Caspian J. Environ. Sci. 2017, Vol. 15 No. 4 pp. 385~399

CJES
Caspian Journal of Environmental Sciences

[Research]

Optimization of sediment rating curve coefficients using evolutionary algorithms and unsupervised artificial neural network

Tabatabaei M.*, Salehpour Jam A.

Soil Conservation and Watershed Management Research Institute, Agricultural Research, Education and Extension Organization (AREEO), Tehran, Iran

** Corresponding author's E-mail: Taba1345@hotmail.com*

(Received: July 05. 2017 Accepted: Nov. 28. 2017)

ABSTRACT

Sediment rating curve (SRC) is a conventional and a common regression model in estimating suspended sediment load (SSL) of flow discharge. However, in most cases the data log-transformation in SRC models causing a bias which underestimates SSL prediction. In this study, using the daily stream flow and suspended sediment load data from Shalman hydrometric station on Shalmanroud River, Guilan Province, Iran, SRC equation was derived, and then, using evolutionary algorithms (genetic algorithm and particle swarm optimization algorithm) it was calibrated again. Worth mentioning, before model calibration, to increase the generalization power of the models, using self-organizing map (an unsupervised artificial neural network for data clustering), the data were clustered and then by data sampling, they were classified into two homogeneous groups (calibration and test data set). The results showed that evolutionary algorithms are appropriate methods for optimizing coefficients of SRC model and their results are much more favorable than those of the conventional SRC models or SRC models corrected by correction factors. So that, the sediment rating curve models calibrated with evolutionary algorithms, by reducing the RMSE of the test data set of 5754.02 ton day^{-1} (in the initial SRC model) to 1681.21 ton day^{-1} (in the calibrated models by evolutionary algorithms) increased the accuracy of suspended sediment load estimation at a rate of 4072.81 ton day^{-1}. In total, using evolutionary algorithms in calibrating SRC models prevents data log-transformation and use of correction factors along with increasing in the accuracy of molding results.

Key words: Clustering, Genetic and PSO Algorithms, Sediment Rating Curve, Self-Organizing Map, Suspended Sediment Load.

INTRODUCTION

It is necessary to have adequate up-to-date information about the suspended sediment load (SSL) of rivers and monitor them continually in order to be aware of the watershed sediment yield condition, the amount of erosion as well as changes in the river bed and river bank, the quality of water, along with optimum design and favorable performance of water resource structures (Tayfur 2012; Nourani *et al.* 2016; Buyukyildiz & Kumcu, 2017; Vercruysse *et al*. 2017; Sarkar *et al*. 2017; Salehpour Jam *et al.* 2017). Regarding the existing limitations (cost of sampling, time, etc.), the SSL is often estimated indirectly using sediment rating curve (SRC) model. The standard model of SRC is obtained through the following exponential regression equation (Ulke *et al.* 2009):

$$SSL_{(t)} = aQ_{(t)}^{b} \quad (1)$$

Where, $Q_{(t)}$ is the mean flow discharge ($m^3 s^{-1}$); SSL (t) is the suspended sediment discharge (ton day^{-1}); a and b are the constant coefficients of the regression equation. In equation 1, SSC (mg l^{-1}) can be used instead of SSL (ton day^{-1}). To use the SRC regression model, the coefficients (a and b) should be calculated optimum. This is firstly done through the taking logarithm of variables of flow discharge and sediment discharge as well as formulating a linear regression equation between them. Then, the linear regression coefficients are calculated using least square method. Once the coefficients and sediment discharge are calculated, the obtained values for the sediment discharge should be back-transformed (an anti-log is taken of them) in order to be used. Studies have shown that the distribution of remaining values (the difference between the observed and computed values of sediment discharge) in this way is not normal, and the mean distribution is greater than zero (Kao *et al.* 2005). In other words, when calculating a and b coefficients, a kind of bias appears in the SRC regression model and makes the estimated values of SSL lower than its corresponding observed values (Ferguson 1986). This problem is most obvious in flood discharges and causes more errors. To correct the bias resulting from the logarithmic transformation, different correction factors have been introduced so far (FAO, Quasi-Maximum Likelihood Estimator, Minimum Variance Unbiased Estimator, etc.), and all of them aim at increasing the values calculated through SRC model. However, these factors sometimes cause another bias in the form of an overestimation besides making the results with the same data zero (Kao *et al.* 2005). In recent years, the application of computational intelligence methods in estimation of environmental variables such as suspended sediment load and modeling the complex hydrological processes, such as rainfall-runoff has been rapidly rising. (Kalteh 2008; Gholami *et al*. 2015; Chen & Chau 2016; Kisi & Zounemat-Kermani 2016; Buyukyildiz & Yurdagul Kumcu 2017). Also, meta-heuristic algorithms (or evolutionary algorithms) such as genetic algorithms (GAs) and particle swarm optimization (PSO) have been commonly used in solving problems related to water resource engineering. Kisi *et al.* (2017) used PSO and differential evolution (DE) algorithms as training algorithms of ANNs (ANN-PSO and ANN-DE) for modeling groundwater qualitative parameters, i.e., SO_4 and SAR. Cheng *et al.* (2002) could calibrate parameters of a rainfall-runoff model of Xinanjiang watershed automatically with multiple objectives (including time to peak, peak rate, and total volume of flood) using GA and fuzzy algorithm. In another study, Hejazi *et al.* (2008) calibrated parameters of a distributed rainfall-runoff model using multi-objective GA. Tayfur (2009) optimized parameters of some empirical equations and could estimate the longitudinal dispersion coefficient of a river. Kisi *et al.* (2012) used the genetic programming (GP) model in order to estimate the amount of daily suspended sediment in two stations in the Cumberland River in America. Kuok *et al.* (2010) applied the PSO algorithm to optimize parameters of neural network model of daily rainfall-runoff in Sungai Bedup watershed, Malaysia. They showed that the neural network training through the above method was successful. Guo & Wang (2010) used radial basis function (RBF) neural network whose parameters was optimized based on PSO algorithm to estimate SSL of Yangtze river. In another similar research, in relation to the application of evolutionary algorithms in the modeling and monitoring of water quality of rivers, Altunkaynak (2009) could optimize SRC coefficients of Mississippi River located in St. Louis, MO using GA. The results of the study showed the priority of SRC model optimized by GA over its conventional model. In another similar study conducted by Mohammad Rezapour *et al.* (2016) using genetic and PSO algorithms, the relationship between flow discharge and sediment discharge was optimized. The comparison of the results of models showed that the SRC models optimized by evolutionary algorithms had better results than the conventional SRC models. Swain and

Sahooh (2017) also used the combination of the regression methods and genetic algorithm (GA) to optimize three regression models between turbidity concentration (Tu) and Landsat surface reflectance (Ls) (Tu-Ls); total suspended solid (TSS) as well as Tu (Tss-Tu) and six heavy metals (HV) and also TSS (HV-TSS). In another field, to predict non-deposition sediment transport, Ebtehaj & Bonakdar (2016) used PSO and imperialist competitive algorithms (ICA) for estimating the densimetric Froude number. The results showed that the algorithm ICA is superior to the algorithm PSO. Clustering and sampling them play an important role in building similar homogenous data sets (such as calibration, cross-validation, and test data set) for data-driven models (such as regression, neural network, and Neuro-fuzzy models). The failure to use similar homogenous data in the mentioned three sections has much direct effect on the precision and final efficiency of designing models and reduces its power generalization (May *et al.* 2010). In the present study, self-organizing map clustering method (SOM) was used to build two similar homogenous data sets of calibration and test the models regarding drastic changes in sediment discharge data during the statistical period. Regarding the foregoing, the objectives and innovations of this study are summarized as follows:

A. Estimation of daily SSL of Shalman River using the traditional SRC model and the SRC model modified by traditional correction factors.

