 1985–1998 (154 months), and the period of 2000–2005 (69 months) was chosen for the testing data set. The time step of the training and testing data is monthly. The selected time series for the training and testing data sets do not have missing values. The statistical properties of training and testing data sets are listed in Table 1.

Selection of Inputs and Determination of Model Structure

In this study each water-quality parameter is modeled at time step t based on the values of the same water quality at previous time steps $(t-1, t-2, t-3)$ and the values of other water-quality parameters

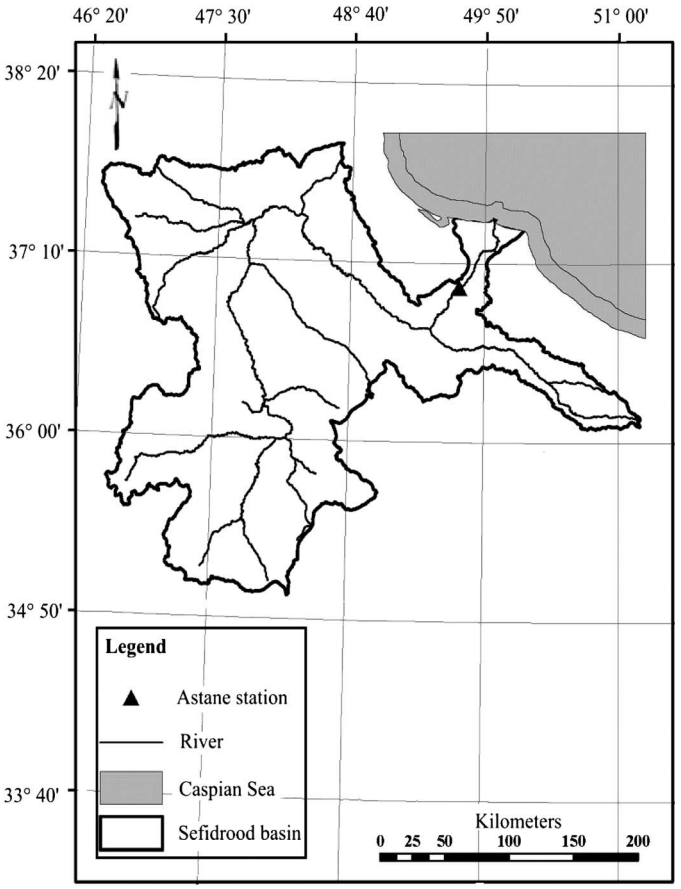

Fig. 4. Locations of the Sefidrood River basin and the Astane station

including river flow discharge at the current and previous time steps $(t, t-1, t-2, t-3)$. Each time step t refers to 1 month.

The scale effect of different input variables was removed by performing standardization, which leads to dimensionless quantities. An input variable was standardized by subtracting its mean from its raw value and the difference was divided by the variable's standard deviation. Next, the principal components of the standardized input matrix were obtained by PCA. The cumulative percentage of the variance of each principal component is presented in Fig. 5. In this study 10 principal components were input to the GA-LSSVR algorithm and GP method because, according to Fig. 5, 10 calculated principal components explain 99.99% of the variance of the input matrix, and consequently, the next 25 components only justify 0.01% of the variance of the input matrix.

Results of the Sensitivity Analysis

The sensitivity of the LSSVR method to coefficients selection was evaluated by means of a sensitivity analysis of the coefficients γ and σ. The interval 0.01 to 100 was considered for both coefficients and the LSSVR method was implemented for the selected coefficients. The calculated results from sensitivity analysis of the LSSVR method on water-quality parameter Cl^- are listed in Table 2.

