



# A Wavelet Based Data Mining Technique for Suspended Sediment Load Modeling

Vahid Nourani<sup>1,2</sup> · Amir Molajou<sup>3</sup> · Ali Davanlou Tajbakhsh<sup>1</sup> · Hessam Najafi<sup>1</sup>

Received: 8 October 2018 / Accepted: 17 February 2019 /

Published online: 22 February 2019

© Springer Nature B.V. 2019

## Abstract

The suspended sediment load (SSL) modeling generated within a catchment is a significant issue in the environmental and water resources planning and management of watersheds. The estimation methods of SSL are limited by the important parameters and boundary conditions (which are based on the flow and sediment properties). In this situation, soft computing approaches have proven to be an efficient tool in modelling the sediment load of rivers. In this study, the hybrid Wavelet-M5 model was introduced to model SSL of two different rivers (Lighvanchai and Upper Rio Grande) at both daily and monthly scales. In this way, first, the runoff and suspended sediment load time series were decomposed using the wavelet transform to several sub-time series to handle the non-stationary of the runoff and sediment time series. Then, the obtained sub-series were applied to M5 model tree as inputs. The obtained results for the Upper Rio Grande River at daily time scale, showed the better performance of Wavelet-M5 model in comparison with individual Artificial Neural Network (ANN) and M5 models so that the obtained Nash-Sutcliffe efficiency (NSE) was 0.94 by the hybrid Wavelet-M5 model while it was calculated as 0.89 and 0.77 by Wavelet-ANN (WANN) and M5 tree models, respectively. Also, the obtained NSE for the Lighvanchai River at monthly time scale was 0.90 by the hybrid Wavelet-M5 model while it was calculated as 0.78 and 0.69 by Wavelet-ANN (WANN) and M5 tree models in the verification step, respectively.

**Keywords** Suspended sediment load modeling · Runoff · Decision tree · M5 model tree · Wavelet transform · Artificial neural network

---

✉ Vahid Nourani  
vnourani@yahoo.com

Amir Molajou  
amolajou@yahoo.com

Ali Davanlou Tajbakhsh  
alidavanlou@yahoo.com

Hessam Najafi  
hessamnajafi71@gmail.com

Extended author information available on the last page of the article

## 1 Introduction

Suspended Sediment Load (SSL), known as a main proportion of total sediment load, has strong impacts on recreational and ecological issues, is considered as a significant variable in a river modeling. Estimation of SSL is considered as a challenging task because it is strictly related to the highly non-linear and complex interactions mechanisms of river streamflow (Sivakumar and Wallender 2005). The SSL of a river can be mannered as a function of hydro-meteorological parameters, which is known as a costly and complicated process. Besides, the accuracy of obtained results cannot be guaranteed by utilization of classic hydro-mechanic approaches (ALP and Cigizoglu 2007; Yang et al. 2009).

Classic time series models - such as Auto Regressive Integrated Moving Average (ARIMA) - are broadly utilized to forecast hydro-meteorological variables including SSL (Salas et al. 1980; Singer and Dunne 2001; Yang et al. 2009; Sharghi et al. 2018a, 2018b). However, the defect that can be attributed to classic time series models is that they are assumed as fundamentally linear models handling the stationary data and restricted to capture the non-linear and non-stationary characteristics of the hydrological data sets.

On the other hand, to handle the complex and non-linear characteristics of sediment time series, non-linear Artificial Intelligent (AI)-based models, especially Artificial Neural Network (ANN), have been widely implemented to design the functional relationship between inputs and outputs. Due to the remarkable advantages of ANN, such as benefiting black-box property (no need to prior knowledge), handling the non-linear property of the studied process by applying a non-linear function and the ability of analyzing multi-variate inputs with different characteristics, the application of ANN as an AI-based method has been investigated at different fields of engineering, including SSL modeling (Mustafa et al. 2012; Azad et al. 2018; Salimi et al. 2018).

Due to the multi-resolution nature of original raw SSL time series, the efficiency of mentioned models (i.e. ANN) to forecast highly non-stationary, autoregressive and seasonal SSL time series are decreased significantly. Under this situation, applying a suitable data pre-processing tool (wavelet transform) can be an appropriate solution to overcome the mentioned problems. Hidden frequencies and significant temporal information of original SSL time series can be extracted by wavelet transform. Hence, several studies have investigated the ability of wavelet transform in decomposing seasonal SSL time series into sub-time series at various temporal scales (levels) to extract inherent features (Partal and Cigizoglu 2008; Shiri and Kisi 2010; Belayneh et al. 2014).

