Indian Journal of Science and Technology, *Vol 7(10), 1488–1497, October 2014*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Daily Suspended Sediment Load Estimation by Meta-heuristic Optimization Approaches and Fuzzy-C-Means Clustering Method

(Case Study: Siera Hydrometry Station - Karaj River)

Mahmoodreza Tabatabaei[1*], Karim Solaimani[2], Mahmoud Habibnejad Roshan[2] and Ataollah Kavian[2]

[1*]Department of Watershed Management and Research Instructor of Soil Conservation and Watershed Management Research Institute (SCWMRI), Tehran, Iran; Taba1345@hotmail.com
[2]Department of Watershed Management Engineering, Faculty of Natural Resources, Sari Agricultural Sciences and Natural Resources University, Sari, Iran

Abstract

Accurate Suspended Sediment Load (SSL) estimation is very important for water resources quantity and quality studies. In this regard, Sediment Rating Curve (SRC) is a common regression model in predicting SSL of discharge. Studies in this field have shown that the data log-transformation in SRC model causing a bias which underestimates SSL prediction. In this study, using data from the daily flow discharge and suspended sediment discharge of Karaj Dam watershed at Siera station, a 30–year period (1981 to 2011), SRC equation derived, and then, using meta-heuristic algorithms (genetic algorithm and particle swarm optimization algorithm) it was calibrated again. Before modeling to increase the generalization power of the models, using fuzzy clustering method, the data were clustered and then by doing data sampling, they were classified into two homogeneous groups (calibration and test data set). The results show that meta-heuristic 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. In this relation, the sediment rating curve models calibrated with meta-heuristic algorithms, by reducing the RMSE of the test data set of 3718.87 ton/day (in the initial SRC model) to 2615/119 ton/day (in the calibrated models by meta-heuristic algorithms) increased the accuracy of suspended sediment load estimation at a rate of 1103.68 ton/day. However, the SRC model corrected by FAO factor decreased the efficiency of initial SRC model by increasing the RMSE of the test data set to 4128/73 ton/day. Using meta-heuristic algorithms in calibrating SRC models also prevents data log-transformation and use of correction factors and increases the accuracy of results.

Keywords: Fuzzy-C-Means Clustering, Genetic and PSO Algorithms, Karaj Dam, Sediment Rating Curve, Suspended Sediment Load

1. 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 and changes in the river bed and river bank, the quality of water, and optimum design and favorable performance of water resource structures[1–5]. 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[6]:

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

where, $Q_{(t)}$ is the mean flow discharge (m^3/s), SSL(t) is the suspended sediment discharge (ton/day), and a and b are the constant coefficients of the regression equation.

**Author for correspondence*

In Equation 1, SSC (mg/l) can be used instead of SSL (ton/day).

To use the SRC regression model, the coefficients (a and b) should be calculated optimally. This is firstly done through taking logarithm of variables of flow discharge and sediment discharge and formulating a linear regression equation between them, and 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[7]. In other words, when calculating a and b coefficients, a kind of bias appears in SRC regression model and makes the estimated values of SSL lower than its corresponding observed values[8]. This problem is more 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, QMLE (Quasi-Maximum Likelihood Estimator), MVUE (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 different[7].

In recent years, 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. Cheng et al.[9] could calibrate parameters of 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.[10] calibrated parameters of a distributed rainfall-runoff model using multi-objective GA. Tayfur[11] optimized parameters of some empirical equations and could estimate longitudinal dispersion coefficient of a river. Kisi et al.[12] used the Genetic Programming (GP) model in order to estimate the amount of daily suspended sediment in two stations in Cumberland River in America. Kuok et al.[13] applied PSO algorithm to optimize parameters of neural network model of daily rainfall-runoff in Sungai Bedup watershed in Malaysia. They showed that the neural network training through the above method was successful. Guo and Wang[14] used Radial Basis Function (RBF) neural network whose parameters were optimized based on PSO algorithm to estimate SSL of Yangtze river. Muhammadi et al.[15] used neural network, Adaptive Neuro-Fuzzy Inference System (ANFIS) and sediment rating curve models to estimate suspended sediment concentration of Karaj river at Siera hydrometric station of Karaj dam watershed in Iran. Water temperature, flow discharge, suspended sediment concentration and water depth were the model inputs. The comparative analysis of the results showed that the ANFIS model has superiority over the other models for estimating daily suspended sediment concentration. Few studies that are mentioned below have been performed on the use of evolutionary algorithms in optimization of SRC coefficients so far. Altunkaynak[16] could optimize SRC coefficients of Mississippi river located in St. Louis, MO using GA. Results of the study showed the priority of SRC model optimized by GA over its conventional model. Other similar studies conducted by Rezapour et al.[17] and Ebrahimi et al.[18] indicated the priority of meta-heuristic methods over SRC regression model.