B. Optimization of SRC model's coefficients using evolutionary algorithms (GA and PSO algorithm) and re-estimation of SSL.

C. Comparison of traditional SRC models (part A) with optimized models (part B) in terms of SSL estimation as accurate as possible.

MATERIALS AND METHODS

In this study, MATLAB 7.11 software was used to implement GA and PSO algorithms, cluster the data, and calculate the cluster validity index. The data were statistically analyzed using SPSS 19 and MATLAB software programs.

The study area and used data

The present study was performed in the Shalman watershed at Shalman hydrometric station, which is located between longitude 49°56'-50°18' E and latitude 36°54'- 37°14' N in Guilan Province, Iran (Fig. 1).

The watershed has the area of 48021 hectares and mean elevation of 522 m above sea level. The data used in this study included 841information records of hydrometric data of instantaneous flow discharge and sediment discharge in Shalman Hydrometric Station during the 34 years (1972-2006). The statistical parameters [mean ($\overline{X}$), standard deviation (S_x), coefficient of variation (C_V), skewness coefficient (C_{sx}), overall minimum (X_{min}) and maximum (X_{max})] of the whole data set in this period are presented in Table 1.

According to the statistical data in Table 1, the sediment discharge has a high skewness and coefficient of variation, as the variation between its maximum and minimum is very high. This result along with other calculated statistics revealed the complexity of SSL modeling of the river.

Preparation of homogenous data for calibrating and evaluating the models

To build the SRC models as accurate as possible, the calibration data of the models should represent the data of the entire statistical period. Moreover, to evaluate the models and its results, the test data should be similar to those of calibration (in terms of statistical parameters) and have the same distribution. To do so, the SOM clustering method was used to cluster the data, and proportional allocation method was used to sample the clusters to prepare two homogenous and similar sets of data (calibration and test data sets). The number of optimal clusters was determined using Davies-Bouldin index. To analyze the results of clustering, besides comparing the statistical parameters (mean, standard deviation, skewness, etc.) together, the similarity of data

distribution (in calibration and evaluation) was examined using Two-Sample Kolmogorov-Smirnov Test (KS). All these stages are briefly described below:

Data clustering using self-organizing map (SOM)

Data clustering is a common method in analysis of statistical data in which similar data are classified into different clusters in a way that the samples in each cluster are similar to one another but different from samples of other clusters (Yar Kiani 2009).

The self-organizing map (SOM) is an unsupervised artificial neural network proposed by Kohonen (1982). One of the most important application of the SOM is its ability in the clustering of data. Generally, SOM networks learn to cluster groups of similar input data from a high dimensional input space in a non-linear fashion onto a low dimensional (most commonly two-dimensional) discrete lattice of neurons in an output layer (Kohonen 2001). The typical structure of an SOM consists of two layers: an input layer and a Kohonen or output layer (Fig. 2).

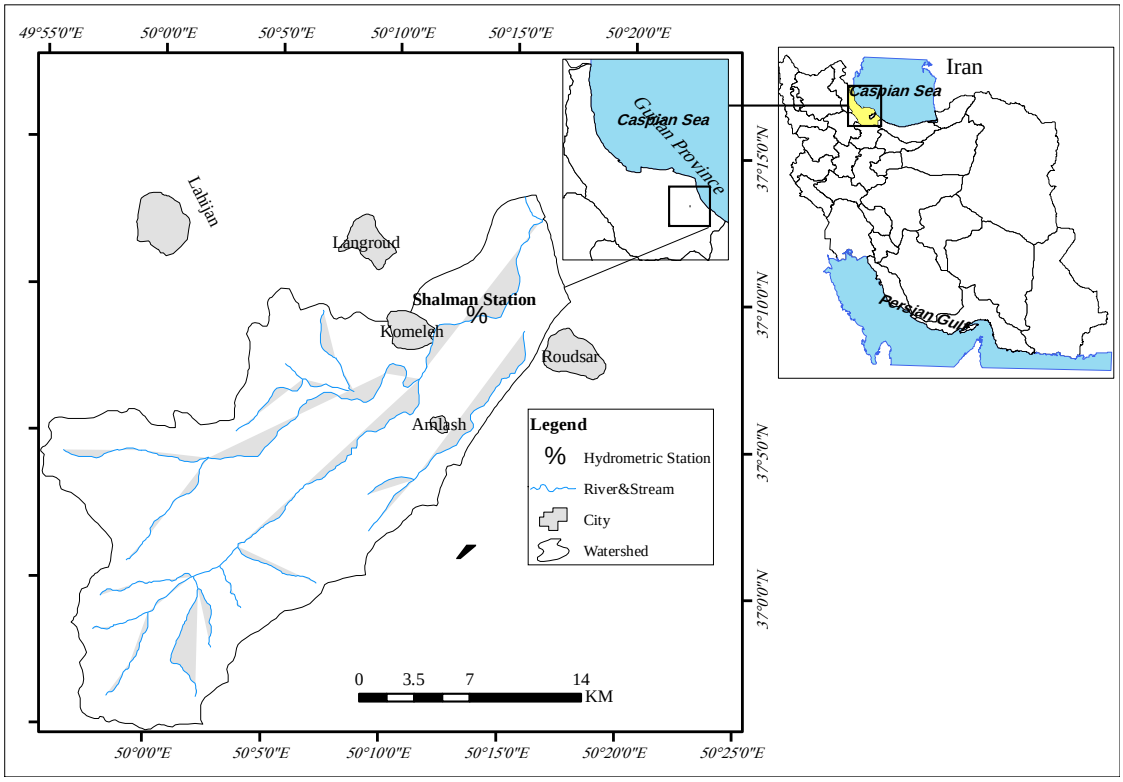

Fig. 1. The location of Shalman watershed and Shalman Hydrometric Station.

Table 1. Statistical characteristics of the data used during the study.