It is noticeable that the specific features of SSL time series can be represented by the obtained wavelet-based sub-time series (Kuo et al. 2010). Previous studies showed that the performance of both short and long terms predictions is improved by employing wavelet technique. This improvement is more sensible in larger time scales (for example, seasonally/monthly than daily/hourly) because the seasonal (periodic) patterns are more dominant in the large-scale time series than the small-scale time series in most of hydrological process (Shiri and Kisi 2010; Nourani et al. 2018a; Sharghi et al. 2018b).

Some deficiencies can be attributed to hybrid WANN model despite of its reliable efficiency. It is believed that the network can identify the important data, hence the ANN users supply a large number of data as inputs. This can lead into the complex calculations, error, non-convergence and over-training (Nourani et al. 2018a).

Due to the scatter of time series data and mentioned deficiencies of WANN model, input data classification can be taken into account as an appropriate tool for complexity reduction of used

dataset. Decision tree (known as a hierarchical clustering method) is one of the efficient tools of data mining in classification, clustering and regression issues. Decision tree tries to accumulate the most similar observations (the basis of division is minimizing the available entropy among the sub-groups data) and provides a suitable regression for each accumulation. In fact, decision tree has a position between linear and non-linear models (as a multi-linear model) by providing piecewise linear functions. It also should be mentioned that the relationships presented by decision tree are simpler and more understandable by unprofessional users with regard to the complicated non-linear methods (Solomatine and Xue 2004; Nourani et al. 2018b).

M5 model tree is a subset of decision tree, machine learning and data mining methods which can detect the useful information from a dataset (Quinlan 1992). M5 model tree, unlike the other algorithms of decision tree, assigns a multi-variate linear regression model instead of fitting a constant value to the leaf node, so it is analogous to piecewise linear functions. In recent years, M5 model tree has evolved considerably in the classification and forecasting issues. M5 benefits can be stated as being more intelligible and much simpler in training stage than ANN, requires no trial and error, acceptable performance in dealing with missing data and large multi-dimensional problems (Solomatine and Xue 2004; Nourani et al. 2018a, 2018b). Numerous studies have reported comparable performances of M5 and ANN in hydrological simulation, particularly sediment modeling (Bhattacharya and Solomatine 2006; Senthil Kumar et al. 2011; Goyal 2014).

Owing the merits of both M5 and Wavelet tools in dealing with hydrological processes as an innovation, a hybrid model utilizing the wavelet transform and data mining features is proposed in this study for SSL simulation. A model with a set of linear regressions (multi-linear model) benefiting the wavelet transform may be reliable to handle the non-stationary nature of the process instead of using complex non-linear models. In this way, it is tried to eliminate the available trend of the main time series data of the studied rivers using the wavelet transform due to its ability to mitigate the effects of non-stationary. As the next step, the obtained wavelet-based sub-time series are classified by M5 model tree and finally, the appropriate regression models are presented for the classes. To evaluate the performance of Wavelet-M5 model while facing different hydro-climatological behaviors, the proposed hybrid model is applied on Lighvanchai and Upper Rio Grande with quite different behaviors. For this purpose the daily and monthly data are considered to perceive the efficiency of the model in dealing with autoregressive and seasonal characteristics of the process.

## 2 Materials and Methods

### 2.1 Study Area and Data Set

In the current research, the data for two cases studies, the Lighvanchai and Upper Rio Grande rivers, were used to implement and examine the proposed methodology. They are located in northwest Iran at Azerbaijan province and West of USA, at Colorado and Mexico States, respectively. The time series data of 29 years, from 1988 to 2016 for Lighvanchai River (Lighvanchai station located at 37°50' North latitude and 46°26' East longitude in the northern slope of Sahand Mountain); and 40 years, from 1977 to 2016 for Upper Rio Grande River (Otowi Bridge station located at 35°52' North latitude and 106°08' West longitude), were employed in the modeling process, in which the first 75% years and the rest 25% years were considered as training and validation purposes, respectively. The used SSL and water

discharge time series were derived from Iran Water and Power Resources Development Co. (IWPC) and the United States Geological Survey's website (USGS-<https://cida.usgs.gov/sediment/>). Table 1 shows the statistical characteristics of SSL data for both watersheds at daily and monthly time scales.