According to Table 2, when γ and σ are nearly equal to 10 and 100, respectively, the RMSE value decreases drastically, but the LSSVR method does not provide acceptable results for other values of these two coefficients. Another trend is that the accuracy of estimation of the LSSVR method increases at first by increasing the

Table 1. Statistical Characteristics of the Training and Testing Data Sets of the Sefidrood River Water-Quality Parameters at Astane Station

Parameter	Data set	Minimum	Average	Maximum	Standard deviation	Coefficient of variation (%)
Na^+ (meq/L)	Training	0.05	5.78	15.65	2.64	45.72
	Testing	0.19	3.74	10.13	2.19	58.57
K^+ (meq/L)	Training	0.01	0.09	0.22	0.04	49.14
	Testing	0.01	0.08	0.15	0.03	39.76
Mg^{2+} (meq/L)	Training	0.20	2.31	5.50	0.93	40.37
	Testing	0.38	2.08	5.52	1.04	49.77
SO_4^{2-} (meq/L)	Training	0.34	2.83	7.38	1.27	44.88
	Testing	0.21	1.93	4.12	1.00	51.89
Cl^- (meq/L)	Training	0.80	5.72	15.60	2.65	46.37
	Testing	0.20	4.27	10.90	2.45	57.31
pH	Training	6.70	7.74	8.80	0.37	4.81
	Testing	6.42	7.54	8.42	0.46	6.07
EC (μs/cm)	Training	244.43	1,221.70	2,336.00	381.49	31.23
	Testing	252.00	1,024.41	2,018.00	393.96	38.46
TDS (mg/L)	Training	263.00	772.59	1,472.00	232.36	30.08
	Testing	159.00	641.75	1,271.00	243.81	37.99

values of the coefficients, and then decreases. For example, given a constant value of γ equal to 1, by increasing σ up to 10 the accuracy increases, but then it decreases. It is seen in Table 2, however, that R^2 increases monotonically as σ increases. In addition, it follows from Table 2 that the LSSVR method achieved better results than the average of Cl^- only 40% of the time. Therefore, the results shown in Table 2 indicate high sensitivity of the LSSVR method to the selection of coefficients γ and σ. It is concluded that there is an optimal point, and that the closer the values of coefficients to that optimal point are, the higher the accuracy of estimation. Furthermore, the accuracy of estimation decreases with increasing distance

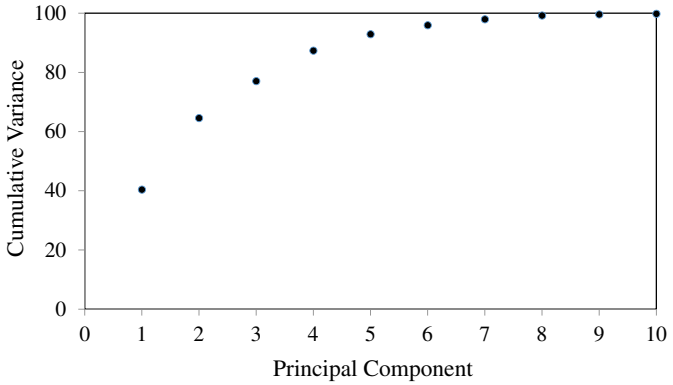

Fig. 5. Percentage of the cumulative variance of the principal components

Table 2. Results of Statistical Criteria for the Testing Data Sets Considering Different Values of γ and σ for Cl^-

		γ				
Σ	Statistical criteria	0.01	0.1	1	10	100
0.01	RMSE	2.84	2.84	2.84	2.84	2.84
	R^2	0.11	0.1	0.12	0.11	0.13
	NS	−0.35	−0.35	−0.35	−0.35	−0.34
0.1	RMSE	2.84	2.83	2.8	2.78	2.78
	R^2	0.35	0.35	0.34	0.34	0.34
	NS	−0.34	−0.33	−0.31	−0.29	−0.29
1	RMSE	2.79	2.51	2.07	1.83	1.96
	R^2	0.61	0.63	0.7	0.72	0.64
	NS	−0.29	−0.05	0.27	0.44	0.35
10	RMSE	2.67	1.82	0.92	0.82	1.11
	R^2	0.95	0.95	0.95	0.95	0.93
	NS	−0.19	0.44	0.85	0.88	0.79
100	RMSE	2.8	2.52	1.37	0.75	0.93
	R^2	0.96	0.95	0.96	0.96	0.88
	NS	−0.31	−0.06	0.68	0.9	0.9

Table 3. Optimal Values of the Coefficients γ and σ Calculated with the GA

Parameter	σ	γ
Na^+	58.31	48.75
K^+	3.03	13.43
Mg^{2+}	18.88	69.87
SO_4^{2-}	15.61	57.61
Cl^-	15.77	58.04
pH	38.92	0.58
EC	21.09	75.98
TDS	17.17	66.75