Data Set	Data Type	$\overline{X}$	S_x	X_{max}	X_{min}	C_{sx}	C_v
Whole data	Flow, Q_w ($m^3 s^{-1}$)	12.54	29.84	349.08	0.01	58.55	2.37
	SSL, Q_s (ton day^{-1})	819.5	6525.98	144852.8	0.004	314.08	7.96

In the SOM structure, input layer contains one neuron for each variable (x_i for I = 1,2,..., n) (e.g., flow discharge, suspended sediment load, etc.) in the data set and it is fully connected to the Kohonen layer through adjustable weights (w_{ji} for j = 1,2, ..., m).

The process of network learning is formed of three phases of the competition, co-operation

and adaptation. In competitive phase, by introducing a data pattern (an input vector) to the SOM network, the Euclidean distances of the data to the neurons of output layer are calculated and each neuron of the output layer that has the least distance is selected as a winner neuron or neuron which is the closest neuron to the input vector.

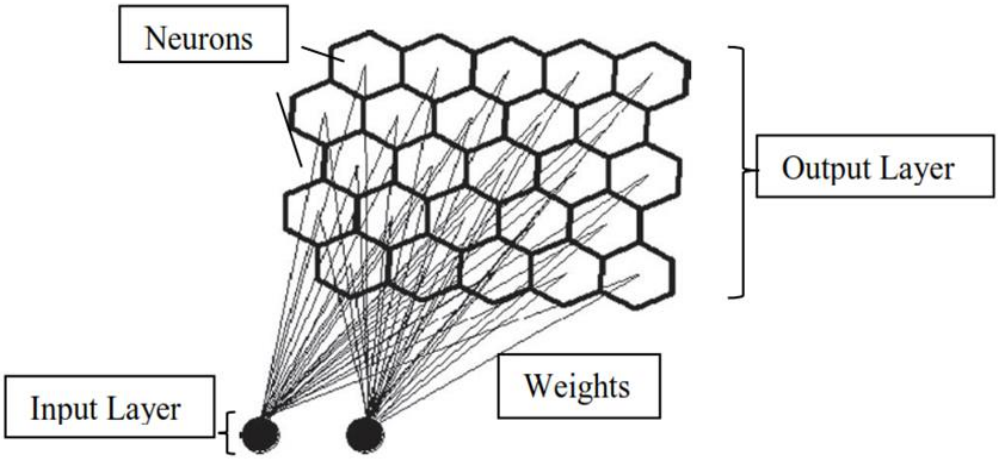

Fig. 2. A 5×5 two-dimensional self-organizing map (modified from Kalteh *et al.* (2008).

This neuron is also called best matching unit (BMU). Notably, at the first, the weight of neurons in the output layer is randomly defined, but during the process of learning, it is more similar to the vector values of input variables. The Euclidean distance is calculated according to the following equation (Bowden *et al.* 2002):

$$D_j = \left|x - w_j\right| = \left[\sum_{i=1}^{N}(x_i - w_{ji})^2\right]^{\frac{1}{2}} \quad , \quad j = 1, 2, \ldots, M \qquad (2)$$

Where: D_j, jth neuron distance of the output layer to x input vector ($X = (x_i; i = 1,2,3, \ldots, N) \in R^n$), N, the number of input vector variables, M, the number of neurons in the output layer, W_{ji}, neurons weight of the output layer and ||.|| is the Euclidean distance. After determining the BMU, its weight and the weight of its other neighboring neurons, depending on their distances from the BMU (co-operation phase), are updated according to the equation (3) (adaptation phase).

$$w_{ji}(t+1) = w_{ji}(t) + \theta(t) * \eta(t) * [x_i(t) - w_{ji}(t)] \qquad (3)$$

Where: t, time, θ(t), a function transforming the distance between neighboring neurons of the BMU to a ratio of the neighborhood and η (t) is the learning rate. The process of learning the SOM network is continued by presenting new input data vectors to the SOM network, and during this process the connection weights are adjusted until they remain unchanged. A full description of the self-organizing map process was proposed by Kohonen (1982).

Cluster validity index (Determining the optimal number of clusters)

The indexes evaluating the quality of clustering, regardless of the algorithm used in them, examine the clusters in terms of two parameters: 1- Intra-cluster Similarity (Cluster Compactness) and 2- Inter-cluster Dissimilarity (Cluster Separation). A suitable clustering method (in which number of clusters are optimum) is that in which the value of the two parameters is high (Kaufman *et al.* 2009). Most of indexes evaluating the quality of clustering use the distance criterion to calculate intra-cluster compactness and intra-cluster separation (May *et al.* 2010).

There are various methods to determine the optimal number of clusters (Dunn index, silhouette index, Davies-Bouldin index, validation index, etc.) of which Davies-Bouldin index was used in this study due to its efficiency and easy implementation in

MATLAB software. The index is briefly described below:

Davies-Bouldin index: It calculates mean similarity between two clusters that are most similar (Yar Kiani 2009). The lower calculated value of the index increases the quality of clustering. The index uses the inter-cluster similarity that is defined based on the dispersion of a cluster and inter-cluster dissimilarity. Equation 4 (Yar Kiani 2009):

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (4)$$

Where, R_{ij}: similarity between i and j clusters; S_i and S_j: dispersion of i and j clusters; and d_{ij}: distance between the centers of the two clusters. In Equation 4, dispersion of a cluster and the distance between two clusters are calculated respectively through equations 5 and 6:

$$d_{ij} = d(v_i + v_j) \quad (5)$$

Where, d_{ij}: distance between i and j clusters; and V_i and V_j: centers of i and j clusters.

$$s_i = \frac{1}{|c_i|} \sum_{x \in c_i} d(x, v_i) \quad (6)$$

Where, $|c_i|$ is the number of data in the i^{th} cluster. Finally, Davies-Bouldin index is calculated through Equation 7:

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \quad (7)$$

Where, DB: Davies-Bouldin index; n_c: number of clusters; and R_i: the highest inter-cluster similarity that is calculated using Equation 8:

$$R_i = Max(R_{ij})_{j=1...n_c, i \neq j}, \quad i = 1,...n_c \quad (8)$$

Cluster sampling method

To prepare two sets that were as homogenous and similar as possible (calibration and test data sets), the proportional allocation method was used for sampling the clusters. In this method, the number of samples varies with the size of the cluster, as the size of a cluster increases, the number of samples increases too, and vice versa (May *et al*. 2010). Equation 9:

$$nh = n \frac{Nh}{\sum_{i=1}^{H} Nj} \quad (9)$$

Where, nh: number of samples drawn from h cluster; n: number of required data; Nh: number of data in h cluster; and Nj: number of data in other clusters. In the present study, 80% of the data were used for making the calibration set, and the remaining 20% of the data were used for making the test sets.