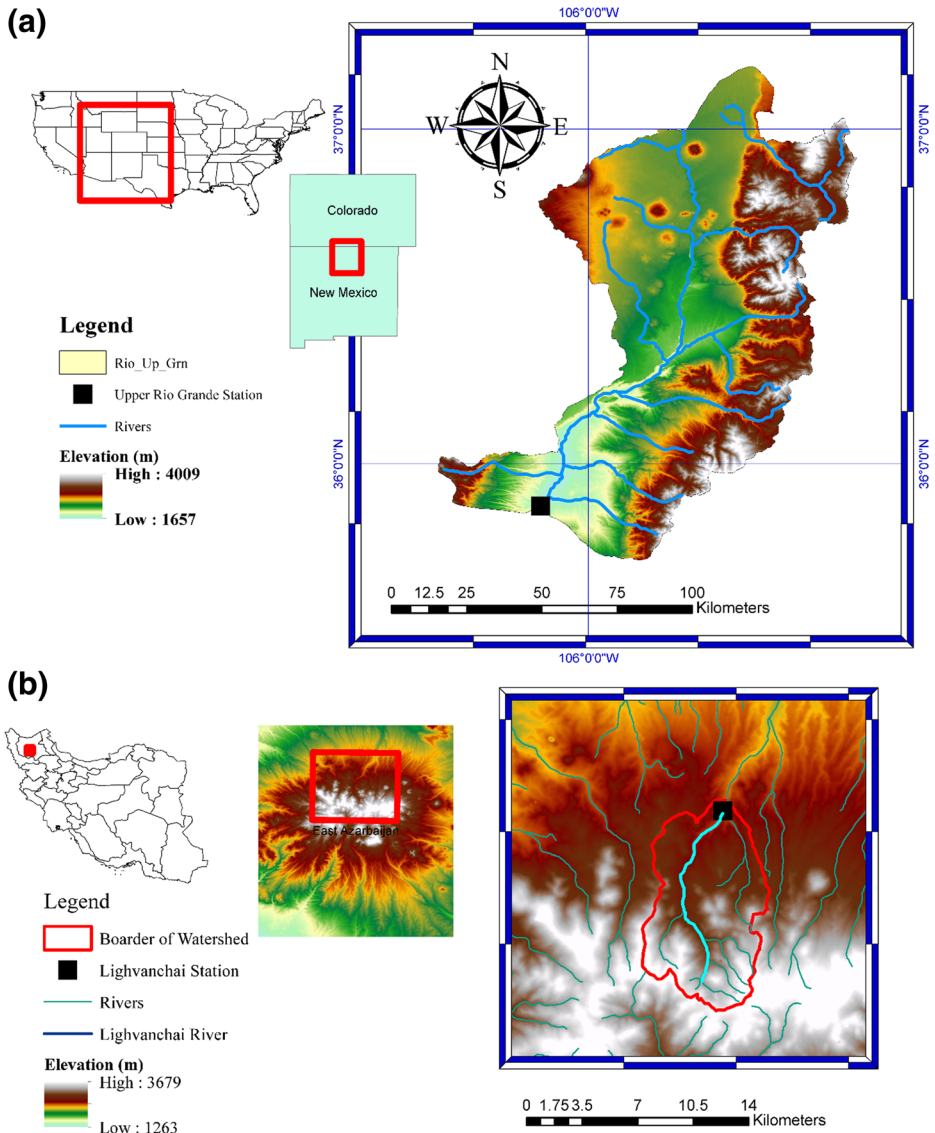
### 2.1.1 Case 1: Lighvanchai River

Lighvanchai River located at Azerbaijan province in the North-West of Iran. The related watershed is one of the major sub-tributaries of Ajichai River which discharges to Urmia Lake. Lighvanchai watershed is located between 37°43' and 37°50' North latitude and 46°22' and 46°28' East longitude in the northern slope of Sahand Mountain. The watershed area is approximately 142 km<sup>2</sup> (Fig. 1). Watershed altitude is varying between 1263 m and 3679 m above the sea level.

Azerbaijan province has a partly mountainous climate and a dry steppe with permanent lack of precipitation. The rainfall peaks in winter and spring. The dominant climate of Lighvanchai watershed is rainy and sub-humid having four distinct seasons. During the autumn and winter, the region is under the effect of middle latitude westerlies, and most of the precipitations that happen over the area during this period is initiated by depressions moving over the area, after developing in the Mediterranean Sea on a subdivision of the polar jet stream in the upper troposphere. The watershed contains medium vegetative land cover as a rural region. The