Table 4. Calculated Statistical Results with the GA-LSSVR Algorithm and the GP Method for the Training and Testing Data Sets

		Training			Testing		
Parameter	Method	RMSE	R^2	NS	RMSE	R^2	NS
Na^+ (meq/L)	GA-LSSVR	0.92	0.93	0.87	1.20	0.94	0.69
	GP	1.31	0.87	0.75	1.19	0.91	0.72
K^+ (meq/L)	GA-LSSVR	0.03	0.63	0.36	0.02	0.64	0.38
	GP	0.03	0.54	0.25	0.02	0.64	0.40
Mg^{2+} (meq/L)	GA-LSSVR	0.64	0.72	0.52	0.66	0.79	0.58
	GP	0.69	0.67	0.44	0.75	0.75	0.47
SO_4^{2-} (meq/L)	GA-LSSVR	0.76	0.79	0.63	0.67	0.84	0.54
	GP	0.94	0.68	0.45	0.82	0.78	0.32
Cl^- (meq/L)	GA-LSSVR	0.86	0.94	0.89	0.71	0.96	0.91
	GP	1.31	0.87	0.75	1.15	0.92	0.79
pH	GA-LSSVR	0.02	0.99	0.99	0.31	0.75	0.52
	GP	0.34	0.41	0.16	0.37	0.68	0.35
EC (μs/cm)	GA-LSSVR	69.08	0.98	0.96	61.79	0.98	0.97
	GP	74.32	0.98	0.96	72.86	0.99	0.96
TDS (mg/L)	GA-LSSVR	39.35	0.98	0.97	55.84	0.98	0.94
	GP	57.41	0.97	0.94	44.41	0.99	0.96