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 designed models and reduces its generalization[19]. Fuzzy C- Means Clustering (FCM) was used in the present study to build two similar homogenous data sets for calibration and test of 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 Karaj river using the traditional SRC model and the SRC model modified by FAO correction factor.
B. Optimization of SRC model's coefficients using meta-heuristic 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.

It should be mentioned that PSO meta-heuristic algorithm was first used in this study for optimization of SRC coefficients.

2. Materials and Methods

In this study, MATLAB 7.11 software was used to implement GA and PSO algorithms, cluster the data, and calculate cluster validity index. The data were statistically analyzed using SPSS19 and MATLAB software programs.

2.1 The Study Area and Used Data

The present study was performed in Karaj river watershed, Siera Hydrometric Station. The watershed located at east longitude of 51°–51°35' and north latitude of 35°53'–36°11' in 30–60 km away from north and northwest of Tehran, Iran (Figure 1).

The watershed has the area of 84213 hectares and mean elevation of 2827 m above sea level. The region involved soils with various thicknesses ranging from young undeveloped entisols to medium-developed inceptisols with different parent material[20]. The statistics used in this study included 611 information records of hydrometric data of instantaneous flow discharge and sediment discharge in Siera Hydrometric Station during 30 years (1981–2011). Table 1 shows statistical characteristics of the data used in this period. 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.

2.2 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, FCM 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:

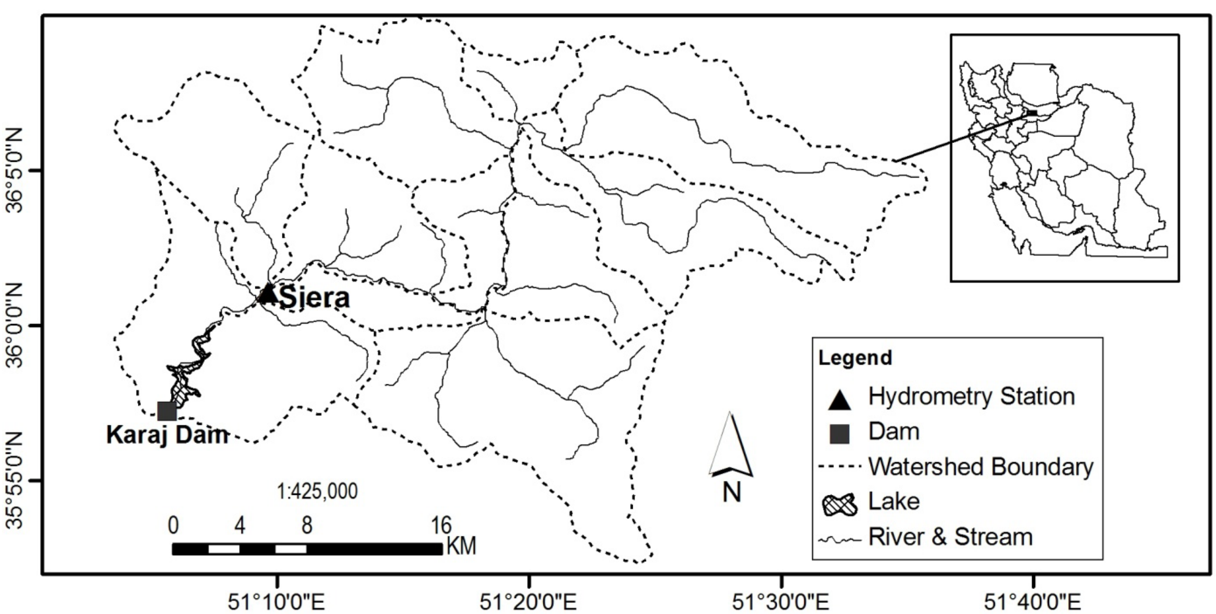


Figure 1. The location of Karaj watershed and Siera Hydrometric Station.