**Table 1** Statistics of the observed time series for Lighvanchai and upper Rio Grande Rivers

Scale	Time series	Statistical parameter	River					
			Lighvanchai		Upper Rio Grande			
			Calibration	Validation	Calibration	Validation		
Daily	SSL ( <i>Ton/Day</i> )	Mean	9.177	18.083	5001.003	5146.007		
		Maximum	264.847	395.883	349,187.495	444,762.968		
		Minimum	0.011	0.015	9.321	21.941		
		Standard deviation	18.451	29.900	13,400.179	16,622.151		
		Coefficient of Variation (dimensionless)	2.011	1.653	2.679	3.230		
	Runoff ( <i>m<sup>3</sup>/s</i> )	Mean	0.771	0.825	43.466	33.435		
		Maximum	8.650	6.050	339.802	254.002		
		Minimum	0.005	0.010	5.522	6.570		
		Standard deviation	0.852	0.890	42.775	29.152		
		Coefficient of Variation (dimensionless)	1.104	1.080	0.984	0.872		
		Monthly	SSL ( <i>Ton/Month</i> )	Mean	279.323	554.580	152,119.545	156,908.614
				Maximum	6243.396	4022.993	2,145,732.433	1,920,140.992
				Minimum	20.733	22.480	1743.410	1203.431
				Standard deviation	515.479	789.383	272,406.721	310,578.964
Coefficient of Variation (dimensionless)	1.845			1.423	1.791	1.979		
Runoff ( <i>m<sup>3</sup>/s</i> )	Mean	23.469	25.288	1322.250	1018.953			
	Maximum	161.460	113.000	7364.929	5323.567			
	Minimum	3.900	3.509	312.307	374.915			
	Standard deviation	23.640	24.136	1219.574	828.308			
	Coefficient of Variation (dimensionless)	1.007	0.954	0.922	0.813			



**Fig. 1** Location and the digital elevation models of the study areas a) Lighvanchai River and b) Upper Rio Grande River

topography is steep with average slope of 11%. Consequently the soils are disposed to erosion to some extended.

### 2.1.2 Case 2: Upper Rio Grande River

The Rio Grande (or Rio Bravo in Mexico) is an interstate and international stream. It rises in Colorado and flows southward for more than 643 km across New Mexico, and then forms the boundary between Texas and the United States of Mexico for about 1930 km to its mouth. The Upper Rio Grande stream runs 1100 km from its headwaters in Colorado through New Mexico

and northern Mexico to Ft. Quitman, Texas. Along its river corridor, it is a primary source of irrigation water for food, fiber and feed production and is used as a source for municipal supply by the cities of Albuquerque, Las Cruces, El Paso and Ciudad Juarez. The upper Rio Grande has a drainage area of about  $10,000 \text{ km}^2$ , less than a fifth of the water-producing area of the Rio Grande basin.

Considering the long-lasting concerns over the water of the upper Rio Grande, extensive records of streamflow are available for quite a lot of places along the main streamline and its main branches. With respect to the recent report of Natural Resources Conservation Service of the United States Department of Agriculture, the land cover of Upper Rio Grande is mostly evergreen forest (about 42% of total) and grasslands-herbaceous (about 40% of total) of the watershed's area which denotes that the area is almost green (see Fig.1).

With the precision in the standard deviation values of the runoff and SSL on the daily scale (Table 1), it could be concluded that Upper Rio Grande runoff and SSL data are more scattered than Lighvanchai. It can be inferred that the standard deviation of Lighvanchai data were close to 0 than Upper Rio Grande. It indicates that the data were closer to the mean value with a little dispersion (most daily runoff values about 0). The difference of standard deviation beside the correlation coefficient values for the runoff and SSL can justify the difference in the behavior of the case studies. For more analysis, the Auto Correlation Function (ACF) plots of daily and monthly data sets for both case studies have been presented in Fig. 2. As it can be seen in ACF plots, both case studies have seasonal and autoregressive patterns. By comparing both watersheds' characteristics (see Table 1 and ACF plots), Lighvanchai follows a regular pattern and linear behavior, thus can be considered as a well-dominated seasonal watershed while Upper Rio Grande follows the irregular and more non-linear pattern due to its denser land cover and bigger area. It can be inferred that the Upper Rio Grande watershed can be considered as a wild watershed with regard to the Lighvanchai watershed in generating the sediment load.

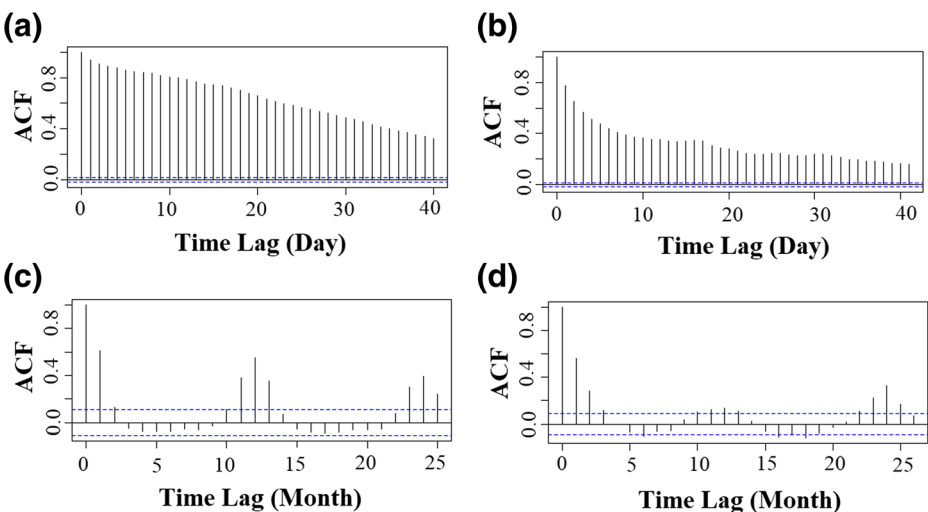


Fig. 2 Daily and monthly ACF plots for a, c) Lighvanchai River and b, d) Upper Rio River