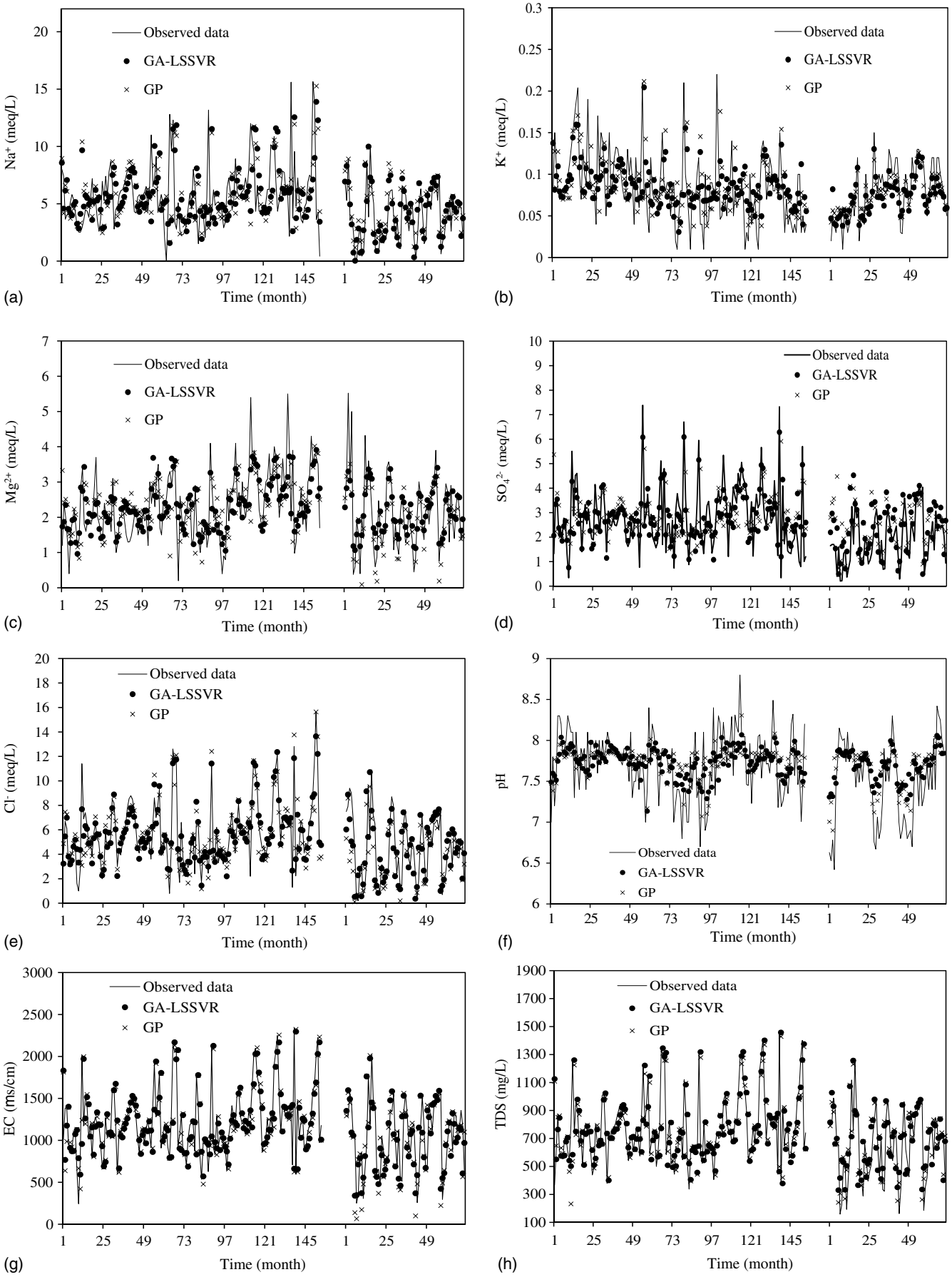

Fig. 6. Observed water-quality parameters and calculated water-quality parameters using the GA-LSSVR algorithm and the GP method: (a) Na^+; (b) K^+; (c) Mg^{2+}; (d) SO_4^{2-}; (e) Cl^-; (f) pH; (g) EC; (h) TDS

from the optimal point. This optimal point can be obtained with the GA.

According to Table 2, LSSVR is very sensitive to choosing the coefficients of γ and σ so that choosing the best coefficients (10, 100) instead of the worst coefficients (0.1, 0.01) for (γ, σ) in Table 2 improves the results of RMSE, R^2, and NS by 73, 860, and 360%, respectively. Selecting the LSSVR coefficient manually is a slow and computationally burdensome endeavor. Therefore, the GA-LSSVR algorithm is herein developed to find the optimal coefficients automatically, thus making the application of the developed algorithm fast and practical.

Results of the GA-LSSVR Algorithm and the GP Method

The calculated optimal coefficients with the GA-LSSVR algorithm are listed in Table 3, where it is seen that the coefficient γ varies between 0.58 (corresponding to pH) and 75.98 (corresponding to EC), and the coefficient σ ranges between 3.03 (corresponding to K^+) and 58.31 (corresponding to Na^+). The ranges of the coefficients of the LSSVR method indicate that the coefficients γ and σ vary for each water-quality parameter.

Results calculated with the GA-LSSVR algorithm for the training and testing data sets with respect to different water-quality parameters are listed in Table 4. The GP method results are also displayed in Table 4 for comparison purposes.

The results for the statistical criteria of the GA-LSSVR shown in Table 4 indicate that the water-quality parameters EC and TDS have more R^2 equal to 0.98 than the other water-quality parameters. Electrical conductivity has the largest value of NS (equal to 0.97) among the parameters. Therefore, the GA-LSSVR algorithm can model EC more accurately than the other parameters listed in Table 4. In contrast, the GA-LSSVR algorithm cannot model K^+ as accurately as other water-quality parameters given that it reached R^2 and NS equal to 0.64 and 0.38, respectively, which are the worst values among those for other water-quality parameters.

The results in Table 4 establish that GP can model TDS more accurately than other parameters given that it features R^2 and NS equal to 0.99 and 0.96, respectively. On the other hand, GP cannot model K^+ and SO_4^{2-} as accurately as other water-quality parameters due of its low R^2 and NS, respectively. The value of R^2 for K^+ is equal to 0.64, which is the worst value of R^2 among the water-quality parameters, and the value of NS for SO_4^{2-} is equal to 0.32, which is the worst among the water-quality parameters.

