Environmental Science and Pollution Research (2018) 25:35693–35706
https://doi.org/10.1007/s11356-018-3533-6

RESEARCH ARTICLE

Modeling daily suspended sediment load using improved support vector machine model and genetic algorithm

Mitra Rahgoshay[1] · Sadat Feiznia[2] · Mehran Arian[3] · Seyed Ali Asghar Hashemi[4]

Received: 25 July 2018 / Accepted: 18 October 2018 / Published online: 24 October 2018

Abstract

Prediction of sediment volume and sediment load is always one of the important issues for decision-makers of watershed basins. The present study investigated the daily suspended sediment load in a watershed basin using the improved support vector machine method. Since in most of the previous studies, the coefficients of the support vector machine method had been calculated based on trial and error, in the present study, the combination of the support vector machine and the genetic algorithm is used. In the first step, the unknown parameters of the support vector machine are calculated and then, the sediment load simulation is performed. Two case studies in the present work involve two earth dams in Semnan Province called Veynakeh and Royan. Furthermore, multivariate adaptive regression spline (MARS) and MT tree model (M5T) methods are used for comparison. The results indicated that the input combination of discharge data at the current time and one, two, and three previous days has the best performance for all models. Also, the support vector machine-genetic algorithm (SVM-GA) model has a lower root mean square error (RMSE) and mean absolute error (MAE) compared to the MARS and M5T models for both stations. In addition, comparing observational data with simulation data based on the R^2 coefficient suggested that the SVM-GA model offers more accurate results than the other two methods. Accordingly, the SVM-GA method used in this study has a high potential for simulating sediment volume.

Keywords Sediment · Genetic algorithm · Support vector machine · Tree model

Introduction

One of the important problems in relation to watershed basin is sedimentation. Sediments have the potential to transport nutrients and contaminants (Liu et al. 2018; Yilmaz et al. 2018). Also, various environmental, hydrological, and hydraulic issues are associated with sedimentation and are affected by the volume of sediments (Gholami et al. 2018; Adib and Mahmoodi 2017). Also, construction of hydraulic structures in different parts of a watershed basin is influenced by sediment volume and, thus, the calculation of sediment load is crucial (Choubin et al. 2018; Moeeni and Bonakdari 2018). The process of erosion and sedimentation as an intensifying process leads to the loss of fertile soil of agriculture

Responsible editor: Marcus Schulz

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11356-018-3533-6) contains supplementary material, which is available to authorized users.

✉ Sadat Feiznia
sfeiz@ut.ac.ir

Mitra Rahgoshay
mitra.rahgoshay@srbiau.ac.ir

Mehran Arian
mehranarian@srbiau.ac.ir

Seyed Ali Asghar Hashemi
sa.hashemi@areo.ir

[1] Department of Earth Sciences, Science and Research Branch, Islamic Azad University, Tehran, Iran

[2] Faculty of Natural Resources, University of Tehran, Karaj, Iran

[3] Department of Earth Sciences, Science and Research Branch, Islamic Azad University, Tehran, Iran

[4] Department of Watershed Management, Agricultural and Natural Resources Research and Education Center, Semnan, Iran

and irreparable damage to water construction projects, including accumulation of sediments behind dams and reduction of their useful volume, destruction of structures, damages to beaches and ports, reduction of the capacity, elevation of the maintenance cost of irrigation canals, etc. (Lin et al. 2018; Kumar et al. 2018). So calculating sediment volume for decision-makers in watersheds is very important. In recent years, researchers have used artificial intelligence methods to calculate sediment volumes (Moeeni and Bonakdari 2018; Zamani et al. 2018; Negm et al. 2018; Adarsh and Reddy 2018; Hatten et al. 2018). Given that many parameters affect the volume of sediments, therefore, software computation methods, with the massive amount of data received, have a high potential to predict the calculation of sediment volumes. They are also highly accurate and have high adaptability to hydrological and hydraulic conditions (Wu et al. 2018). In addition, statistical methods and regression models are the next priorities for predicting sediment volume. If the quality of these methods can be improved based on some unknown parameters using evolutionary algorithms, we can use regression models as a powerful tool for predicting the volume of sediments (Lang et al. 2018; Ahilan et al. 2018).

History

Neural network and wavelet models have been used to predict the daily sediment volume (Rajaee et al. 2010). The study was based on daily precipitation data as well as daily discharge to predict daily suspended sediment load. The results indicated that the correlation coefficient of the neural network and wavelet was greater than that of the support vector model. Also, the root mean square error (RMSE) was significantly reduced based on the neural network and wavelet method.

Nourani et al. (2012) used the improved neural network method based on genetic algorithm to simulate sediments. The calculation of the number of hidden layers and the number of hidden neurons in the neural network was performed based on the genetic algorithm. The results revealed that the improved neural network method based on daily discharge data reduced the RMSE error by 30% compared to the tree model.

Kisi and Shiri (2012) used genetic programming models as well as tree models to predict sediment load. The results showed that the genetic programming model based on daily discharge values with a delay of 1 day, 2 days, and 3 days had a greater ability to predict the daily suspended sediment load compared to the tree model.

Kisi (2012) used the least square support vector machine to predict daily suspended sediment load. The results indicated that the method had a higher correlation coefficient in comparison with the simple support vector method in calculating the daily suspended sediment load.

Liu et al. (2013) used the wavelet support vector and the wavelet neural network to predict the daily suspended sediment load. Daily precipitation and daily discharge data were obtained based on both models. The results indicated that the wavelet neural network method reduces the absolute mean error between simulation and observational data.

Singh et al. (2014) used different neural networks to predict the daily sediment load. The results revealed that the multi-layer neural network offered results with a higher correlation coefficient compared to the radial neural network. The number of hidden layers and hidden neurons in the neural network was also calculated based on the particle swarm algorithm.

Afan et al. (2015) used the forward and backward neural network to calculate the daily suspended sediment load. The results showed that the existence of precipitation data with a long delay would not have any significant effect on the improvement of the results. The performance of the forward neural network was also more accurate than that of the backward neural network.

Skardi et al. (2015) used the neural network method as well as the artificial bee colony algorithm to predict suspended sediment load. The results indicated that the improved neural network method based on the artificial bee colony algorithm had accurate results based on the exact determination of the number of hidden layers and the weight used in the network.

Kumar et al. (2016) employed the improved support vector method and a least square vector support method to calculate the daily suspended sediment load. The results showed that the improved vector support method based on the genetic algorithm and the results of calculating unknown parameters were precise, as the correlation coefficients of the improved vector support method were significantly higher than those of the least square support vector method.

Nourani et al. (2016) predicted the daily suspended sediment load using the improved least square support vector machine based on genetic algorithm. The results indicated that the genetic algorithm could generate high accuracy in the calculation of the daily suspended sediment load by accurately calculating the parameters of the least square support vector machine, where the RMSE error was 20% less than that of the genetic programming method for calculating the suspended sediment load.

Adib and Mahmoodi used a multi-layer neural network method and a genetic algorithm to calculate the suspended sediment load. The results demonstrated that the calculated weights of the network based on the genetic algorithm led to a more precise structure of the network. This caused the simulated daily sediment load values to have high correlation coefficients.

Malik et al. (2017) utilized fuzzy methods, genetic programming, and the spline model to calculate the daily suspended sediment load. The fuzzy methods used in this study were improved with the help of the particle swarm algorithm. The results revealed that the absolute mean error of the fuzzy method by receiving daily discharge data was higher than that of the genetic programming and tree model, as the mean absolute error based on the improved fuzzy method was lower compared to other methods.

Objectives and innovations

The present study calculates the daily suspended sediment load based on the improved support vector method. Past research has shown that the aforementioned method has a good ability in simulating sediment volumes. However, the unknown parameters in previous studies have been computed based on trial and error (Sahraei et al. 2018; Yilmaz et al. 2018; Liang et al. 2017). The trial-and-error method is not a precise method for calculating the parameters of regression models as these parameters have a great effect on the final accuracy of the results. Therefore, the present study intends to calculate the unknown parameters of the vector support model using the genetic algorithm and then use the improved model to calculate the sediment volume. In other words, the values of the unknown parameters of the regression model are included in the genetic algorithm as the decision variable, then the optimal values of the variables are calculated. Accordingly, two homogeneous earth dams in Semnan Province, Iran, whose sediment volume is important for the management of watershed basins and construction of various hydraulic projects, are calculated. Also, the results are compared with the tree and spline models which are used as models for calculating sediments (Talebi et al. 2017; Roushangar and Ghasempour et al. 2017).