## 2.2 Proposed Hybrid Methodology

The proposed hybrid Wavelet-M5 model consists of four steps. In the first step, the SSL and water discharge data are collected (Step 1). Applying an appropriate data pre-processing tool can improve the efficiency of any data-driven method (Sharghi et al. 2018a, 2018b). Wavelet analysis is one of the proposed method that can be helpful as an effective data pre-processing when dealing with seasonal and non-stationary processes and data (Step 2).

One of the main capabilities of wavelet transform is its ability to decompose the main time series into several sub-time series. Each of the obtained sub-time series has a specific feature (representing a specific frequency or seasonal period). In order to reduce the complexity of the original time series, wavelet-based decomposition is employed to analyze the seasonal feature of obtained sub-time series separately. It should be noted that there is hysteresis between the hydrographs of runoff and sediment. There would be many functions that can be related to the features of the main time series regarding to the relationship defines a wavelet function. According to previous studies, the db4 mother wavelet is more appropriate than other functions in order to simulate the hydrological process (Sharghi et al. 2018a, 2018b; Nourani et al. 2018a).

In the third step, the data are classified into homogeneous clusters to optimize the structure of the model (Step 3). M5 model tree is based on the tree classification method to establish a relationship between independent and dependent variables. This model can be used for both quantitative and qualitative data types, unlike the other decision tree algorithms which are used only for qualitative data. The patterns involved in the dataset are finally extracted at the fourth stage of the proposed methodology (Step 4).

### 2.2.1 Wavelet Transform

The time-scale wavelet transforms of a continuous time signal,  $x(t)$ , is defined as (Addison et al. 2001):

$$T(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} g^* \left( \frac{t-b}{a} \right) x(t) . dt \quad (1)$$

where  $a$  and  $b$  define the dilation factor and temporal translation of the function  $g(t)$ , which permits for the study of the signal around  $b$ , \*corresponds to the complex conjugate and  $g(t)$  is known as the wavelet function or mother wavelet.

The important feature of the wavelet transform, obtained from the basic function, is providing a time-scale localization of the process. This issue would be in contrast with the classical trigonometric functions of Fourier analysis. The wavelet transform seeks the connections between the signal and wavelet function. This assessment is determined at different scales of  $a$  and locally around the time of  $b$ . The result shows a wavelet coefficient ( $T(a, b)$ ) contour map known as a scalogram.

In practice, dealing with a discrete time signal to produce  $N^2$  coefficients from a dataset of length  $N$ , based on the trapezoidal rule, using a logarithmically uniform spacing discretization of  $a$  with a correspondingly coarser resolution of the  $b$  locations, the discrete mother wavelet transform will have the form of (Addison et al. 2001):

$$g_{m,n}(t) = \frac{1}{\sqrt{a_0^m}} g \left( \frac{t-nb_0a_0^m}{a_0^m} \right) \quad (2)$$

Where  $a_0$ ,  $b_0$ ,  $m$  and  $n$  show the specified fine dilation, location parameter, integers that control the wavelet dilation and translation, respectively. It is noticeable where  $a_0 > 1$ ,  $a_0$  is considered as 2 and where  $b_0 > 0$ ,  $b_0$  usually sets as 1.

### 2.2.2 M5 Model Tree

M5 model tree, introduced by Quinlan (1992), is a subset of tree-based models which tries to predict the amounts of numerical variables using the regression models. It benefits a multi-linear regression instead of assigning a constant value to the terminal leaf, so it is analogous to the piecewise linear functions. M5 model tree can train multi-dimensional tasks efficiently. This ability leads to the popularity of M5 model tree and caused the more usages at different fields of engineering. Also, the superiority of M5 over the other linear models is that the model trees are simple and more accurate than the regression trees (Nourani et al. 2018a, 2018b).

M5 tree model divides the input dataset into the collection of set  $T$  and the set  $T$  is split into the several subsets by the leaves through evaluating the split criterion. This method usually manufactures overelaborate structures, known as over-fitting, which must be pruned.