It is seen in Table 4 that the GA-LSSVR algorithm provides better results than the GP method because the GA-LSSVR algorithm models water-quality parameters Mg^{2+}, SO_4^{2-}, Cl^-, pH, and EC with lower RMSE values. According to other criteria including R^2 and NS, similar results were obtained. More precisely, the GA-LSSVR algorithm has 13, 16, 34, 18, and 15% lower RMSE values than the GP method for Mg^{2+}, SO_4^{2-}, Cl^-, pH, and EC, respectively. Concerning the water-quality parameters Na^+, K^+, and TDS, the differences between the GP method and the GA-LSSVR algorithm are negligible, the GP method being slightly more accurate, so that the GP method achieved improvements in the estimation of the NS equal to 4, 5, and 2% for the quality parameters Na^+, K^+, and TDS, respectively, compared with the GA-LSSVR algorithm. The RMSE values from the GP method for water-quality parameters Na^+ and TDS exhibited improvements equal to 0.8 and 20%, respectively, compared with the GA-LSSVR algorithm. In addition, R^2 values from GA-LSSVR algorithm were improved 3, 5, 8, 4, and 10% relative to the GP method for the water-quality parameters of Na^+, K^+, Mg^{2+}, SO_4^{2-}, Cl^-, and pH, respectively. The GP method improves R^2 by 1% for water-quality parameters EC and TDS relative to the GA-LSSVR.

In general, given the positive values of the NS statistic, it can be concluded that the GA-LSSVR algorithm and GP method are able to model water-quality parameters well, as listed in Table 4. The modeling results for water-quality parameters achieved by the GA-LSSVR algorithm and the GP method are depicted in Fig. 6.

Orouji et al. (2013) modeled water-quality parameters including Na^+, K^+, Mg^{2+}, SO_4^{2-}, Cl^-, pH, EC, and TDS using default (predefined) models and the GP model in the Sefidrood River at Astane station. The RMSE values of the best models for the testing data sets associated with the cited water-quality parameters were 2.1, 0.02, 0.85, 0.93, 2.18, 0.33, 404.15, and 246.15, respectively. By comparing the results of Orouji et al. (2013) and Table 4, it is concluded that the selection of model inputs relying on the PCA method improves the estimation results from 6 to 84%. In other words, the accuracy of modeling water-quality parameters Na^+, K^+, Mg^{2+}, SO_4^{2-}, Cl^-, pH, EC, and TDS in the Sefidrood River at Astane station were improved by 42, 0, 22, 33, 47, 6, 84, and 77%, respectively, by applying the GA-LSSVR algorithm.

Given that monitoring of water-quality parameters in rivers is time-consuming and expensive, the GA-LSSVR algorithm can be efficiently implemented to estimate water-quality parameters with associated reduction in cost.

Conclusion

The GA, which is a widely used optimization algorithm, was employed in this study to optimize the coefficients of the LSSVR method. A sensitivity analysis of the coefficients of the LSSVR method established that the accuracy of this method depends strongly on the correct selection of these coefficients. An optimization algorithm is required to calculate the LSSVR coefficients. The GA-LSSVR algorithm was developed and applied to find the coefficients of the LSSVR method automatically. The GP method, which is well known and widely used in data-driven studies, was also utilized for comparison purposes. The results indicate the superiority of the GA-LSSVR algorithm over the GP method in such a way that the use of the GA-LSSVR algorithm led to improvements in several statistics including R^2, RMSE, and the NS statistic of water-quality parameters respectively equal to 3, 14, and 21% compared to the GP.

References

Ahmadi, M., Bozorg-Haddad, O., and Loáiciga, H. A. (2015). "Adaptive reservoir operation rules under climatic change." *Water Resour. Manage.*, 29(4), 1247–1266.

Akbari-Alashti, H., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A. (2014). "Multi-reservoir real-time operation rules: A new genetic programming approach." *Proc. Inst. Civ. Eng.: Water Manage.*, 167(10), 561–576.

Anandhi, A., Srinivas, V. V., Nanjundiah, R. S., and Nagesh Kumar, D. (2008). "Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine." *Int. J. Climatol.*, 28(3), 401–420.

Aytek, A., and Alp, M. (2008). "An application of artificial intelligence for rainfall-runoff modeling." *J. Earth Syst. Sci.*, 117(2), 145–155.

Azamathulla, H. M. (2012). "Gene expression programming for prediction of scour depth downstream of sills." *J. Hydrol.*, 460–461, 156–159.