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

Data Set	Data Type	$\overline{X}$	S_x	C_v	C_{sx}	X_{max}	X_{min}
Whole data	Flow, Q_w (m^3/s)	17.14	16.91	0.99	2.27	136.17	2.63
	SSL, Q_s (ton/day)	1517.54	5433.4	3.58	7.08	62958.91	0.74

2.2.1 Data clustering using FCM

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. Clustering algorithms can be divided into two groups: hard clustering algorithms and soft clustering algorithms[21]. In hard clustering algorithms, each sample belongs only to one cluster, while, in soft clustering algorithms (like fuzzy clustering), each sample with a specific degree of membership may belong to different clusters. In fuzzy clustering algorithms, each sample's degree of membership is obtained based on the distance between the sample and center of the cluster in which the sample is placed. The nearer the sample to the center of the cluster, the higher the degree of sample membership. Therefore, in this clustering method, the objective is to minimize the objective function (Equation 2), in other words, maximize samples' degree of membership[21].

$$J\,m = \sum_{i=1}^{N}\sum_{j=1}^{C} U_{ij}^{m}\left\|x_i - c_j\right\|^2 \quad 1 \le m \le \infty \tag{2}$$

where, m: any real number greater than 1; U_{ij}: membership degree of X_i in j cluster; X_i: i^{th} data of the dimension d; C_j: center of the cluster's dimension d; and $\|.\|$ is any criterion (such as Euclidean distance) indicating a similarity between a measured data and the cluster's center. The fuzzy clustering is performed through frequent optimization of the above function. This process is done through updating U_{ij} membership degree and C_j cluster's center using the following equations[21]:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left[\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_j\right\|}\right]^{\frac{2}{m-1}}} \tag{3}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{4}$$

The process is repeated when:

$$\max_{ij}\left\{\left|u_{ij}^{k+1} - u_{ij}^{k}\right|\right\} < \varepsilon \tag{5}$$

where, ε: the termination criterion; and k: the kth step of the repetition process.

2.2.2 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 is optimum) is that in which the value of the two parameters is high[22]. Most of indexes evaluating the quality of clustering use the distance criterion to calculate intra-cluster compactness and intra-cluster separation[19]. 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 mostly similar[21]. 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 6:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{6}$$

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 6, dispersion of a cluster and the distance between two clusters are calculated respectively through equations 7 and 8:

$$d_{ij} = d\left(v_i + v_j\right) \tag{7}$$

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

$$s_i = \frac{1}{|c_i|} \sum x \in c_i \; d\left(x, v_i\right) \tag{8}$$

where, $|c_i|$ is number of data in the ith cluster. Finally, Davies-Bouldin index is calculated through Equation 9:

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \tag{9}$$

where, DB: Davies-Bouldin index; n_c: number of clusters; and R_i: the highest inter-cluster similarity that is calculated using Equation 10:

$$R_i = Max(R_{ij})_{j=1 \dots n_c, i \neq j}, i = 1, \dots n_c \tag{10}$$

2.2.3 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 cluster, as the size of a cluster increases, the number of samples increases too, and vice versa[19]. Equation 11:

$$nh = n \frac{Nh}{\sum_{i=1}^{H} Nj} \tag{11}$$

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.

2.2.4 Statistical Analysis of the Data Obtained from Clustering

Besides comparison of statistical data (mean, standard deviation, skewness, etc.), the nonparametric two-sample KS test (due to the abnormal distribution of data) was used to examine and compare homogeneity and similarity of the data in calibration and test data sets. The KS test was performed at error level of 1% (α = 1%) using Equation 12 and MATLAB software[23]:

$$D_C = Max\left|\frac{F(n_{i1})}{n_1} - \frac{F(n_{i2})}{n_2}\right| \tag{12}$$

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.