The information gain ratio is calculated by reduction of the standard deviation criterion at both previous and after testing in M5 model tree. At first, the standard deviation of the obtained amounts of classes in  $T$  is measured. But if  $T$  includes very few samples or their amounts do not change considerably,  $T$  is divided the test results.  $T_i$  shows the subset of samples corresponding to the  $i^{\text{th}}$  outcome of a particular test. If the standard deviation ( $sd(T_i)$ ) of the ultimate amounts of samples in  $T_i$  is assumed as a measure of error, the expected reduction in error can be calculated by (Quinlan 1992):

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \quad (3)$$

M5 model tree selects one that maximizes the expected error reduction. The WEKA software (Waikato Environment for Knowledge Analysis) was used in this study to investigate the relationships and present the tree model.

### 2.3 WANN Model

The basic of WANN model is quite similar to the ANN model. Both the ANN and WANN models include of three layers, training of error back propagation algorithm. The input layer of WANN model is included wavelet-based water discharge and SSL sub-series. It should be noticed that wavelet transform deals with different time scales. The approximation sub-series ( $Q_a(t)$  or  $S_a(t)$ ), that known as large-scale sub-signal, and  $d_{i\text{th}}$  or  $d_{j\text{th}}$  detailed sub-series ( $Q_{d\text{ith}}(t)$  or  $S_{d\text{jth}}(t)$ ) which state short-scale sub-signals are the components of the wavelet transform which follow the superposition principle (the combination of sub-series sets the main time series). It should be mentioned that there are different types of mother wavelet which are used in accordance with the type of the process. In this study, “db4” mother wavelet was used which is more appropriate for the hydrological processes (Nourani et al. 2014). After decomposition of time series by wavelet transform, the obtained sub-series are fed into an ANN model (Feed Forward Neural Network) model.



## 2.4 Efficiency Criteria

The efficiency of the models was evaluated by the Nash-Sutcliffe efficiency (NSE), root mean square error (RMSE) and mean absolute percentage error (MAPE) as (Azad et al., 2017):

$$NSE = 1 - \frac{\sum_{i=1}^N (S_i - \hat{S}_i)^2}{\sum_{i=1}^N (S_i - \bar{S}_i)^2} \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (S_i - \hat{S}_i)^2}{N}} \quad (5)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|S_i - \hat{S}_i|}{S_i} \quad (6)$$

Where  $DC$ ,  $RMSE$ ,  $MAPE$   $N$ ,  $S_i$ ,  $\hat{S}_i$ ,  $\bar{S}_i$  are Nash-Sutcliffe efficiency, root mean square error, mean absolute percentage error, number of observations, observed SSL data, calculated SSL values and mean of observed SSL data, respectively.

## 3 Results and Discussion

The influence of various factors such as rainfall, runoff, watershed and river conditions are involved in the SSL simulation and therefore selection of the most effective factors as input variables is an important step for sediment modeling. Previous studies have shown that sediment phenomenon is known as a Markovian or autoregressive process (its present value has the highest correlation with its previous values). Therefore, it can be stated that the effect of mentioned factors can be considered by the antecedent sediment values indirectly (Partal and Cigizoglu 2008; Senthil Kumar et al. 2011). Thus, the current SSL value would be a function of antecedent runoff values up to time step  $m$  and sediment up to time step  $n$  values as:

$$Q_t = f(Q_{t-1}, Q_{t-2}, \dots, Q_{t-m}, S_{t-1}, S_{t-2}, \dots, S_{t-n}) \quad (7)$$

The iteration epoch number for the model training is highly related to the behavior of the river response. According to the explanations given in the case studies, the data patterns of the case studies are quite different. The recognition of this fundamental difference can be helpful in justification of the obtained results.

Prior to the Wavelet-M5 modeling, the SSL was simulated by the classic ANN, seasonal-based WANN and sole M5 models. All models were trained by the calibration data and validated using the verification dataset.

The type of neural network used in the ANN and WANN models was a three-layer feed-forward perceptron (with an input, a hidden layer and an output layers) trained by the back propagation (BP) algorithm. The optimal structure (the number of neurons in input and hidden

layers) of ANN was obtained by trial and error as well as the number of optimal training epoch. It is noticeable that the low number of training iterations can lead into incomplete training and on the other hand, the large number of epoch may cause over-fitting issue. The Levenberg-Marquardt scheme of BP algorithm was used to train ANN due to its higher convergence rate (Sharghi et al. 2018a, 2018b). The network training process was stopped when the error rate was increased in the verification data. The sigmoid Tangent activation function was used as the non-linear kernel of neural networks in this research. The results of ANN modeling are presented in Table 2 for both case studies at daily and monthly scales. It should be mentioned that only the results of the best structures (optimal structure) have been presented in the tables.