Azamathulla, H. M., and Ghani, A. A. (2011). "Genetic programming for predicting longitudinal dispersion coefficients in streams." *Water Resour. Manage.*, 25(6), 1537–1544.

Azamathulla, H. M., Ghani, A. A., Leow, C. S., Chang, C. K., and Zakaria, N. A. (2011). "Gene-expression programming for the development of a

stage-discharge curve of the Pahang river." *Water Resour. Manage.*, 25(11), 2901–2916.
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.
Bhagwat, P. P., and Maity, R. (2013). "Hydroclimatic streamflow prediction using least square-support vector regression." *ISH J. Hydraul. Eng.*, 19(3), 320–328.
Bozorg-Haddad, O., Afshar, A., and Mariño, M. A. (2011). "Multireservoir optimisation in discrete and continuous domains." *Proc. Inst. Civ. Eng.: Water Manage.*, 164(2), 57–72.
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 layouts and design optimization in protection of flood areas." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000864, 04015004.
Bozorg-Haddad, O., Fallah-Mehdipour, E., Mirzaei-Nodoushan, F., and Mariño, M. A. (2014a). "Discussion of 'GA-based support vector machine model for the prediction of monthly reservoir storage.'" *J. Hydrol. Eng.*, 10.1061/(ASCE)HE.1943-5584.0001086, 1430–1437.
Bozorg-Haddad, O., Moravej, M., and Loáiciga, H. (2014b). "Application of the water cycle algorithm to the optimal operation of reservoir systems." *J. Irrig. Drain. Eng.*, 10.1061/(ASCE)IR.1943-4774.0000832, 04014064.
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.
Camdevyren, H., Demyr, N., Kanik, A., and Keskyn, S. (2005). "Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs." *Ecol. Modell.*, 181(4), 581–589.
Chapra, S. C. (2008). *Surface water-quality modeling*, Waveland Press, Long Grove, IL.
Chiu, D. Y., and 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." *Exp. Syst. Appl.*, 36(2), 1240–1248.
Deb, K., and Deb, D. (2014). "Analysing mutation schemes for real-parameter genetic algorithms." *Int. J. Artif. Intell. Soft Comput.*, 4(1), 1–28.
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.
Fallah-Mehdipour, E., Bozorg-Haddad, O., and Mariño, M. A. (2014). "Genetic programming in groundwater modeling." *J. Hydrol. Eng.*, 10.1061/(ASCE)HE.1943-5584.0000987, 04014031.
Fallah-Mehdipour, E., Bozorg-Haddad, O., Orouji, H., and Mariño, M. A. (2013c). "Application of genetic programming in stage hydrograph routing of open channels." *Water Resour. Manage.*, 27(9), 3261–3272.
Farhangi, M., Bozorg-Haddad, O., and Mariño, M. A. (2012). "Evaluation of simulation and optimization models for WRP with performance indices." *Proc. Inst. Civ. Eng.: Water Manage.*, 165(5), 265–276.
Ghavidel, S. Z. Z., and Montaseri, M. (2014). "Application of different data-driven methods for the prediction of total dissolved solids in the Zarinehroud basin." *Stochastic Environ. Res. Risk Assess.*, 28(8), 2101–2118.
Hashmi, M. Z., Shamseldin, A. Y., and Melville, B. W. (2011). "Statistical downscaling of watershed precipitation using gene expression programming (GEP)." *Environ. Modell. Software*, 26(12), 1639–1646.
Izadifar, Z., and Elshorbagy, A. (2010). "Prediction of hourly actual evapotranspiration using neural networks, genetic programming, and statistical models." *Hydrol. Process.*, 24(23), 3413–3425.
Jahandideh-Tehrani, M., Bozorg-Haddad, O., and Mariño, M. A. (2015). "Hydropower reservoir management under climate change: The Karoon reservoir system." *Water Resour. Manage.*, 29(3), 749–770.
Johnson, R. A., and Wichern, D. W. (1982). *Applied multivariate statistical analysis*, Prentice Hall, Englewood Cliffs, NJ.
Kisi, O., Dailr, A. H., Cimen, M., and Shiri, J. (2012). "Suspended sediment modeling using genetic programming and soft computing techniques." *J. Hydrol.*, 450–451, 48–58.
Kisi, O., and Shiri, J. (2010). "A comparison of genetic programming and ANFIS in forecasting daily, monthly and daily stremflows." *Proc., Int. Symp. on Innovations in Intelligent Systems and Applications*, INISTA, Kayseri, Turkey.