Statistical analysis of the data obtained from clustering

Besides, comparison of statistical data (mean, standard deviation, skewness, etc.), the nonparametric two-sample Kolmogorov-Smirnov test (due to the abnormal distribution of data) was used to examine and compare homogeneity and the similarity of the data in calibration and test data sets. The KS test was performed at error level of 1% (α = 1%) using Equation 10 and MATLAB software (Mansourfar 2009):

$$D_C = Max \left| \frac{F(n_{i1})}{n_1} - \frac{F(n_{i2})}{n_2} \right| \quad (10)$$

Where, $F(ni_1)$ and $F(ni_2)$: the cumulative frequency of the variable x in the two sets; and DC: the test statistic, absolute maximum of the difference between relative cumulative frequency of the two data sets.

Preparation of sediment rating curve models (SRC and SRC-FAO models)

The sediment rating curve model (SRC model) was prepared on the basis of Equation 1 and least square method using homogenized data of the calibration data set. Moreover, the FAO correction factor was used to modify the SRC model (SRC-FAO model). The FAO correction factor introduced by Jones *et al.* (1981) for

decreasing bias (underestimation) and increasing values calculated in SRC model using Equation 11:

$$CF = \frac{\overline{Qs}}{\overline{(Qw)^b}} \quad (11)$$

Where, CF: FAO correction factor; $\overline{Qs}$: mean sediment discharge of observational samples (mg l^{-1} or ton day^{-1}); $\overline{Qw}$: mean flow discharge of observational samples (m^3 s^{-1}); and b: the parameter used in the SRC model (Equation 1). After calculating FAO correction factor (CF), the CF substitutes the parameter a in Equation 1.

Using genetic algorithm in the optimization of coefficients of the SRC model (SRC-GA model)

The GA is a nonlinear search and optimization method inspired by biological processes of natural selection and survival of the fittest species. This searching method has relatively few assumptions and do not rely on any mathematic properties of function (continuity and differentiability) (Tayfur 2012). In this method, a population of potential responses is obtained through selecting a random set out of initial solutions, which are actually a set of initial responses of the problem (initial population).

Thereafter, individuals of the population compete with each other to survive and make better responses based on the objective function (Equation 12); consequently, the quality and quantity of the appropriate responses increase in next generations using three genetic operators, including selection, reproduction, and mutation; and this process continues up to the convergence of the algorithm and finding the optimal final response (here a and b coefficients in the SRC regression model).

$$OF(\gamma) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(SSL_o - SSL_e)^2} \quad (12)$$

Where, γ: vector of SRC coefficients (values of a chromosome's genes); SSL_o and SSL_e: values of observational and calculated suspended sediment discharge (ton.day^{-1}); and n: number of calibration data.

When using GA, roulette wheel selection method (weighting method based on the cost of the chromosome) was used to select parents for reproduction; the blending method was used to reproduce; and uniform random number generation method was used for genetic mutations. Noteworthy, GA was used with calibration data, and SRC model coefficients after optimization were used in the SSL estimation of the test data set.

In total, to use a continuous genetic algorithm in this study, we determined an initial population of 50, reproduction of 75%, mutation of 15%, and maximum number of reproductions of 500.

Using a particle swarm optimization algorithm in optimizing coefficients of the SRC model (SRC-PSO model)

PSO consists of a group of particles (individuals) which refine their knowledge of the search space (Kisi *et al.* 2017). In this algorithm, each solution (a and b coefficients in this study) called a particle is assumed as a bird in the migrating swarm pattern and its adequacy is determined by an objective function (like Equation 14).

In PSO algorithm, particles cooperate with one another to reach a common goal, and thus, this method is more effective than that in which particles act separately (Shahriar *et al.* 2011).

In this method, the collective behavior does not only depend on individuals' behavior in the society, but also associates with the manner of interaction among individuals in a group in a way that particles scatter in the searching space and then gradually moves toward successful areas (optimum solutions) to achieve the best solutions under the influence of their own knowledge and their neighbors' knowledge.

In PSO algorithm, firstly, some particles with random location and speed are created; then, these particles modify their movement toward the goal based on the best previous location of themselves and their neighbors in each repetition.

After consequent repetitions, the problem converges to the optimum solution. The speed (V) and location (X) of each particle are modified through equations 13 and 14, respectively (Shahriar Shahhoseini *et al.* 2011; Kisi *et al.* 2017):

$$V_i(t+1) = \omega V_i(t) + C_1 * rand_1(pbest_i(t) - x_i(t)) + C_2 * rand_2(gbest_i(t) - x_i(t)) \quad (13)$$

$$x_i(t+1) = x_i(t) + V_i(t+1) \quad (14)$$

In the above equations, gbest shows the best location obtained from the population of particles; pbest is the best location of the particle itself experienced up to now; t is the number of repetitions; $rand_1$ and $rand_2$ are random numbers in the interval [0 and 1]; and C_1 and C_2 coefficients are respectively cognitive parameter (personal experience) and social parameter (collective experience) that determine the slope of moving when searching for a location.

The value of these two coefficients is determined in the interval [0 and 2], mostly 2 or 1.49 for both coefficients. In the above equations, ω is the inertia coefficient that decreases linearly and is defined in the interval [0 and 1] (Shahriar Shahhoseini *et al.* 2011).

To use the PSO algorithm in this study, the number of initial particles, C_1 and C_2 coefficients, inertia coefficient, and the number of reproductions up to the final convergence were 50, 2, 0.9, and 500 respectively.

Evaluating the efficiency of models

To evaluate the results obtained from different models of SRC (the conventional SRC model, SRC-FAO, SRC-GA and SRC-PSO) and compare their results with those of observational sediment data (data of the test set), graphic drawings and error measurements were performed. Moreover, for each model, the scatter plot of the observational data was drawn using calculated data of the model, and we determined the linear regression equation and correlation coefficient (R^2) of the best fit line (Equation 15). To analyze the measurement error of models, root mean square error (RMSE), mean absolute error (MAE) and Nash-Sutcliffe (NS) were used through equations 16 to 18:

$$R^2 = \left[\frac{\sum_{i=1}^{n}(S_O - \bar{S}_O)(S_M - \bar{S}_M)}{\sqrt{\sum_{i=1}^{n}(S_O - \bar{S}_O)^2 \sum_{i=1}^{n}(S_M - \bar{S}_M)^2}} \right]^2 \quad (15)$$

$$RMSE = \sqrt{\frac{1}{n}} \sum_{i=1}^{n}(S_M - S_O)^2 \quad (16)$$

$$MAE = \frac{\sum_{i=1}^{n}\left|(S_O - S_M)\right|}{n} \quad (17)$$

$$NS = 1 - \frac{\sum_{i=1}^{n}\left(S_M - S_o\right)^2}{\sum_{i=1}^{n}\left(S_O - \bar{S}_O\right)^2} \quad (18)$$

In the above equations, S_o and S_M are observed and estimated suspended sediment discharge, respectively, n is the number of data introduced to the model and $\bar{S}_o$ and $\bar{S}_M$ are the means of observed and estimated suspended sediment discharge.