Methods

Support vector machine

The support vector machine is used to analyze the time series regression of variables in order to predict and simulate the exact variables. The linear form of the support vector model is based on the following relationship (Kisi et al. 2017):

$$f(x) = W^{\mathrm{Tr}}.x + b \tag{1}$$

where $f(x)$ is the objective variable estimated by SVM, W represents the input weight coefficient, b is the bias, and Tr denotes the transient symbol. SVM tries to decrease the difference between observational and simulation data. Therefore, based on an optimization process, the SVM reduces the objective function which is minimization of error. This error function ignores errors that are less than the threshold ε.

$$Minimize \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\left(\xi_i^- + \xi_i^+\right) \tag{2}$$

$$\begin{aligned} &subject(to)(w_i.x_i + b) - y_i < \varepsilon + \xi_i^+ \\ &y_i - (w_i.x_i + b) \le \varepsilon + \xi_i^- \end{aligned} \tag{3}$$

where C is the penalty coefficient, ξ_i^- and ξ_i^+ are the penalties for training data whose prediction error is outside the permissible range, m denotes the number of training data, w shows the weight, x represents the input variable, and y is the observational variable. The values of w and b are calculated from Eqs. 2 and 3. Then, the values of the above parameters are substituted in Eq. 1 to calculate $f(x)$. The SVM model can be used to predict and analyze nonlinear time series. Thus, Eq. 1 is rewritten according to the following relation:

$$f(x) = w^{\mathrm{Tr}}.K(x, x_i) + b \tag{4}$$

where $K(x, x_i)$ is the kernel function. The kernel functions are different. Previous studies have shown that the radial basis kernel functions are successful compared to other kernel functions where the simulated and observational results have high compatibility (Bharti et al. 2017; Sahraei et al. 2018):

$$K(x, x_i) = \exp\left(-\frac{|x - x_i|^2}{2\gamma^2}\right) \tag{5}$$

γ denotes the kernel parameter. Also, the SVM method has unknown parameters ε and C. All three parameters were calculated in the previous research based on the trial-and-error process, which did not result in high accuracy (Kisi et al. 2017, Sahraei et al. 2018; Yilmaz et al. 2018; Liang et al. 2017).

Genetic algorithms

Genetic algorithms are one of the most successful evolutionary algorithms in solving various optimization problems such as image processing, mathematical functions, engineering optimization problems, and other practical problems (Gil et al. 2018; Mousavi-Avval et al. 2017). First, a primary population of different solutions is generated randomly and, in an iterative process, subsequent populations are generated to improve the objective function. At each stage, people from the current population are selected to generate children or the next generation. Accordingly, individuals with a better performance are more likely to be selected. Selected individuals generate the next population based on the mutation and combination operator. The following relations are used for the combination operator:

$$Pop_i^{\mathrm{new}} = \alpha pop_i^{\mathrm{old}} + (1-\alpha)pop_j^{\mathrm{old}} \tag{6}$$

$$Pop_j^{\mathrm{new}} = \alpha Pop_j^{\mathrm{old}} + (1-\alpha)Pop_i^{\mathrm{old}} \tag{7}$$

where Pop_i^{new} represents the *i*th child, pop_i^{old} is the *i*th parent, Pop_j^{old} shows the *j*th parent, Pop_j^{new} denotes the *j*th child, and α is a random number. The mutation is applied according to the following relation:

$$Pop_{j,i}^{\mathrm{new}} = Var_{j,i}^{\mathrm{low}} + \beta\left(Var_{i,j}^{h_i} - Var_{j,i}^{\mathrm{low}}\right) \qquad (8)$$

where $Pop_{j,i}^{\mathrm{new}}$ is the *i*th new gene in the *j*th chromosome, $Var_{i,j}^{h_i}$ shows the lower limit of the *i*th gene in the *j*th chromosome, and β is a random number between 0 and 1. In combination, the generation of both new individuals occurs by changing the gene between the two individuals. The mutation operator is used to alter the chromosomes and to transform the genes to create diversity.

Support vector method and genetic algorithm

Given that various parameters are associated with unknown values, calculated in the previous research based on the trial-and-error process (Kisi et al. 2017; Kisi et al. 2017, Sahraei et al. 2018; Yilmaz et al. 2018; Liang et al. 2017), the genetic algorithm is used in the present study to calculate parameters with unknown vector support values:

Random parameters of the swarm genetic algorithm of the initial population, the probability of mutation and combination are determined. Then, an initial value is defined for the values of the unknown parameters of the SVM method.

The input and output parameters are specified and based on the highest correlation between the input parameters and the suspended sediment load, where the best combination of the input data is determined to simulate the amount of sediment.

Then, the input data is learned based on the SVM method. Also, the value of the objective function which in this study is RMSE is calculated.

The convergence criterion is controlled. If the convergence criterion is satisfied, the SVM method, based on the best values of its calculated parameters, performs the test step to calculate the suspended sediment load. Then, the results are extracted; otherwise, the algorithm proceeds to the next step.

The values of the SVM-related parameters including γ, ε, and C as input populations are introduced into a matrix based on the genetic algorithm. Indeed, these parameters are known as decision variables whose optimum value is calculated based on the particle swarm algorithm.

Then, the mutation and combination operators are applied to the population and then go back to step 3. Figure 1 demonstrates the stages of this structure.

Tree model (M5 tree model (M5T))

One of the regression models used by researchers as a popular model is the M5T model. It has a simple procedure to simulate hydraulic and hydrological variables and relatively accurate results (Heddam and Kisi 2018; Talebi et al. 2017).

The decision tree is for displaying a series of rules making a category or quantity. Decision trees are designed through sequential data separation into a separate set of groups, which intends to increase the distance between groups in the process of separation. In the tree construction process, an inference algorithm or division criterion is used to generate a decision tree. The dividing criterion for the model involves estimating the standard deviation of the class values reaching the node as a quantity of error and calculating the expected reduction in this error as the test result of each attribute in that node. Reduced standard deviation is calculated from the following equation:

$$\mathrm{SDR} = \mathrm{sd}(T) - \sum \frac{|T_i|}{|T|}\,\mathrm{sd}(T_i) \qquad (9)$$

where T represents a series of samples reaching the node, T_i denotes the samples of the *i*th output of the series, and sd shows the standard deviation. Due to the data branching process located at the child node, it has a lower standard deviation than that of the mother node and therefore it is more pure. After maximizing all possible splits, the adjective model is selected which maximizes the expected reduction. This division yields the pseudo-tree structure, contributing to greater fitting. To overcome the over-fit problem, the formed tree must be pruned. This is done by replacing the tree with a leaf. So the second step in designing a tree model is to prune a grown tree and replacing the subsidiary trees with linear regression functions.

Multivariate adaptive regression spline

The multivariate adaptive regression spline (MARS) method is known as one of the most applied and successful methods for simulating hydraulic and hydrological variables with a simple process (Deo et al. 2017; Kisi et al. 2017; Golkarian et al. 2018).