Koza, J. R. (1990). "Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems." Dept. of Computer Science, Stanford Univ., Stanford, CA.
Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., and Wei, Y. (2013). "A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction." *Math. Comput. Modell.*, 58(3), 458–465.
Maity, R., Bhagwat, P. P., and Bhatnagar, A. (2010). "Potential of support vector regression for prediction of monthly streamflow using endogenous property." *Hydrol. Process.*, 24(7), 917–923.
Miller, J. F., and Thomson, P. (2000). "Cartesian genetic programming." *Genetic programming*, Springer, Berlin, 121–132.
Noori, R., et al. (2011). "Assessment of input variables determination on the SVM model performance using PCA, gamma test, and forward selection techniques for monthly stream flow prediction." *J. Hydrol.*, 401(3), 177–189.
Noori, R., Ashrafi, K., and Ajdarpour, A. (2008). "Comparison of ANN and PCA based multivariate linear regression applied to predict the daily average concentration of CO: A case study of Tehran." *J. Phys. Earth Space*, 34(1), 135–152.
Noori, R., Karbassi, A., and Salman Sabahi, M. (2010). "Evaluation of PCA and gamma test techniques on ANN operation for weekly solid waste prediction." *J. Environ. Manage.*, 91(3), 767–771.
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. (2014a). "Extraction of decision alternatives in project management: Application of hybrid PSO-SFLA." *J. Manage. Eng.*, 10.1061/(ASCE)ME.1943-5479.0000186, 50–59.
Orouji, H., Bozorg-Haddad, O., Fallah-Mehdipour, E., and Mariño, M. A. (2014b). "Flood routing in branched river by genetic programming." *Proc. Inst. Civ. Eng.: Water Manage.*, 167(2), 115–123.
Ouyang, Y. (2005). "Evaluation of river water quality monitoring stations by principal component analysis." *Water Res.*, 39(12), 2621–2635.
Raghavendra, N. S., and Deka, P. C. (2014). "Support vector machine applications in the field of hydrology: A review." *Appl. Soft Comput.*, 19, 372–386.
Rajaee, T., Mirbagheri, S. A., Zounemat-Kermani, M., and Nourani, V. (2009). "Daily suspended sediment concentration simulation using ANN and neuro-fuzzy models." *Sci. Total Environ.*, 407(17), 4916–4927.
Singh, K. P., Basant, N., and Gupta, S. (2011). "Support vector machines in water quality management." *Anal. Chim. Acta*, 703(2), 152–162.
Soleimani, S., Bozorg-Haddad, O., Saadatpour, M., and Loáiciga, H. A. (2016). "Optimal selective withdrawal rules using a coupled data mining model and genetic algorithm." *J. Water Resour. Plann. Manage.*, 10.1061/(ASCE)WR.1943-5452.0000717, 04016064.
Su, J., Wang, X., Liang, Y., and Chen, B. (2013). "GA-based support vector machine model for the prediction of monthly reservoir storage." *J. Hydrol. Eng.*, 10.1061/(ASCE)HE.1943-5584.0000915, 1430–1437.
Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., and Vandewalle, J. (2002). *Least squares support vector machines*, World Scientific, Singapore.

Tabachnick, B. G., and Fidell, L. S. (2001). *Using multivariate statistics*, Pearson.
Tan, G., Yan, J., Gao, C., and Yang, S. (2012). "Prediction of water quality time series data based on least squares support vector machine." *Proc. Eng.*, 31, 1194–1199.
Tripathi, S., Srinivas, V. V., and Nanjundiah, R. S. (2006). "Downscaling of precipitation for climate change scenarios: A support vector machine approach." *J. Hydrol.*, 330(3), 621–640.
Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer, 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), 294–306.
Yoon, H., Jun, S. C., Hyun, Y., Bae, G. O., and Lee, K. 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.
Yunrong, X., and Liangzhong, J. (2009). "Water quality prediction using LS-SVM with particle swarm optimization." *2nd Int. Workshop on Knowledge Discovery and Data Mining*, IEEE, New York.