RESULTS

Results of data clustering

Optimal number of clusters in the studied data were determined as 8 clusters using SOM clustering and Davies-Bouldin index (Fig. 2). Results of statistical parameters and nonparametric two-sample Kolmogorov-Smirnov test in calibration and test data sets (obtained from data clustering through the proportional allocation method) are respectively shown in Tables 2 and 3.

Table 2. Statistical parameters of the variables used in calibration and test data sets.

Model Variables and Data Set	Statistical Parameters					
	$\overline{X}$	S_x	X_{max}	X_{min}	C_{sx}	C_v
Flow Discharge (Q_w) ($m^3\ s^{-1}$)						
Calibration Set	12.90	30.65	349.08	0.01	6.61	2.37
Test Set	11.10	26.39	298.21	0.01	8.21	2.37
Sediment Discharge (SSL) (Q_s) (ton day^{-1})						
Calibration Set	822.50	6610.89	144852.76	0.01	16.90	8.03
Test Set	807.41	6190.53	78616.51	0.00	12.03	7.66

Table 3. Results of two-sample Kolmogorov-Smirnov test of the data.

Model Variables	Data Sets	P-value	D_c	D_t	h
Flow Discharge (Q_w) ($m^3\ s^{-1}$)	Calibration & Test	0.75	0.06	0.14*	0
Sediment Discharge (SSL) (Q_s) (ton day^{-1})	Calibration & Test	0.87	0.05	0.14*	0

*Significant at the error level (α) = 1%

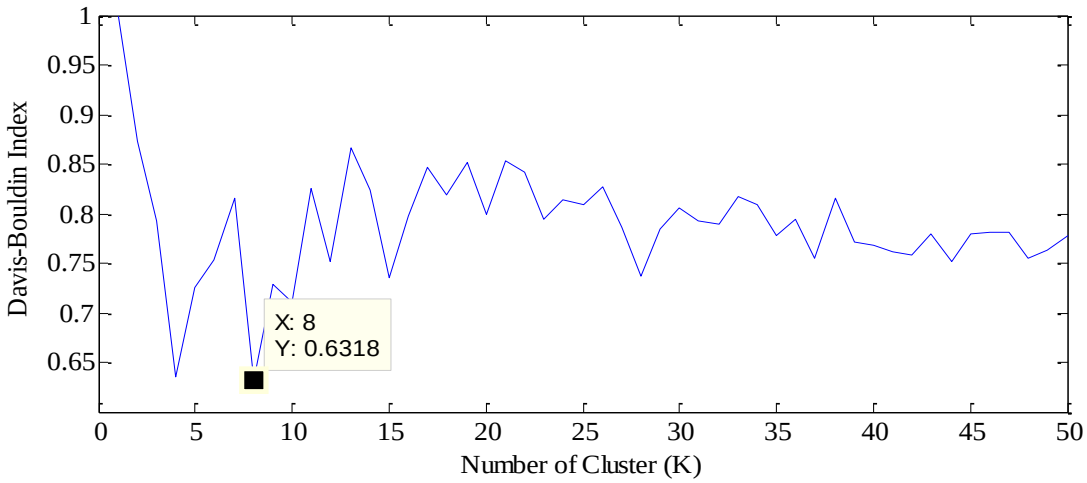

Fig. 2. Determining the optimal number of clusters using SOM clustering and Davies-Bouldin index.

In Table 3, h letter is a statistic for two-sample Kolmogorov-Smirnov test in MATLAB software. When h = 0, it means that it does not reject the null hypothesis (which is that x1 and x2 are from the same continuous distribution) at the significance level of α (α is the desired significance level, e.g. 0.05). The obtained results from the K-S test showed that the distribution of the corresponding data in both data sets (calibration and test data sets) was identical (proof of H_0 hypothesis of the K-S test). These results are also graphically illustrated in Fig. 3. Based on the above results, it could be concluded that the data used in calibration of the models were selected in a way that represented the data of the entire statistical period. It can increase the generalizability of the models.

Results of modeling

Table 4 shows the results of calibration and evaluation of various models of SRC using data of calibration and test data sets. The obtained results show that hybrid models of SRC (SRC-GA and SRC-PSO models) are more favorable than the SRC model and SRC model modified by an FAO factor (SRC-FAO). Also, among the hybrid models, SRC-GA model was selected as the best model because it had slightly more proper performance than SRC-PSO model. In Fig. 4, the fitness of various models of SRC to observational data [flow discharge (QW) and daily sediment discharge (QS) in calibration data set] has been

presented. As well shown in Fig. 4, GA and PSO hybrid models showed better fitness than other models. Furthermore, their difference was very partial, as their curves almost overlay each other.

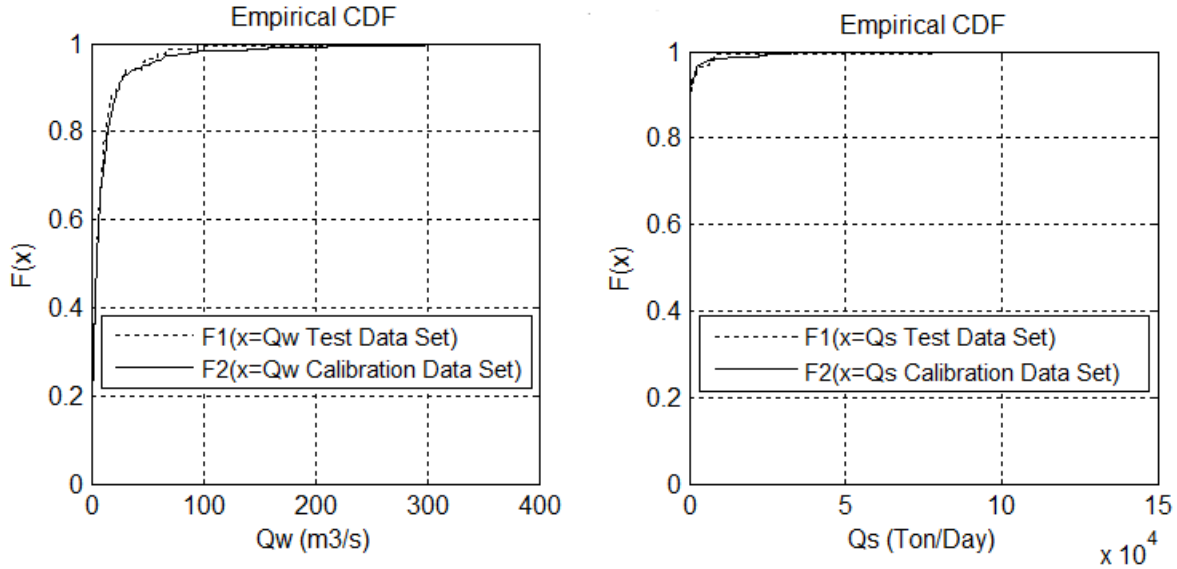

Fig. 3. Comparing the distribution of flow discharge (Q_w) and suspended sediment discharge (Q_s) in the calibration and test data sets using two-sample Kolmogorov-Smirnov test.