The Markov characteristic of the sediment phenomenon, described as Eq. (6), was considered by the classic ANN, while the seasonality was ignored. To handle the seasonal features of the process, the WANN model was also applied for the modeling. For WANN modeling, the SSL and water discharge time series were decomposed at level 8 into 9 sub-time series (one approximation and 8 detailed sub-series) and level 4 into 5 sub-time series (one approximation and 4 detailed sub-series) at the daily and monthly time scales, respectively, by the wavelet transform in order to consider the seasonal pattern of the process at different scales. Due to the relative relationship between runoff and sediment, it was assumed that both time series included the same frequencies, so both time series were decomposed in the same level. The 'db4' wavelet was used due to previous studies and according its form which is similar to the SSL signal (Nourani et al., 2014). The obtained sub-series (only the dominant sub-series) then were considered as the candidate inputs of the ANN model. The results of WANN model are presented in Table 2 for both case studies at daily and monthly scales.

Similarly, M5 algorithm divided the SSL and water discharge data to some classes and then provided a linear regression for each class (multi-linear model) instead of a complex non-linear regression for all input samples. The results of M5 model tree are also presented in Table 2 for both case studies at daily and monthly time scales.

Finally, in the case of Wavelet-M5 modeling, the non-stationary SSL and water discharge time series were decomposed to several short and long term temporal sub-signals by wavelet transform to handle the involved trend in the main time series. Subsequently, all sub-time series were applied as inputs to M5 model tree. As it was argued in methodology, M5 model tree classifies the data samples by attributing a splitting criterion (standard deviation reduction) at the root node and branch, at first. In the following, the samples set is divided into subsets again and the splitting criterion is computed recursively for each branch. The generation of the branch is stopped when all samples at a node have the same clustering attribute at any time. Finally, a linear regression model is fitted on each subset of data samples. The results of Wavelet-M5 model are reported at Table 2 for the case studies in the proposed scales.

The rainfall pattern and behavior of the catchments are the effective factors to investigate the sediment phenomenon and would influence the SSL values. Previous studies have shown that the watershed behavior would approach to more linear response since it experienced more regular pattern. Overall, having a relative knowledge about the behavior of studied catchments would be helpful to apply the suitable model. As it was expected, ANN model could not lead to an acceptable performance in one day ahead forecasting at both case studies, according to the values of NSE and MAPE presented in Table 2. At the monthly scale, this less-desirable performance became weaker (in some cases, the NSE value was less than 50% at the training or verification step). Perhaps the most important reason that can be justified the disability of ANN model regards to the non-stationary nature of the input time series of sediment

**Table 2** Results of ANN, WANN, M5 and Wavelet-M5 models for both Lighvanchai and Upper Rio Grande Rivers at daily and monthly scales

Scale	Basin	Inputs	Output	Model	Hn*	Epoch	NSE		RMSE*		MAPE	
							Train	Verify	Train	Verify	Train	Verify
Daily	Lighvanchai	Q(t-1), S(t-2), S(t-1)	S(t)	ANN	4	30	0.93	0.80	0.02	0.07	1.54	1.19
		Q <sub>d</sub> (t-1), I <sub>dd</sub> (t-1), S <sub>d</sub> (t-1), S <sub>dd</sub> (t-1), S <sub>ds</sub> (t-1)		WANN	6	90	0.98	0.91	0.01	0.05	0.53	0.44
		Q(t-1), S(t-2), S(t-1)		M5	-	-	0.92	0.86	0.01	0.07	1.43	1.10
	Upper Rio Grande	Q <sub>d</sub> (t-1), Q <sub>dd</sub> (t-1), S <sub>d</sub> (t-1), S <sub>dd</sub> (t-1), S <sub>ds</sub> (t-1)		WT-M5	-	-	0.97	0.95	0.006	0.03	0.40	0.32
		Q(t-1), S(t-2), S(t-1)		ANN	3	50	0.62	0.53	0.04	0.06	5.89	6.69
		Q <sub>d</sub> (t-1), Q <sub>dd</sub> (t-1), S <sub>d</sub> (t-1), S <sub>ds</sub> (t-1), S <sub>dd</sub> (t-1)		WANN	5	60	0.89	0.84	0.01	0.02	2.51	2.82
Monthly	Lighvanchai	Q(t-1), S(t-2), S(t-1)	M5	-	-	0.77	0.75	0.01	0.03	3.75	4.24	
		Q <sub>d</sub> (t-1), Q <sub>dd</sub> (t-1), S <sub>d</sub> (t-1), S <sub>ds</sub> (t-1), S <sub>dd</sub> (t-1)	WT-M5	-	-	0.94	0.89	0.004	0.008	1.81	2.26	
		Q(t-1), S(t-1)	ANN	5	40	0.61	0.48	0.07	0.14	1.23	1.05	
	Upper Rio Grande	Q <sub>d</sub> (t-1), Q <sub>ds</sub> (t-1), S <sub>d</sub> (t-1), S <sub>ds</sub> (t-1), S <sub>dd</sub> (t-1)	WANN	9	70	0.97	0.78	0.01	0.09	0.56	0.64	
		Q(t-1), S(t-1)	M5	-	-	0.78	0.69	0.02	0.07	0.83	0.70	
		Q <sub>d</sub> (t-1), Q <sub>ds</sub> (t-1), S <sub>d</sub> (t-1), S <sub>ds</sub> (t-1), S <sub>dd</sub> (t-1)	WT-M5	-	-	0.94	0.90	0.007	0.02	0.34	0.31	
Upper Rio Grande	Q(t-1), S(t-1)	ANN	4	50	0.44	0.30	0.09	0.14	3.15	4.61		
	Q <sub>d</sub> (t-1), S <sub>at</sub> (t-1), S <sub>d</sub> 2(t-1), S <sub>d</sub> 3(t-1), S <sub>d</sub> 4(t-1)	WANN	14	90	0.93	0.77	0.03	0.08	1.28	1.55		
	Q(t-1), S(t-1)	M5	-	-	0.68	0.70	0.04	0.06	2.17	3.15		
		Q <sub>d</sub> 3(t-1), S <sub>at</sub> (t-1), S <sub>d</sub> 2(t-1), S <sub>d</sub> 3(t-1), S <sub>d</sub> 4(t-1)	WT-M5	-	-	0.90	0.86	0.01	0.03	1.71	2.13	