Nonparametric models are proposed if the structure of a model is not known before modeling. Among them, the multivariate adaptive regression spline can be mentioned. Also, if the used model uses the relevant data entirely, the model is general, while if the model divides the data, then the model is called local. This method divides the data into subsets, and in proportion to the complexity of data, in each region, it tries to fit functions called basic functions. The first stage of the model is known as a move forward. At this point, the model first starts with a constant term only and then repeatedly adds basic functions for modeling the data to a constant term. Indeed, a model generated with the highest fitness to the data involved in the modeling process. However, for the data that did not participate in the modeling process, it does not offer good fitness. The second stage is known as the backward movement

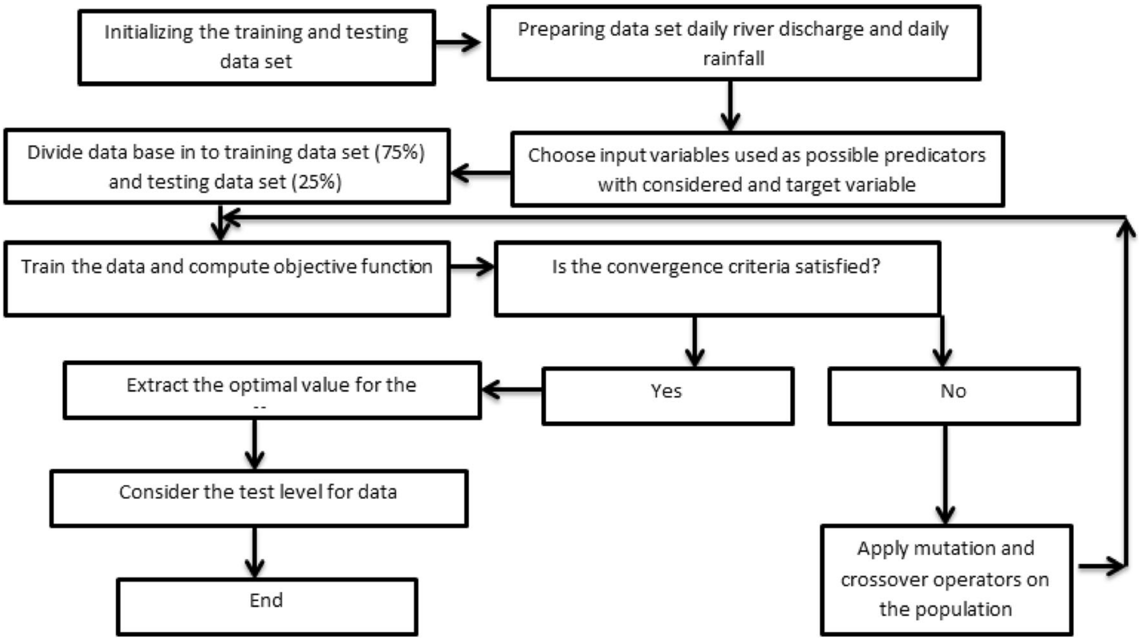
Initializing the training and testing data set
Preparing data set daily river discharge and daily rainfall
Divide data base in to training data set (75%) and testing data set (25%)
Choose input variables used as possible predicators with considered and target variable
Train the data and compute objective function
Is the convergence criteria satisfied?
Extract the optimal value for the --
Yes
No
Consider the test level for data
End
Apply mutation and crossover operators on the population

Fig. 1 SVM and GA for computation of sediment load

stage. This step is for pruning the model and eliminating the basis functions with the least effect on the modeling process. To specify the submodels with the least effect on the modeling, we use generalized cross-validation. This kind of validation is a kind of interaction between fit and complexity of the model in such a way that the model fitness with the data rises in the forward movement step by adding basis functions, and in response, the complexity of the model grows. Adaptive regression is performed according to the following equation:

$$\hat{Y} = \hat{f}(x) = a_0 + \sum_{m=1}^{M} a_m B_m(X) \tag{10}$$

where a_0 is the constant value, M denotes the number of nonzero terms, a_m shows the coefficient of the mth basis function, and $B_m(X)$ represents the mth basis function for the model, which is calculated according to the following equation.

$$B_m(x) = \prod_{i=1}^{k_m} \left[S_{i,m}\left(X_{v(i,m)} - t_{i,m}\right)\right]_+^q \tag{11}$$

k_m is the degree of interaction between the variables; $S_{i,\ m} = \pm 1$ and $X_{v(i,\ m)}$ represent the variable v, where $1 \leq v(i, m) \leq k$; and k is the total number of input variables. Also, $t_{i,\ m}$ shows the node position on each prediction variable of the dependent variable. The crossover validation is generalized based on the following relation:

$$\mathrm{GCV} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[y_i - f(x_i)^2\right]}{\left[1 - \left(\frac{C(M)}{n}\right)^2\right]} \tag{12}$$

where GCV is the generalized cross-validation, y_i represents the actual values of the class, $f(x_i)$ denotes the estimated value for the actual values of the class, n represents the total number of observations, and $C(M)$ is the cost criterion-penalty of the model.

Case study

Two earth dams are considered for estimating the daily suspended sediment load in Semnan Province in Iran. The features of the dams studied include homogeneity, noncore, and the type of earth dam. The goal of the dam is to control flood and store groundwater feeds. Thus, a large amount of sediment has been trapped in these two dams. Royan is located at the longitude 53° 39′ 00 and latitude 35° 43′ 17.28″ and is known as the first dam. Further, the second dam is called Veynakeh with longitude 53° 00′ 11.25″ and latitude 35° 37′ 7.9″. Figure 2 illustrates the position of the dams.

The items selected in the study are more than 50 earth dams. The basis for choosing includes the following criteria:

1. Minimum operation lifetime of 10 years
2. No significant deposition has occurred during the years of operation

This issue is investigated based on adequate evidence in the reservoir, overflow situation, and local information from local people and experts.

Table 1 and Fig. 3 presents the statistical characteristics of the two stations' relevant data between 2000 and 2010. The input data of this study are precipitation and discharge rates.

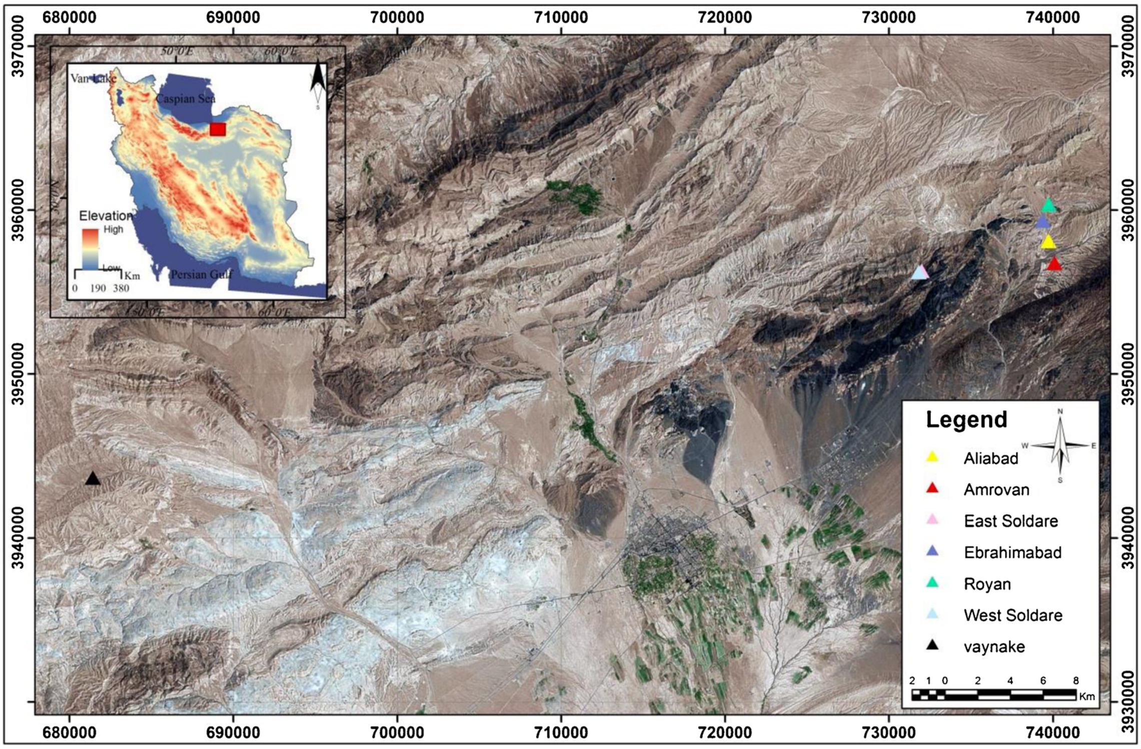

Fig. 2 Location of case study

The recorded statistical values indicate the hydraulic complexity of the flow and sediment. For example, the values of the coefficient of variation of data related to precipitation, discharge, and numerical deposition are significant. The skewness coefficient is also significant for the discharge and sediment data representing the complex conditions of modeling at both stations. The following indicators are also used to evaluate different methods:

- Root mean square error (RMSE)

$$\mathrm{RMSE} = \sqrt{\sum_{t=1}^{N} \frac{\left(X_{\mathrm{obt}} - X_{\mathrm{st}}\right)^2}{N}} \tag{13}$$

- Mean absolute error (MAE)

$$\mathrm{MAE} = \sum_{i=1}^{N} \frac{\left|X_{\mathrm{obt}} - X_{\mathrm{st}}\right|}{N} \tag{14}$$

- Nash Sutcliff

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N} \left(X_{\mathrm{obt}} - X_{\mathrm{st}}\right)^2}{\sum_{I=1}^{N} \left(X_{\mathrm{obt}} - \overline{X}_{\mathrm{obt}}\right)^2} \tag{15}$$

In addition, the following equation is used to calculate the correlation coefficient between the suspended sediment load and various input parameters:

$$\rho_{x,y} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{E\left[\left(X - \mu_x\right)\left(Y - \mu_y\right)\right]}{\sigma_x \sigma_y} \tag{16}$$

where X_{obt} denotes the observational data, X_{st} represents the simulated data, N is the number of data, $\overline{X}_{\mathrm{obt}}$ shows the mean number of observational data, $\rho_{x,\ y}$ is the correlation coefficient, cov(X, Y) is the covariance between the quantitative variables X and Y, μ_x and μ_y indicate the mean of X and Y respectively, and E is the expected value.

Results and discussion

Investigation of sediment load correlation and various input parameters

To simulate the daily suspended sediment load, first, the correlation values of different parameters with the amount of daily sediment load have been investigated. Figure 4 illustrates different parameters with different time delays and their correlation with the daily suspended sediment load. Parameters Q_t, Q_{t-1}, Q_{t-2}, and Q_{t-3} represent the current time discharge as well as discharge of one, two, and three

Table 1 The statistical parameters of the data set of the stations

Station	Data set	Data type	X_{mean}	X_{max}	X_{min}	S_x	C_v	c_{sx}
Royan	Training	Sediment load (Ton)	5.25	7.12	0.30	9.01	1.71	2.24
		Precipitation (mm)	18.14	27.29	9.11	23.4	1.32	3.41
		Discharge (m^3/s)	1.565	3.254	0.90	6.27	4.00	4.45
	Test	Sediment load (Ton)	5.454	6.87	0.34	8.65	1.58	5.12
		Precipitation (mm)	16.12	23.12	7.25	22.11	1.37	3.87
		Discharge (m^3/s)	1.871	3.872	0.70	5..451	2.12	5.12
Veynakeh	Training	Sediment load (Ton)	8.14	12.25	1.12	8.91	1.09	3.11
		Precipitation (mm)	19.12	28.14	12.11	24.12	1.26	2.78
		Discharge (m^3/s)	12.25	9.78	3.44	14.87	1.21	3.67
Veynakeh	Test	Sediment load (Ton)	8.21	12.14	1.11	9.11	1.10	4.55
		Precipitation (mm)	20.12	27.67	12.44	21.24	1.05	5.67
		Discharge (m^3/s)	12.45	14.56	2.98	14.35	1.16	3.12

previous days, respectively. On the other hand, the parameters R_t, R_{t-1}, R_{t-2}, and R_{t-3} show the current time precipitation along with the precipitation of one, two, and three previous days, respectively. It is clearly evident that the parameter Q_t for both stations has the highest correlation with the daily suspended sediment load. The longer the delay in the magnitude of daily flow, the lower the effect on the suspended load. Also, a comparison of correlation coefficients for precipitation values indicates that they are less effective and less correlated compared to the discharge rate. In particular, the 3-day and 2-day delay times to the plain reduced the correlation coefficient between precipitation and the suspended sediment load. Therefore, based on Fig. 4, the following input combinations are recommended for both stations:

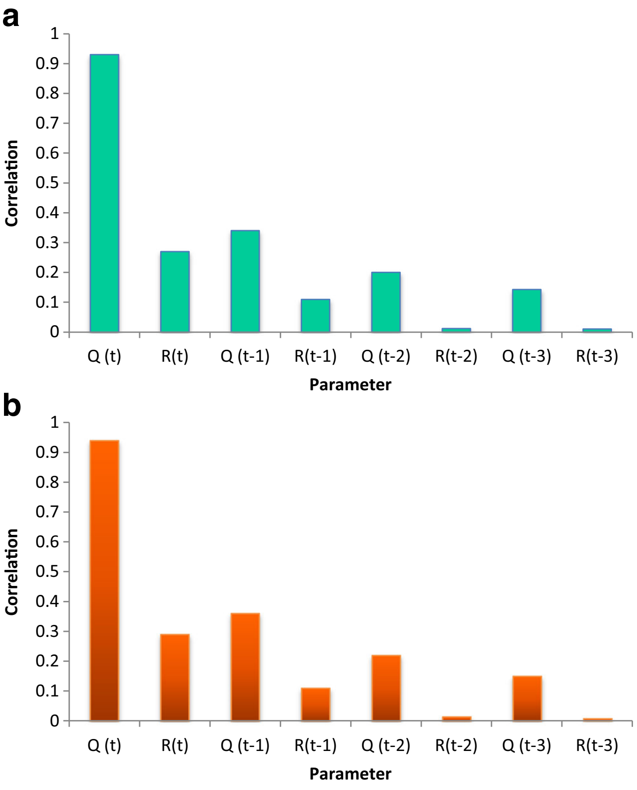

Fig. 3 The computation of correlation for **a** Royan and **b** Veynakeh

$$Q_t, R_t, Q_{t-1}, R_{t-1}, Q_{t-2}, Q_{t-3} \rightarrow (\text{SVM}-\text{GA1}, \text{M5T1}, \text{MARS1})$$
$$Q_t, R_t, Q_{t-1} \rightarrow (\text{SVM}-\text{GA2}, \text{M5T2}, \text{MARS2})$$
$$Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3} \rightarrow (\text{SVM}-\text{GA3}, \text{M5T3}, \text{MARS3})$$
$$Q_t, R_t, Q_{t-1}, Q_{t-2} \rightarrow (\text{SVM}-\text{GA4}, \text{M5T4}, \text{MARS4})$$

Royan station

Table 2a reports the performance of the M5T model at the test and final stages for simulating the daily suspended sediment load. The results indicate that the lowest RMSE is associated with the M5T3 model, which has flow inputs in terms of the current time discharge rate, 1-day-delay discharge rate, 2-day-delay discharge rate, and 3-day-delay discharge rate. It has also a better performance than other M5T models, as the RMSE index in terms of M5T3 is 59%, 38%, and 26% lower than that of the M5T1, M5T2, and M5T4, respectively. Further, the M5T3 has the lowest value for the MAE coefficient and the highest value for the NSE coefficient. The weakest performance among the M5T models belongs to the M5T1, where the present-day precipitation and 1-day delay precipitation data had no positive effect on sediment load simulation. Accordingly, the values of error indices based on M5T1 are higher than those of the other tree models. Table 2b presents the performance of different MARS models. The MARS3 model has a better performance than the other models of MARS, based on lower values for the RMSE and MAE indices. For example, the MAE index for the MARS3 model is 56%, 40%, and 10% lower than that for the MARS1, MARS2, and MARS4 models. Although the MARS1 model has more inputs compared to the MARS3 model, the number of inputs does not guarantee to improve the results, since the inclusion of precipitation data has reduced the quality of the model due to lower correlation with the suspended sediment load. The comparison of the results of the MARS models and of the tree model suggests that the M5T model has a better performance. For example, the best performance of the models based on MARS and M5T has been associated with MARS3 and M5T3 where the RMSE and MAE values for M5T3 have been 19% and 15% lower than those of MARS. Other M5Ts have also a better performance compared to MARS. Table 2c reveals the performance of different models for SVM-GA. The best performance of the models is represented by the lowest values for the RMSE and MAE indices and the highest for NSE. For example, the RMSE index for SVM-GA3 is 51%, 7.7%, and 1.2%, respectively lower compared to the index for SVAM-GA1, SVM-GA2, and SVM-GA1. As with the two previous models, discharge rates have been the best in terms of current time and time delays. Adding precipitation data at the present time causes the SVM-GA1 model to have a worse performance than other models. The comparison of the results of the SVM-GA model with MARS and M5T suggests that the SVM-GA has a superior performance over the two models. For example, SVM-GA3 with MARS3 and M5T3 as the best models for MARS and M5T shows that the RMSE index for SVM-GA3 is lower than that for MARS3 and M5T3 by 27% and 41%. In