Table 4. Results of evaluating various models with calibration and test data sets.

Model Name	Equation	Performance Measures and Data Sets							
		RMSE (ton day^{-1})		MAE (ton day^{-1})		NS		R^2	
		Calibration	Test	Calibration	Test	Calibration	Test	Calibration	Test
SRC	$Q_s = 3.0718Q_w^{1.3248}$	6220.87	5754.02	703.05	710.20	0.11	0.13	0.63	0.94
SRC-FAO	$Q_s = 27.7752Q_w^{1.3248}$	4079.17	2247.07	802.09	638.99	0.62	0.87	0.63	0.94
SRC-GA	$Q_s = 2.7866Q_w^{1.750}$	3964.10	1681.21	471.88	347.31	0.64	0.93	0.64	0.98
SRC-PSO	$Q_s = 4.1783Q_w^{1.6779}$	3960.88	1700.76	502.34	370.02	0.64	0.93	0.64	0.97

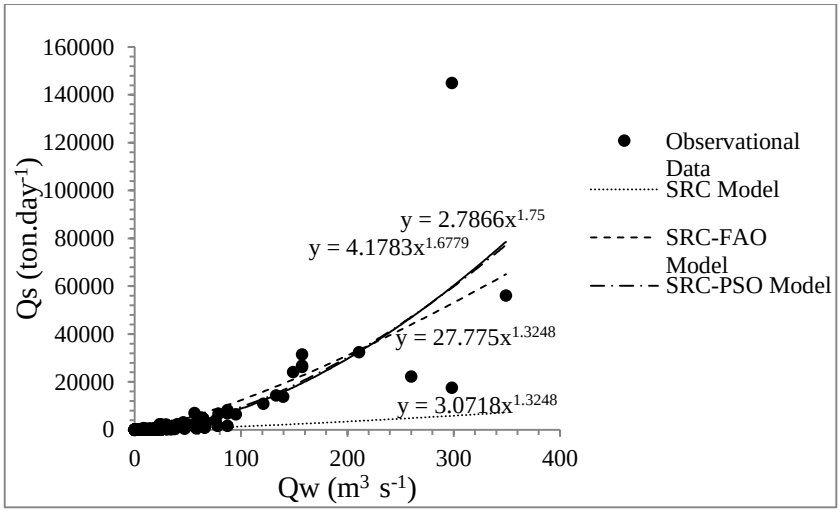

Fig. 4. Fitness of various SRC models to the observational data (calibration data set).

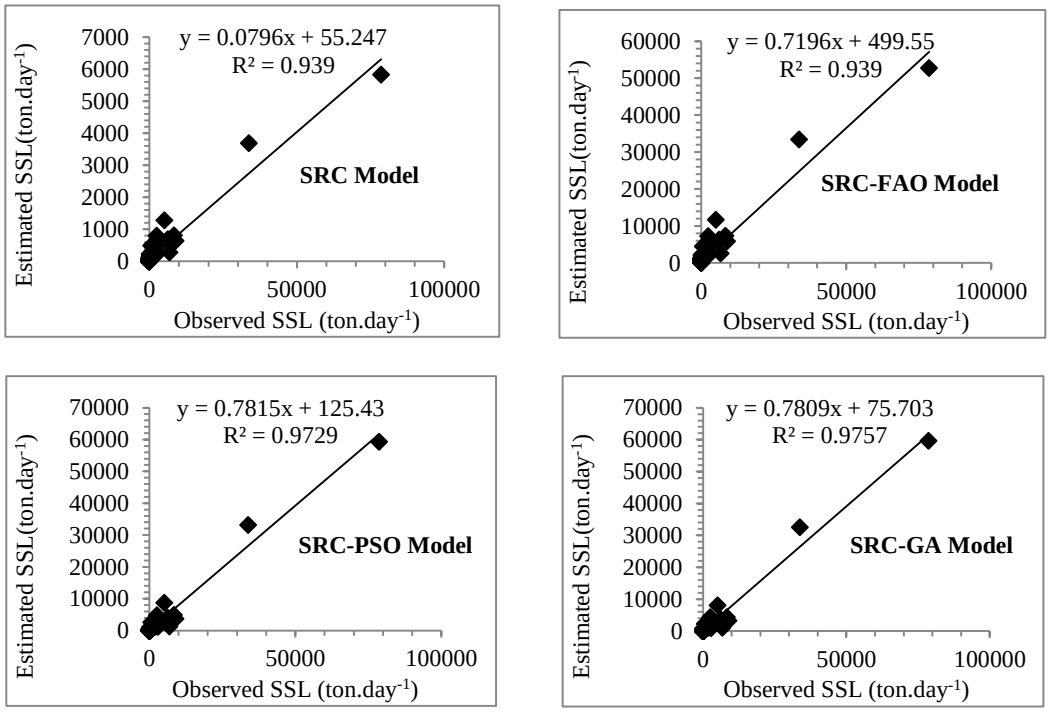

Fig. 5. Scatter plot of observed and estimated suspended sediment load (SSL) (test data set) from different SRC models.

As shown in Fig. 4, some of the data are far from the regression line. This problem can be explained in two parts. First, one of the problems which is associated with sediment data measured at the hydrometric stations is basically the lack of data samples on flood conditions. So, it is common that the quality of these data does not have enough precision to a perfect model calibration. Second, in the sediment rating curve, there is only one predictor variable which is the flow discharge. According to the Rodríguez-Blanco *et al*. (2010), only 19% of the variance in the amount of suspended sediment discharge can be described by flow discharge. So, the poor quality of data on the one hand and use only one predictor variable in the regression model on the other hand cause that the model is not able to simulate all sediment data in low and high flows. Fig. 5, has shown scatter plot and results obtained from simulation of observational suspended sediment discharge of the test data set of different models.

As can be seen in Fig. 5, the slope of fitness line in the evolutionary models (SRC-GA and SRC-PSO) is better than those of FAO-SRC and SRC

models (0.78 against 0.71 in FAO-SRC and 0.07 in SRC models respectively). However, in comparison with PSO-SRC model, the GA-SRC model, by having the less y-intercept and more R^2 was identified as the best model in this study. Fig. 6 shows variations in the value of the cost function (RMSE on calibration data set) in GA and PSO algorithms over different generations (500 generations) up to reaching convergence and determining the optimum value of SRC model coefficients.