2.3 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, FAO correction factor was used to modify the SRC model (SRC-FAO model). The FAO correction factor introduced by Jones et al.[24] for obviating bias (underestimation) and increasing values calculated in SRC model using Equation 13:

$$CF = \frac{\overline{Qs}}{\overline{(Qw)}^b} \tag{13}$$

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

2.4 Using Genetic Algorithm in Optimization of Coefficients of 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 relative few assumptions and do not rely on any mathematic properties of function (continuity and differentiability)[5]. 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). After that, individuals of the population compete with each other to survive and make better responses based on the objective function (Equation 14); 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 SRC regression model).

$$OF(\gamma)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(SSL_O-SSL_e)^2} \qquad (14)$$

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); 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. It should be noted that GA was used with calibration data, and SRC model coefficients after optimization were used in SSL estimation of the test data set. In total, to use continuous genetic algorithm in this study, an initial population of 50, reproduction of 75%, mutation of 15%, and maximum number of reproduction of 500 were determined.

2.5 Using Particle Swarm Optimization Algorithm in Optimizing Coefficients of SRC Model (SRC-PSO Model)

PSO is a social searching algorithm inspired by the social behavior of swarms of birds and fish when looking for food[25]. In this algorithm, each solution (a and b coefficients in this study) called a particle is assumed as a bird in 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[26]. 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 move 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 15 and 16, respectively[26]:

$$V_i(t+1)=\omega V_i(t)+C_1^{*}rand_1(pbest_i(t)-x_i(t)) + C_2^{*}rand_2(gbest_i(t)-x_i(t)) \qquad (15)$$

$$x_i(t+1)=x_i(t)+V_i(t+1) \qquad (16)$$

In the above equations, gbest shows the best location obtained by the population of particle; 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][26]. To use 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 respectively 50, 2, 0.9, and 500.

2.6 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 the regression linear equation and correlation coefficient (R^2) of the best fitness line (Equation 17) were determined. To analyze the measurement error of models, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were used through equations 18 and 19:

$$R^2=\left[\frac{\sum_{i=1}^{n}\left(S_O-\overline{S_O}\right)\left(S_M-\overline{S_M}\right)}{\sum_{i=1}^{n}\left(S_O-\overline{S_O}\right)^2\sum_{i=1}^{n}\left(S_M-\overline{S_M}\right)^2}\right]^2 \qquad (17)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_M - S_O)^2} \tag{18}$$

$$MAE = \frac{\sum_{i=1}^{n}|S_O - S_M|}{n} \tag{19}$$

In the above equations, S_o and S_M are respectively the observational suspended sediment discharge and estimated SSL discharge, and n is the number of data introduced to the model.

3. Results

3.1 Results of Data Clustering

Optimal number of clusters for the studied data was determined as 32 clusters using fuzzy clustering and Davies-Bouldin index (Figure 2).

Results of statistical parameters and nonparametric two-sample KS tests in calibration and test data sets obtained from data clustering through the proportional allocation method are respectively shown in Table 2 and 3.

The results obtained from KS test (Table 3) at error level of 1% (confidence interval of 99%) showed that the distribution of corresponding data in both data sets (calibration and test data sets) was identical (proof of H_0 hypothesis of the KS test). These results are provided in Figure 3 graphically.

Based on the above results, it could be concluded that the data used in calibration of models were selected in a way that represented the data of the entire statistical period, and this increased the generalizability of the models.

3.2 Results of Modeling

Table 4 shows results of calibration and evaluation of SRC's various models using data of calibration and test

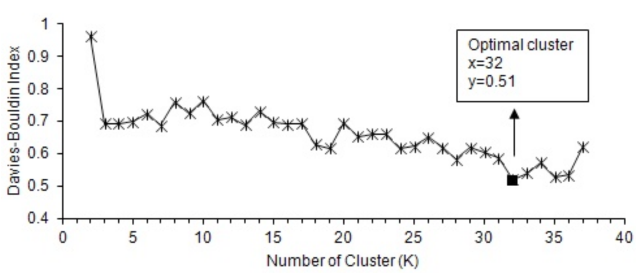


Figure 2. Determining optimal number of clusters using FCM clustering and Davies-Bouldin index.