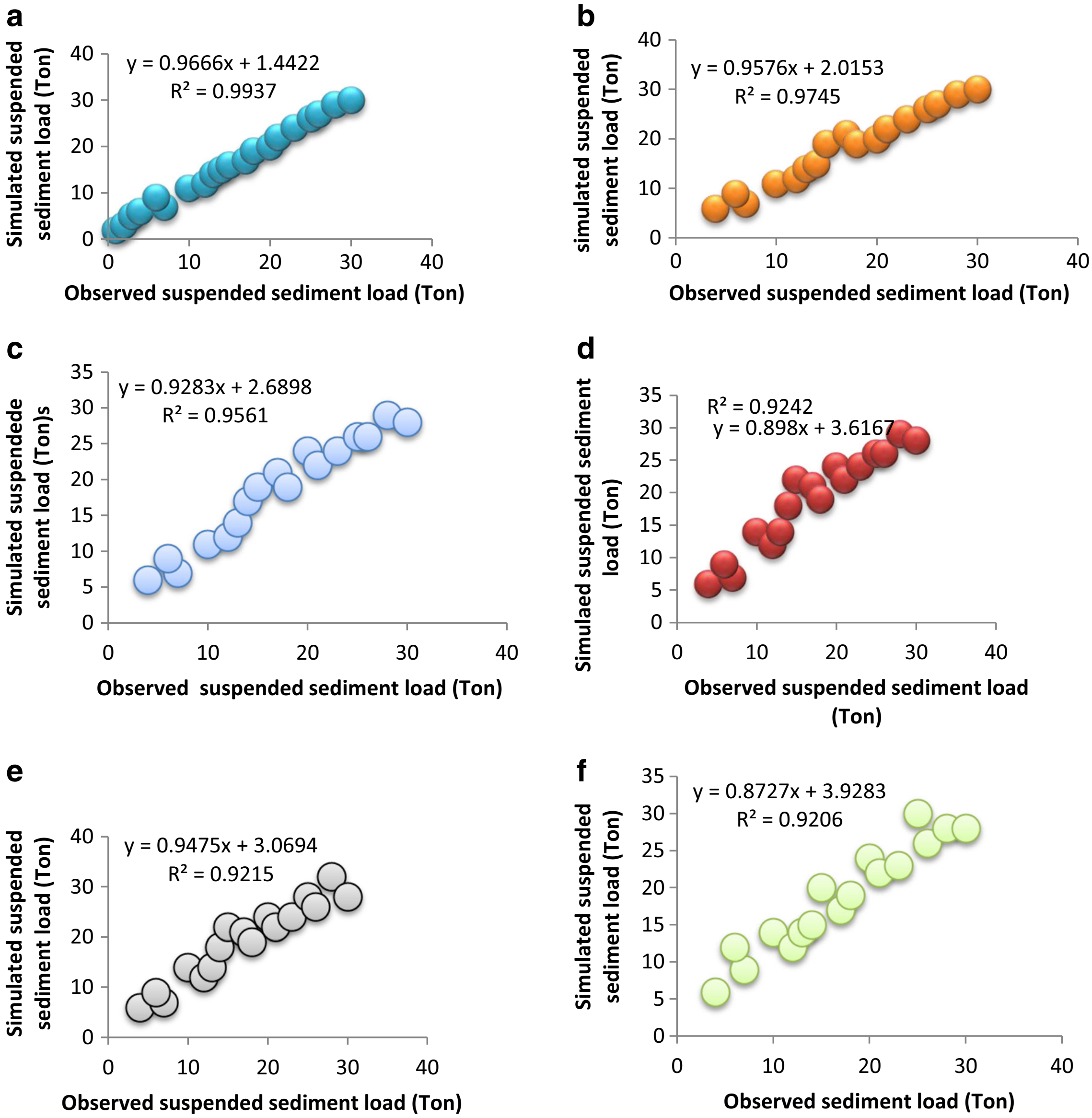

Fig. 4 The computation of R^2 for Rooyan station, **a** SVM-GA3, **b** SVM-GA4, **c** SVM-GA2, **d** SVM-GA1, **e** M5T3, **f** M5T4, **g** M5T2, **h** M5T1, **i** MARS3, **j** MARS4, **k** MARS2, and **l** MARS1

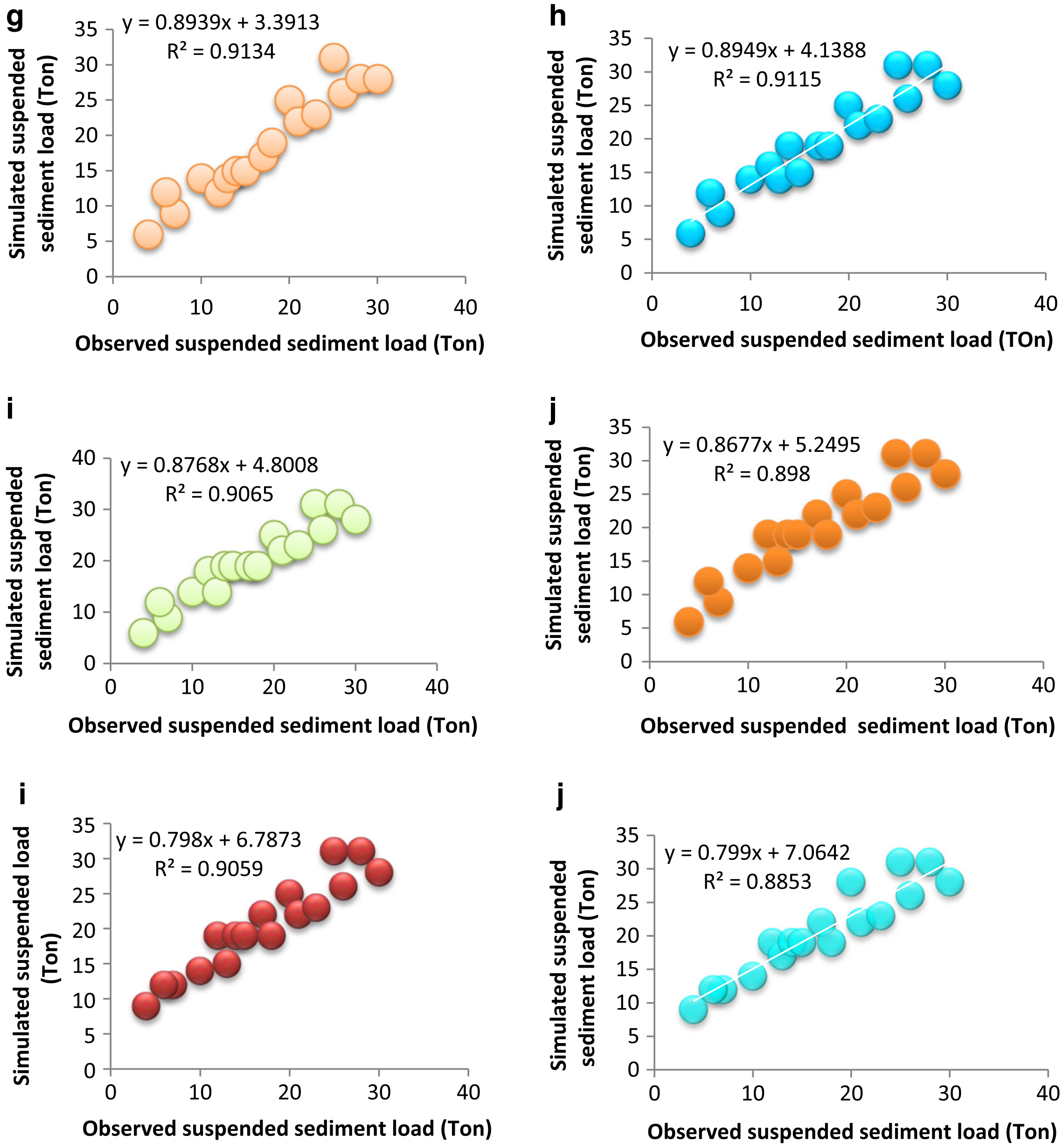

Fig. 4 continued.