DISCUSSION AND CONCLUSION

Accurate suspended sediment load estimation is very essential in planning, designing, operating and favorable performance of water resource structures. The models based on regression methods, such as SRC model, have restricted assumptions such as normality, linearity and constant variance. These models are able to provide only one solution point (a and b coefficients) for estimation of sediment load. On the other hand, the evolutionary algorithms, such as GAs, PSO and etc. can produce more than one solution points that provide the optimal relation between flow discharge and sediment loads. Also, they are not restricted by regression assumptions. Generally, to optimize the coefficients of the SRC model, data log-transformation and least square error method are used in a form of linear regression.

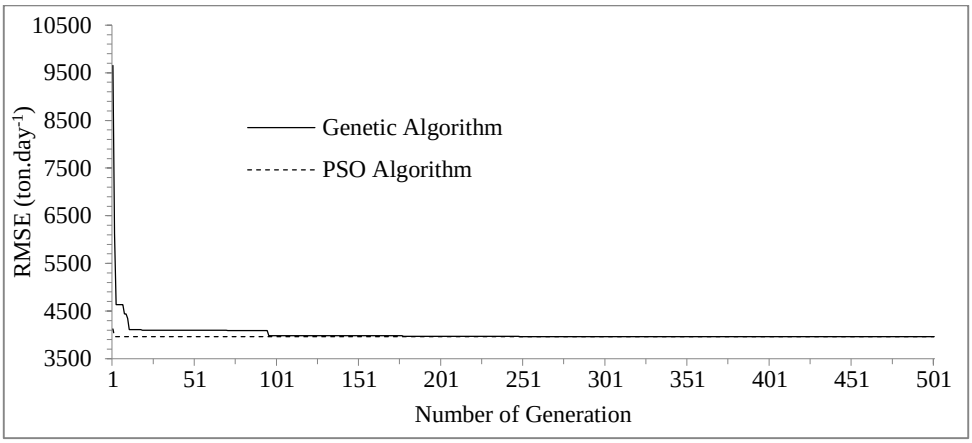

Fig. 6. Diagram of the minimum cost (RMSE) as a function of generations up to reaching convergence in GA and PSO algorithms.

The data log-transformation results in a bias in the calculation of model coefficients and underestimation of SSL (sediment discharge or sediment concentration). This problem is most obvious in high flood discharges, and the model error increases with an increase in the flow discharge. So far, different correction factors have been introduced to correct the bias. However, these factors sometimes cause another error in the form of an overestimation along with different results. In this study, besides the conventional methods (least square error method and the model modified with FAO factor), the SRC model coefficients were optimized through evolutionary algorithms (GA and PSO) and results were much more favorable than those of the conventional methods. The results of this study conformed to those of the studies conducted by Altunkaynak (2009), Mohammad Rezapour *et al.* (2016) and Swain & Sahooh (2017). Using evolutionary algorithms also prevents the data logarithm transformation and use of correction factors and increases the accuracy of results. Moreover, to increase generalizability of data-driven models, the samples used in calibration of models should represent the data of the entire statistical period. To properly evaluate the model and its results, the test data set should be similar to those of calibration data set. This is an important problem and of the fundamental challenges in modeling, as the failure to use similar homogenous data in calibration and test sets may largely affect the results of modeling.

So that, the SOM clustering method can be used to provide similar homogeneous data sets for calibration and evaluation of data-driven models (Li *et al*. 2010).

ACKNOWLEDGEMENT

We are grateful for financial support of Soil Conservation and Watershed Management Research Institute (SCWMRI).

REFERENCES

Altunkaynak, A 2009, Sediment load prediction by genetic algorithms. *Advances in Engineering Software*, 40: 928-934.

Bowden, GJ, Maier, HR & Dandy, GC 2002, Optimal division of data for neural network models in water resources applications. *Water Resources Research*, 38: 2-11.

Buyukyildiz, M & Kumcu, SY 2017, An Estimation of the Suspended Sediment Load Using Adaptive Network Based Fuzzy Inference System, Support Vector Machine and Artificial Neural Network Models. *Water Resources Management*, 31: 1343-1359.

Chen, XY & Chau, KW 2016, A hybrid double feed-forward neural network for suspended sediment load estimation. *Water Resources Management*, 30: 2179-2194.

Cheng, CT, Ou, CP, & Chau, KW 2002, Combining a fuzzy optimal model with a genetic algorithm to solve multi-objective rainfall–runoff model calibration. *Journal of Hydrology*, 268: 72-86.

Ebtehaj, I & Bonakdari, H 2016, Assessment of evolutionary algorithms in predicting non-deposition sediment transport. *Urban Water Journal,* 13: 499-510

Ferguson, RI 1986, River loads underestimated by rating curves. *Water Resources Research,* 22: 74-76.

Gholami, V, Darvari, Z & Saravi, MM 2015, Artificial neural network technique for rainfall temporal distribution simulation (Case study: Kechik region), *Caspian Journal of Environmental Sciences,* 13: 53-60.

Guo, W & Wang, H 2010, PSO optimizing neural network for the Yangtze River sediment entering estuary prediction. In *2010 6th International Conference on Natural Computation,* 4: 1769-1772, IEEE.

Hejazi, MI, Cai, X & Borah, DK 2008, Calibrating a watershed simulation model involving human interference: an application of multi-objective genetic algorithms. *Journal of Hydroinformatics*, 10: 97-111.

Jones, KR, Berney, O, Carr, DP & Barrett, EC 1981. Arid zone hydrology for agricultural development. pp. 190–206.

Kalteh AM 2008, Rainfall-runoff modelling using artificial neural networks (ANNs): modelling and understanding, *Caspian Journal of Environmental Sciences,* 6: 53-58.

Kalteh, AM, Hjorth, P & Berndtsson, R 2008, Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. *Environmental Modelling & Software,* 23: 835-845.

Kao, SJ, Lee, TY & Milliman, JD 2005, Calculating highly fluctuated suspended sediment fluxes from mountainous rivers in Taiwan. *Terrestrial Atmospheric and Oceanic Sciences,* 16: 653.

Kaufman, L & Rousseeuw, PJ 2009, Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons. pp. 227–252.

Kisi, O, Dailr, AH, Cimen, M & Shiri, J 2012, Suspended sediment modeling using genetic programming and soft computing techniques. *Journal of Hydrology*, 450-451: 48-58.

Kisi, O & Zounemat-Kermani, M 2016, Suspended Sediment Modeling Using Neuro-Fuzzy Embedded Fuzzy c-Means Clustering Technique. *Water Resources Management,* 30: 3979-3994.

Kisi, O, Keshavarzi, A, Shiri, J, Zounemat-Kermani, M & Omran, ESE 2017, Groundwater quality modeling using neuro-particle swarm optimization and neuro-differential evolution techniques. *Hydrology Research*, 48: 1508-1519

Kohonen, T 1982, Analysis of a simple self-organizing process. *Biological Cybernetics,* 44: 135-140.