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

Model Variables & Data Set	Statistical Parameters					
	$\overline{X}$	S_x	C_{sx}	X_{max}	X_{min}	C_v
Flow Discharge (Q_w) (m^3/s)						
Calibration Set	17.12	16.91	2.33	136.17	2.63	98.8
Test Set	17.24	17	2	94.42	2.84	98.6
Sediment Discharge (SSL) (Q_s) (ton/day)						
Calibration Set	1544.59	5614.79	7.11	62958.91*	1	363.51
Test Set	1410.43	4665.42	6.27	40832.95	0.74	330.78

*When sampling the clusters, the data limit values were put in the calibration set

sets. In this table, hybrid models of SRC (SRC-GA and SRC-PSO models) present results more favorable than those of the SRC model and modified with FAO factor (SRC-FAO). Among hybrid models, SRC-PSO model was the best model and showed more favorable results than the SRC and SRC-FAO models. As mentioned earlier, the use of FAO factor in SRC model did not only correct the bias (underestimation) of the model but also increased the error of the model through overestimation.

Figure 4 shows fitness of various models of SRC to observational data (flow discharge (Q_W) and daily sediment discharge (Q_S) in calibration data set). As well shown in the figure, GA and PSO hybrid models showed better fitness than other models. Furthermore, their difference was very partial, as their curves almost overlay each other.

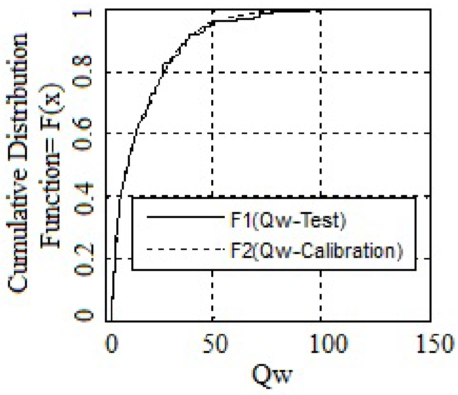
In Figure 5 and 6, scatter plot and results obtained from simulation of observational suspended sediment discharge of the test data set by different models have been shown.

In Figure 6, the data estimated by the SRC-FAO model were highly more than the real values. Moreover, the

Table 3. Results of KS test of the data used in calibration and test sets

Model Variables	Data Sets	P-value	D_c	D_t
Flow Discharge (Q_w) (m^3/s)	Calibration & Test	0.99	0.041	0.164**
Sediment Discharge (SSL) (Q_s) (ton/day)	Calibration & Test	0.7	0.069	0.164**

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



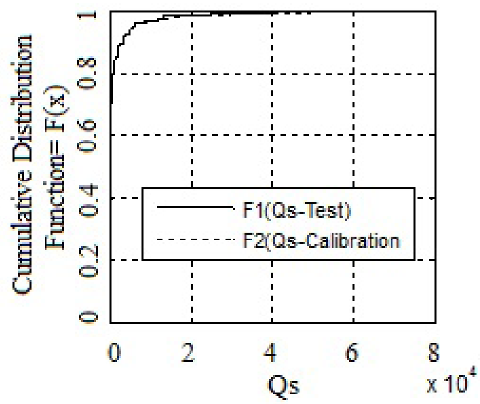


Figure 3. Comparing the distribution of flow discharge (Q_w) and suspended sediment discharge (Q_s) (respectively a and b) in the calibration and test data sets using two-sample KS test.

results of SRC model were much lower than the observational values.

Figure 7 shows variations in the value of cost function (RMSE) in GA and PSO over different generations (500 generations) up to reaching convergence and determining optimum value of SRC model coefficients.