addition, the SVM-GA3 model has a lower MAE index and higher NSE index compared to MARS and M5T models. Also, the optimum coefficient values for the SVM-GA model are shown in Table 2c. In addition, for setting the parameters of the genetic algorithm, parameter changes were used against the changes in the objective function. When the value of a parameter causes the objective function to be minimized, the value of that parameter is selected as the best value; when the chromosome population was 50, the combination rate was 0.6 and the mutation rate was 0.7. Figure 4 displays the performance of different methods in terms of the R^2 coefficient. The results indicate that among the SVM-GA models, the SVR-GA3 model has a greater value for the R^2 coefficient, while the SVR-GA1 model has a lower value for this coefficient. In addition, the comparison of the results suggests that the M5T and MARS models have a better performance compared to the SVM-GA model. The best performance for all models belongs to the third type input, while the worst performance was observed for the first type input. In any case, the SVM-GA3 model had the best performance among the models.

Veynakeh station

Table 3a outlines the performance of the M5T models for the Veynakeh station. The highest NSE coefficient, which represents the performance of a model and its adaptation

Table 2 Evaluation of different methods for Rooyan Station: a M5T, b MARS, and c SVM-GA

Combination	RMSE (ton)	MAE (ton)	NSE
a			
M5T1	0.85	2.985	3.123
M5T2	0.87	2.001	2.212
M5T3	0.91	1.111	1.265
M5T4	0.89	1.615	1.718
b			
MARS1	0.84	2.988	3.224
MARS2	0.86	2.212	2.414
MARS3	0.90	1.312	1.567
MARS4	0.87	1.465	1.819
c			
SVM-GA1 $\gamma = 2.15$, $\varepsilon = 0.028$, $C = 12.8$	0.92	1.812	1.878
SVM-GA2 $\gamma = 2.19$, $\varepsilon = 0.027$, $C = 12.6$	0.94	0.894	0.989
SVM-GA3 $\gamma = 2.14$, $\varepsilon = 0.023$, $C = 12.5$	0.96	0.791	0.912
SVM-GA4 $\gamma = 2.17$, $\varepsilon = 0.021$, $C = 12.2$	0.93	0.912	0.924

to observational data, is related to the M5T3 model, with the NSE coefficient equal to 0.9. The comparison of the results reveals that the MAE index value for the M5T3 model is 64%, 54%, and 37% lower than that of the M5T1, M5T2, and M5T4 models. However, the combination of current time discharge as well as discharges of one, two, and three previous days has the best performance for the M5T model. The second priority to use the M5T model is associated with the fourth input, current time discharge as well as discharges of one and two previous days with the present time precipitation. It causes the M5T4 model to have a more appropriate performance following the M5T3 model. However, although the M5T1 model has more inputs, it has not shown much better performance with delayed precipitation. Table 3b summarizes the performance of the MARS models. The results indicate that the RMSE index for the MARS3 model is 46%, 32%, and 42% lower than that for the MARS1, MARS2, and MARS4 models. In addition, the MAE and NSE indicators also reveal the superior performance of the MARS3 model. Also, when comparing RMSE and MAE indices for the MARS3 and M5T3 models, the results indicate that the RMSE and MAE for the M5T3 model are 34% and 17% lower than those for the MARS3 model. Also, the comparison of the results suggests that all M5T models have a better performance than MARS. The combination of the first data inputs for the MARS model, such as the M5T model, has had a slightly weaker performance compared to the other models. Table 3c represents the various performances of the SVM-GA. The results indicate that the RMSE error value for SVM-GA3 is 50%, 21%, and 1.4% lower than those of SVM-GA1, SVM-GA2, and SVM-GA4, respectively. The NSE and MAE indicators also confirm that the SVM-GA3 model has a better performance than other SVM-GA models. The comparison of RMSE values for the three SVM-GA3, MARS3, and M5T3 models suggests that the RMSE for the SVM-GA3 model is 49% and 22% lower than that for the MARS3 and M5T3 models. The NSE and MAE indicators also confirm that all SVM-GA models have a better performance compared to the MARS and M5T models. Meanwhile, the values of the optimal vector coefficients of the support vector machine are specified in Table 2c. In the case of previous research, where the results of the support vector method were not always superior to those of other methods used, the present study was able to produce good results through the improved method based on genetic algorithm and accurate calculation of the coefficients of the method. Also, the results of the two stations showed that the combination of inputs of discharge based on the current time as well as one and two previous days offered the best performance, while increasing the number of inputs such as first type inputs does not provide any guarantee for the performance of the models. Figure 5 reveals the performance of different models based on the R^2 coefficient. The highest value of this coefficient belongs to the SVM-GA3 model, which is

Table 3 Evaluation of different methods for Veynakeh station: a M5T, b MARS, and c SVM-GA

Combination	RMSE (ton)	MAE (ton)	NSE
a			
M5T1	3.110	2.784	0.86
M5T2	2.411	2.212	0.87
M5T3	1.163	1.001	0.90
M5T4	1.618	1.612	0.89
b			
MARS1	3.312	2.915	0.84
MARS2	2.651	2.321	0.86
MARS3	1.785	1.211	0.89
MARS4	1.911	1.711	0.87
c			
SVM-GA1 $\gamma = 2.12$, $\varepsilon = 0.021$, $C = 12.2$	1.811	1.511	0.92
SVM-GA2 $\gamma = 1.14$, $\varepsilon = 0.025$, $C = 12.8$	1.151	1.001	0.94
SVM-GA3 $\gamma = 3.15$, $\varepsilon = 0.029$, $C = 12.5$	0.899	0.545	0.96
SVM-GA4 $\gamma = 2.17$, $\varepsilon = 0.028$, $C = 12.5$	0.912	0.724	0.93

equivalent to 0.9671, and is larger than that of other SVM-GA models. Also, comparing the performance of the M5T and MARS models with that of SVM-GA, both are weaker than the SVM-GA given the R^2 coefficient. In addition, M5T models have higher R^2 coefficients than MARS models. In any case, the indices used in the present research including MAE and RMSE, which represent the extent of difference between observed and simulated data, decreased by the modified SVM method compared to the tree model and MARS model based on all input data. In addition, the NSE index offered more precise results for both stations based on the improved SVM model, where the index value for all models was close to 1, suggesting better performance of the improved SVM model. Investigation of the R^2 coefficient for the modified SVM model indicated that the model is more accurate than other methods, with the value of the mentioned coefficient being close to one, suggesting desirable performance of the modified SVM model. Indeed, identifying the accurate value of the parameters of the SVM method based on genetic algorithm allows for enhancing the method's accuracy through the optimization process. This, in turn, helps the SVM model present a better performance compared to other previous studies.