Kohonen, T 2001, Self-organizing maps, Vol. 30 of Springer Series in Information Sciences. Ed.: Springer Berlin, pp. 98–114.

Kuok, KK, Harun, S & Shamsuddin, SM 2010, Particle swarm optimization feed-forward neural network for modeling runoff. *International Journal of Environmental Science & Technology,* 7: 67-78.

LiX, Nour, MH, Smith, DW & Prepas, EE 2010, Neural networks modelling of nitrogen export: model development and application to unmonitored boreal forest watersheds. *Environmental Technology*, 31: 495-510.

Mansourfar, K 2009, Advanced statistical methods using applied software. University of Tehran Press. p. 459 (In Persian).

May, RJ, Maier, HR & Dandy, GC 2010, Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Networks*, 23: 283-294.

Mohammad Rezapour, O, Nour jou, P, Zeynali, MJ 2016, Compression of Genetic Algorithm and Particle Swarm Algorithm models for Optimizing Coefficients of Sediment Rating Curve in the estimation of Suspended Sediment in Sistan River (Case Study: Kohak station). *The Iranian Society of Irrigation & Water Engineering*, 6: 76-89 (In Persian).

Muhammadi, A, Akbari, G & Azizzian, G 2012, Suspended sediment concentration estimation using artificial neural networks and neural-fuzzy inference system case study: Karaj Dam. *Indian Journal of Science and Technology,* 5: 3188-3193.

Nourani, V, Alizadeh, F & Roushangar, K 2016, Evaluation of a two-stage SVM and spatial statistics methods for modeling monthly river suspended sediment load. *Water Resources Management,* 30: 393-407.

Rodríguez-Blanco, ML, Taboada-Castro, MM, Palleiro-Suárez, L, Taboada-Castro, MT 2010, Temporal changes in suspended sediment transport in an Atlantic catchment, NW Spain. *Geomorphology*, 123: 181-188.

Salehpour Jam, A, Tabatabaei, M, Sarreshtehdari, A 2017, Pedological criterion affecting desertification in alluvial fans using AHP-ELECTRE I technique, case study: southeast of Rudeshoor watershed area. *ECOPERSIA*, 5: 1711-1730.

Sarkar, A, Sharma, N & Singh, RD 2017, Sediment Runoff Modelling Using ANNs in an Eastern Himalayan Basin, India. In *River System Analysis and Management*. Springer Singapore. pp. 73-82.

Shahriar Shahhoseini, H, Moosavi, M, Mollajafari, M 2011, Evolutionary algorithms - Fundamentals, Applications, Implementation. Tehran: *Press Center of Iran University of Science and Technology*, pp. 413–439 (In Persian).

Swain, R & Sahoo, B 2017, Mapping of heavy metal pollution in river water at daily time-scale using spatio-temporal fusion of MODIS-aqua and Landsat satellite imageries. *Journal of Environmental Management,* 192: 1-14.

Tayfur, G 2009, GA-optimized model predicts dispersion coefficient in natural channels. *Hydrology Research*, 40: 65-78.

Tayfur, G 2012, Soft computing in water resources engineering: Artificial neural networks, fuzzy logic and genetic algorithms. WIT Press. pp. 56–89.

Ulke, A, Tayfur, G & Ozkul, S 2009, Predicting suspended sediment loads and missing data for Gediz River, Turkey. *Journal of Hydrologic Engineering,* 14: 954-965.

Vercruysse, K, Grabowski, RC & Rickson, RJ 2017, Suspended sediment transport dynamics in rivers: Multi-scale drivers of temporal variation. *Earth-Science Reviews,* 166: 38-52.

Yar Kiani, A 2009 Intelligent Systems. Tehran: *Press Center of Poyesh Andisheh*, pp. 14–43. (In Persian).

بهینه‌سازی ضرایب منحنی سنجه رسوب با استفاده از الگوریتم‌های تکاملی و شبکه عصبی مصنوعی بدون ناظر

طباطبائی م.*، صالح پورجم الف.

پژوهشکده حفاظت خاک و آبخیزداری، سازمان تحقیقات، آموزش و ترویج کشاورزی، وزارت جهاد کشاورزی، تهران، ایران

(تاریخ دریافت: ۹۶/۰۴/۱۴ تاریخ پذیرش: ۹۶/۰۹/۰۷)

چکیده

منحنی سنجه رسوب، یک مدل رگرسیونی مرسوم در برآورد بار رسوب معلق از دبی جریان است. با وجود این، در اغلب موارد تبدیل لگاریتمی داده‌ها در مدل‌های منحنی سنجه رسوب سبب بروز خطا شده که منجر به کم برآوردی بار رسوب معلق میشود. در این مطالعه، با استفاده از داده‌های دبی جریان روزانه و بار رسوب معلق ایستگاه هیدرومتری شلمان واقع در رودخانه شلمان‌رود در استان گیلان، ایران، مدل منحنی سنجه رسوب اقتباس و پس از آن این مدل با استفاده از الگوریتم‌های تکاملی (الگوریتم ژنتیک و الگوریتم بهینه‌سازی ازدحام ذرات) مجدداً واسنجی شد. لازم به ذکراست که به منظور افزایش قدرت تعمیم‌دهی مدل‌ها و قبل از واسنجی آن‌ها، با استفاده از روش نگاشت خودسازمان‌ده (یک شبکه عصبی بدون ناظر برای خوشه‌بندی داده‌ها)، داده‌ها خوشه‌بندی شده و سپس با استفاده از نمونه‌گیری از آن‌ها، داده‌ها به دو دسته همگن و مشابه (مجموعه داده‌های واسنجی و آزمون) طبقه‌بندی شدند. نتایج نشان داد که الگوریتم‌های تکاملی، روش‌های مناسبی برای بهینه‌سازی ضرایب مدل منحنی سنجه رسوب هستند و نتایج آن‌ها به مراتب بهتر از مدل‌های سنتی منحنی سنجه رسوب یا منحنی سنجه تصحیح شده با ضرایب تصحیحی است، به نحوی که مدل منحنی سنجه رسوب واسنجی شده با الگوریتم‌های تکاملی با کاهش مقدار RMSE داده‌های آزمون از ۵۷۵۴/۰۲ تن در روز (در مدل اصلی منحنی سنجه) به ۱۶۸۱/۲۱ تن در روز (در مدل‌های واسنجی شده با الگوریتم‌های تکاملی) صحت برآورد رسوب معلق را به میزان ۴۰۷۲/۸۱ تن در روز افزایش داده است. در مجموع، استفاده از الگوریتم‌های تکاملی در واسنجی مدل‌های منحنی سنجه رسوب، مانع از تبدیل لگاریتمی داده‌ها و استفاده از ضرایب تصحیح شده و همچنین سبب افزایش صحت نتایج شبیه‌سازی می‌شود.

* مولف مسئول