4. Discussion and Conclusion

The SRC regression model is a common method for estimation of rivers' SSL. To optimize coefficients of the model, data log-transformation and least square error method are used in a linear regression model. The data transformation results in a bias in calculation of coefficients and underestimation of SSL (sediment discharge or sediment concentration). The problem is more 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 meta-heuristic methods (GA and PSO) and showed results much more favorable than those of the conventional

Table 4. Results of evaluating various models of SRC using the data of calibration and test data set

Model Name	Equation	Performance Measures and Data Sets					
		RMSE (ton/day)		MAE (ton/day)		R^2	
		Calibration	Test	Calibration	Test	Calibration	Test
SRC	$Q_s=1.1352Q_w^{1.9883}$	4555.03	3718.87	1176.51	1077.52	0.61	0.66
SRC-FAO	$Q_s=5.4511Q_w^{1.9883}$	7811.98	4128.73	2998.94	1900.94	0.61	0.66
SRC-GA	$Q_s=1.3216Q_w^{2.1898}$	3460.5	2619.53	1115.67	1065.95	0.62	0.68
SRC-PSO	$Q_s=1.2601Q_w^{2.2004}$	3419.42	2615.19	1109.41	1060.15	0.62	0.68

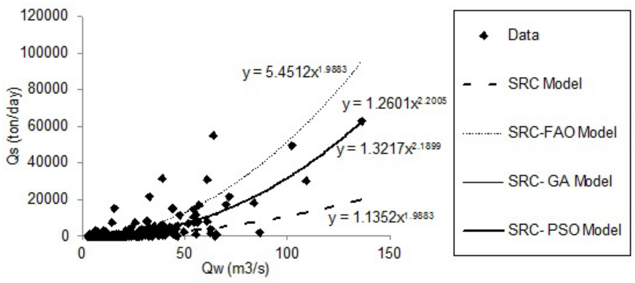


Figure 4. Fitness of various models of SRC to observational data (calibration data).

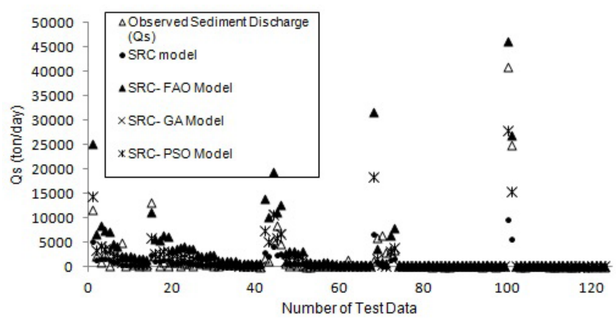


Figure 5. Results of simulation of suspended sediment discharge by different SRC models.

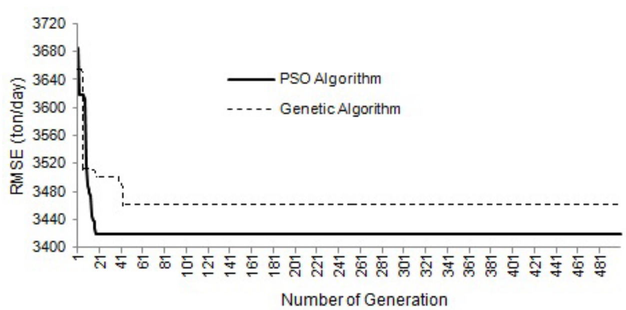


Figure 6. Diagram of the minimum cost as a function of generations up to reaching convergence in GA and PSO algorithm.

methods. Results of this study (on optimization of the SRC model coefficients using GA) conformed to those of the studies conducted by Altunkaynak[16], Rezapour et al.[17], and Ebrahimi et al.[18]. Furthermore, optimization of the SRC model coefficients using PSO algorithm was first introduced by the present study and can be used as an appropriate method for optimizing coefficients of SRC model. PSO method also prevents 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 should be similar to those of calibration. This is an important problem and of fundamental challenges in modeling, as the failure to use similar homogenous data in calibration and test sets may largely affect results of modeling. In this regard, FCM clustering can be used to provide similar homogenous data for calibration and evaluation of data-driven models.