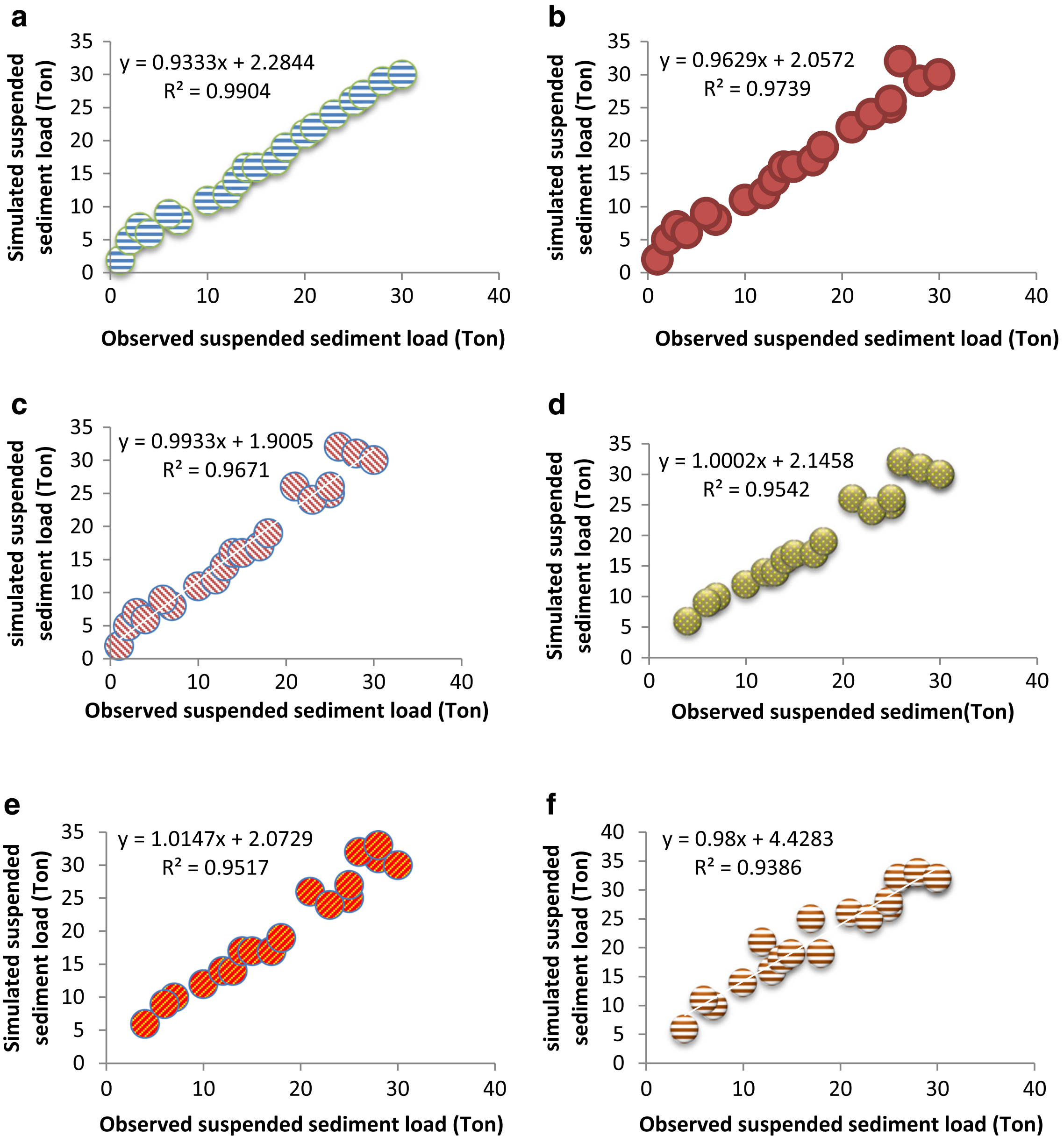

Fig. 5 The computation of R^2 for Veynakeh station, **a** SVM-GA3, **b** SVM-GA4, **c** SVM-GA2, **d** SVM-GA1, **e** M5T3, **f** M5T4, **g** M5T2, **h** M5T1, **i** MARS3, **j** MARS2, **k** MARS4, and **l** MARS1

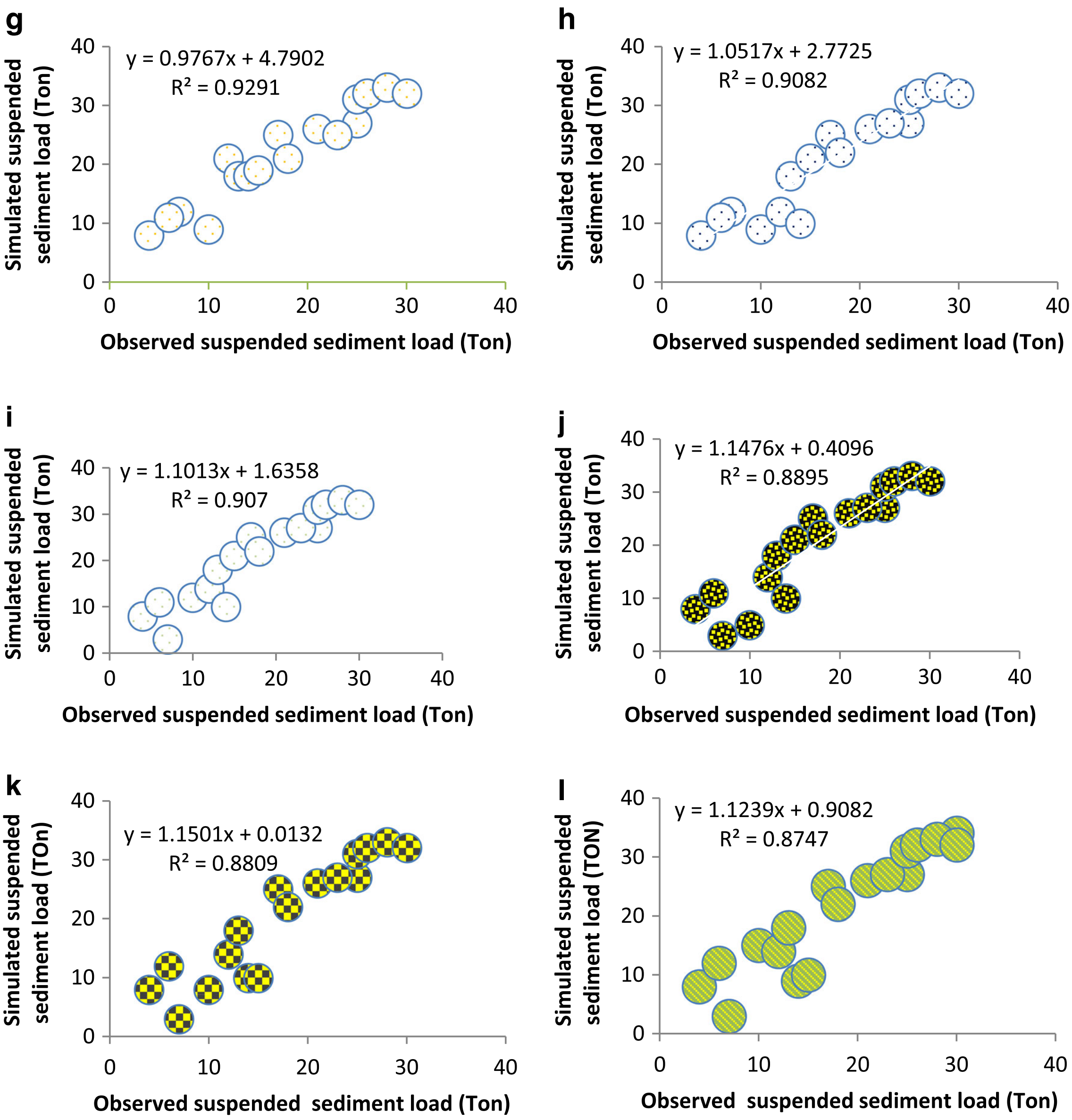

Fig. 5 continued.

Conclusion

The importance and effectiveness of sediment volume in a basin prompted the present study to investigate the models for predicting daily suspended sediment load. Since regression methods have unknown parameters, therefore, the exact calculation of parameters with an unknown value is very important. The support vector method is one of the regression methods which calculates unknown parameters in previous studies based on trial and error. However, in the present paper, these variables were included in the genetic algorithm. Then, based on the definition of the RMSE objective function, the best value for the parameters of the support vector models was calculated. A case study was conducted in Semnan Province with two earth dams called Royan and Veynakeh in order to calculate the daily suspended sediment load. Next, the results were compared with the findings of the MARS and M5T methods. The results indicated that inputs including the current time discharge as well as discharges of one, two, and three previous days had a better performance for all models. Further, adding precipitation inputs based on the current time and the day before led to reduced quality of the models in the simulation of suspended sediment load. Also, the correlation coefficients and RMSE and MAE errors for the SVM-GA3 model suggested that the model has the best performance among all SVM-GA, MARS, and M5T models. The comparison of the results revealed that the M5T model claims the second position in terms of good performance followed by the SVM-GA models. Also, the MARS models offered a good performance following the SVM-GA and M5T models. Future studies can be used to develop the SVM method with other advanced evolutionary algorithms such as bat and shark algorithms to develop the SVM method.

Author contribution A new method for the support vector machine was used to simulate the sediment load. All the authors contributed to the manuscript development by writing, discussing changes for clarity, and technical usefulness and correction.

References

Adarsh S, Reddy MJ (2018) Multiscale modelling of daily suspended sediment load using MEMD-SLR Coupled approach. In: Handbook of research on predictive modeling and optimization methods in science and engineering. IGI Global, pp 264–275

Adib A, Mahmoodi A (2017) Prediction of suspended sediment load using ANN GA conjunction model with Markov chain approach at flood conditions. KSCE J Civ Eng 21(1):447–457

Afan HA, El-Shafie A, Yaseen ZM, Hameed MM, Mohtar WHMW, Hussain A (2015) ANN based sediment prediction model utilizing different input scenarios. Water Resour Manag 29(4):1231–1245

Ahilan S, Guan M, Sleigh A, Wright N, Chang H (2018) The influence of floodplain restoration on flow and sediment dynamics in an urban river. J Flood Risk Manage 11:S986–S1001

Bharti B, Pandey A, Tripathi SK, Kumar D (2017) Modelling of runoff and sediment yield using ANN, LS-SVR, REPTree and M5 models. Hydrol Res 48(6):1489–1507

Choubin B, Darabi H, Rahmati O, Sajedi-Hosseini F, Kløve B (2018) River suspended sediment modelling using the CART model: a comparative study of machine learning techniques. Sci Total Environ 615:272–281

Deo RC, Kisi O, Singh VP (2017) Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. Atmos Res 184:149–175

Gholami V, Booij MJ, Tehrani EN, Hadian MA (2018) Spatial soil erosion estimation using an artificial neural network (ANN) and field plot data. Catena 163:210–218

Gil JM, Montes JFA, Alba E, Aldana-Montes JF (2018) Optimizing ontology alignments by using genetic algorithms

Golkarian A, Naghibi SA, Kalantar B, Pradhan B (2018) Groundwater potential mapping using C5. 0, random forest, and multivariate adaptive regression spline models in GIS. Environ Monit Assess 190(3):149

Hatten JA, Segura C, Bladon KD, Hale VC, Ice GG, & Stednick JD (2018) Effects of contemporary forest harvesting on suspended sediment in the Oregon Coast Range: Alsea Watershed Study Revisited. Forest Ecology and Management 408:238–248

Heddam S, Kisi O (2018) Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 559:499–509

Kisi O (2012) Modeling discharge-suspended sediment relationship using least square support vector machine. J Hydrol 456:110–120

Kisi O, Shiri J (2012) River suspended sediment estimation by climatic variables implication: comparative study among soft computing techniques. Comput Geosci 43:73–82

Kisi O, Parmar KS, Soni K, Demir V (2017) Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, and M5 model tree models. Air Qual Atmos Health 10(7):873–883

Kumar D, Pandey A, Sharma N, & Flügel WA (2016) Daily suspended sediment simulation using machine learning approach. Catena 138: 77–90

Kumar R, Kumar R, Singh S, Singh A, Bhardwaj A, Kumari A, Saha A (2018) Dynamics of suspended sediment load with respect to summer discharge and temperatures in Shaune Garang glacierized catchment, Western Himalaya. Acta Geophys:1–12

Lang Z, Li Y, Hu Y, Li B, & Wang J (2018) A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework. Theoretical and Applied Climatology 1–13

Liang Z, Li Y, Hu Y, Li B, Wang J (2017) A data-driven SVR model for long-term runoff prediction and uncertainty analysis based on the Bayesian framework. Theor Appl Climatol:1–13

Lin S, Qi J, Jones JR, Stevenson RJ (2018) Effects of sediments and coloured dissolved organic matter on remote sensing of chlorophyll-a using Landsat TM/ETM+ over turbid waters. Int J Remote Sens 39(5):1421–1440

Liu QJ, Shi ZH, Fang NF, Zhu HD, Ai L (2013) Modeling the daily suspended sediment concentration in a hyperconcentrated river on the Loess Plateau, China, using the wavelet–ANN approach. Geomorphology 186:181–190

Liu CG, Li ZY, Hao Y, Xia J, Bai FW, Mehmood MA (2018) Computer simulation elucidates yeast flocculation and sedimentation for efficient industrial fermentation. Biotechnol J 13

Malik A, Kumar A, Piri J (2017) Daily suspended sediment concentration simulation using hydrological data of Pranhita River Basin, India. Comput Electron Agric 138:20–28

Moeeni H, Bonakdari H (2018) Impact of normalization and input on ARMAX-ANN model performance in suspended sediment load prediction. Water Resour Manag:1–19

Mousavi-Avval SH, Rafiee S, Sharifi M, Hosseinpour S, Notarnicola B, Tassielli G, Renzulli PA (2017) Application of multi-objective genetic algorithms for optimization of energy, economics and environmental life cycle assessment in oilseed production. J Clean Prod 140:804–815

Negm A, Elsahabi M, Abdel-Nasser M, Mahmoud K, Ali K (2018) Impacts of GERD on the accumulated sediment in Lake Nubia using machine learning and GIS techniques

Nourani V, Kalantari O, Baghanam AH (2012) Two semidistributed ANN-based models for estimation of suspended sediment load. J Hydrol Eng 17(12):1368–1380

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 Resour Manag 30(1):393–407

Rajaee T, Nourani V, Zounemat-Kermani M, Kisi O (2010) River suspended sediment load prediction: application of ANN and wavelet conjunction model. J Hydrol Eng 16(8):613–627

Roushangar K, & Ghasempour R (2017) Estimation of bedload discharge in sewer pipes with different boundary conditions using an evolutionary algorithm. International Journal of Sediment Research, 32(4):564–574

Sahraei S, Alizadeh MR, Talebbeydokhti N, Dehghani M (2018) Bed material load estimation in channels using machine learning and meta-heuristic methods. J Hydroinf 20(1):100–116

Singh A, Imtiyaz M, Isaac RK, Denis DM (2014) Assessing the performance and uncertainty analysis of the SWAT and RBNN models for simulation of sediment yield in the Nagwa watershed, India. Hydrol Sci J 59(2):351–364

Skardi MJE, Afshar A, Saadatpour M, Solis SS (2015) Hybrid ACO–ANN-based multi-objective simulation–optimization model for pollutant load control at basin scale. Environ Model Assess 20(1):29–39

Talebi A, Mahjoobi J, Dastorani MT, Moosavi V (2017) Estimation of suspended sediment load using regression trees and model trees approaches (case study: Hyderabad drainage basin in Iran). ISH J Hydraul Eng 23(2):212–219

Wu L, Peng M, Qiao S, Ma XY (2018) Effects of rainfall intensity and slope gradient on runoff and sediment yield characteristics of bare loess soil. Environ Sci Pollut Res 25(4):3480–3487

Yilmaz B, Aras E, Nacar S, Kankal M (2018) Estimating suspended sediment load with multivariate adaptive regression spline, teaching-learning based optimization, and artificial bee colony models. Sci Total Environ 639:826–840

Zamani B, Koch M, Hodges BR, Fakheri-Fard A (2018) Pre-impoundment assessment of the limnological processes and eutrophication in a reservoir using three-dimensional modeling: Abolabbas reservoir, Iran. J Appl Water Eng Res 6(1):48–61