5. References

1. Azamathulla HM, Cuan YC, Ghani AA, Chang CK. Suspended sediment load prediction of river systems: GEP approach. Arab J Geosci. 2013 Sep; 6(9):3469–80.
2. Cobaner M, Unal B, Kisi O. Suspended sediment concentration estimation by an adaptive neuro-fuzzy and neural network approaches using hydro-meteorological data. J Hydrology. 2009; 367(1):52–61.
3. Collins AL, Walling DE. Documenting catchment suspended sediment sources: problems, approaches and prospects. Progr Phys Geogr. 2004 June; 28(2):159–96.
4. Kisi O, Shiri J. River suspended sediment estimation by climatic variables implication: Comparative study among soft computing techniques. Comput Geosci. 2012 June; 43:73–82.
5. Tayfur G. Soft Computing in Water Resources Engineering-Artificial Neural Networks, Fuzzy Logic and Genetic Algorithms. Southampton: WIT Press; 2012.
6. Ulke A, Tayfur G, Ozkul S. Predicting suspended sediment loads and missing data for Gediz River, Turkey. J Hydrol Eng. 2009 Sep; 14(9):954–65.
7. Kao S, Lee T, Milliman JD. Calculating highly fluctuated suspended sediment fluxes from mountainous rivers in Taiwan. TAO. 2005 August; 16(3):653–75.
8. Ferguson RI. River loads underestimated by rating curves. Water Res. 1986; 22(1):74–6.
9. Cheng CT, Ou CP, Chau KW. Combining a fuzzy optimal model with a genetic algorithm to solve multi-objective rainfall–runoff model calibration. J Hydrology. 2002 Nov; 268(1-4):72–86.
10. Hejazi M, Cai X, Borah D. Calibrating a Watershed Simulation Model Involving Human Interference – An Application of Multi-objective Genetic Algorithms. J Hydroinformatics. 1998; 10(1):97–111.
11. Tayfur G. GA-optimized model predicts dispersion coefficient in natural channels. Nord Hydrol. 2009; 40(1):65–78.
12. Kisi O, Dailr AH, Cimen M, Shiri J. Suspended sediment modeling using genetic programming and soft computing techniques. J Hydrology. 2012 July 11; 450–451:48–58.

13. Kuok KK, Harun S, Shamsuddin SM. Particle Swarm Optimization Feedforward Neural Network for Hourly Rainfall-runoff Modeling in Bedup Basin, Malaysia. Int J Environ Sci & Tech. 2010; 9(10):9–18.
14. Guo W, Wang H. PSO optimizing neural network for the Yangtze river sediment entering estuary prediction. Sixth International Conference on Natural Computation (ICNC) 2010; 2010 August 10-12; Yantai, Shandong, China.
15. Muhammadi A, Akbari G, Azizian G. Suspended sediment concentration estimation using artificial neural networks and neural-fuzzy inference system case study: Karaj Dam. Indian Journal of Science and Technology. 2012 August; 5(8):3188–93.
16. Altunkaynak A. Sediment load prediction by genetic algorithms. Adv Eng Software. 2009 Sep; 40(9):928–34.
17. Rezapour OM, Shui LT, Dehghani AA. Genetic Algorithm Model for the Relation between Flow Discharge and Suspended Sediment Load (Gorgan River in Iran). **Electron J** Geotech Eng. 2011; 16:539–53. Available from: http://www.ejge.com/2011/Ppr11.039/Abs11.039.htm
18. Ebrahimi H, Jabbari E, Ghasemi M. Application of Honey-Bees Mating Optimization algorithm on Estimation of Suspended Sediment Concentration. World Appl Sci J. 2013; 22(11):1630–8.
19. May RJ, Maier HR, Dandy GC. Data splitting for artificial neural networks using SOM-based stratified sampling. Neural Networks. 2010 March; 23(2):283–94.
20. Salajegheh A, Fathabadi A. Estimation of the suspended sediment load of Karaj river using fuzzy logic and neural networks. J Range and Watershed Management. 2009; 62:271–82.
21. Kiani AY. Intelligent Systems. Tehran: Press Center of Poyesh Andisheh; 2009.
22. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons; 2009.
23. Mansourfar K. Advanced Statistical Methods Using Applied Software. Tehran: Press Centre of University of Tehran; 2009.
24. Jones KR, Berney O, Carr DP, Barrett EC. Arid Zone Hydrology. Rome: FAO Irrigation and Drainage Paper 37; 1981.
25. Moqaddam TR, Norouzi N, Kamali SM, Salamat-Bakhsh A. Meta-heuristic Algorithms. Tehran: Press Center of Islamic Azad University-South Tehran Branch; 2012.
26. Shahhoseini SH, Moosavi M, Mollajafari M. Evolutionary Algorithms. Tehran: Press Cent of Iran University of Science and Technology; 2011.