\* Hnr: Hidden neuron, \* RMSE is for normalized data, \* WT-M5: Wavelet-M5

phenomenon, which may be handled by employing the wavelet transform as an appropriate data pre-processing tool.

As it mentioned the previous studies have shown that M5 model tree is comparable to neural network (Solomatine and Xue 2004; Bhattacharya and Solomatine 2006). This study, which has been conducted for two different cases, has indicated that if the studied case shows a regular behavior, two mentioned models (i.e. ANN and M5 model tree) will be comparable while the watershed (river) experiences a non-linear behavior, these two models will not be comparable and the non-linear model will be more efficient. Also, the models were compared to each other as it was cited.

Another point that should be noted is the different performance of each model in daily and monthly scales. Clearly, ANN dealt with more samples of the input data in the daily scale rather than the monthly scale. This would make the network better in training and increase the accuracy of the model with regard to the monthly scale ( $(NSE_{Train\ or\ Verify})_{Daily} > (NSE_{Train\ or\ Verify})_{Monthly}$ , according to Table 2). Therefore, another defect that can be attributed to ANN model is its dependency to the number of input samples. Another considerable issue is the difference between NSEs in the training and verification steps, which is more highlighted at the monthly scale where the number of calibration samples is more than the verification data.

The different nature of daily and monthly time series should also be maneuvered in visitation of the sediment modeling. The monthly time series not only contain fewer samples than the daily time series, their seasonal behavior is much remarkable than the Markovian characteristic. Thus, WANN could handle both autoregressive (Markovian) and seasonal characteristics of the process. Therefore according to Table 2, WANN showed an acceptable performance for both daily and monthly modeling.

Despite the WANN efficiency with regard to the classic ANN, the volume of calculation is increased dramatically in this model due to the increment of the input data. Another issue noticed, is the NSEs difference in the training and verification steps. Although the accuracy of WANN model is increased somewhat in the verification step, its accuracy is still far from the training stage (because the non-linear models, particularly the WANN model, are strongly dependent to the number of input data and the number of samples is usually fewer in the verification step than the training step).

After modeling the process by the models with non-linear kernel (i.e., ANN and WANN), the multi-linear M5 model tree was also applied to the data. It should be noted that the classic linear models such as ARIMA fit just one linear relation to the whole time series of a nonlinear stochastic process. But M5 model tree divides the non-linear space of the input dataset into several classes (clusters) so that each class can then be represented by a simple linear regression. In fact, such multi-linear models can represent the non-linear behavior of a process by splitting the non-linear space into several linear sub-spaces while gaining the simplicity, superposition principle and non-magnification of the computation error.

Besides, the accuracy of the model is the most important criterion of evaluating the model performance. It turns out that the performance of M5 model tree was better than ANN model according to Table 2 ( $(NSE_{Train\ or\ Verify})_{M5} < (NSE_{Train\ or\ Verify})_{ANN}$ ). Unwittingly, it was questioned whether the non-stationary of the input series would influence the performance of M5 model tree as well as ANN and whether wavelet-based data pre-processing could be helpful in this regard?

Therefore similar to WANN model, the wavelet data pre-processing was applied to the input time series and then, M5 model tree was applied to the sub-series obtained through the

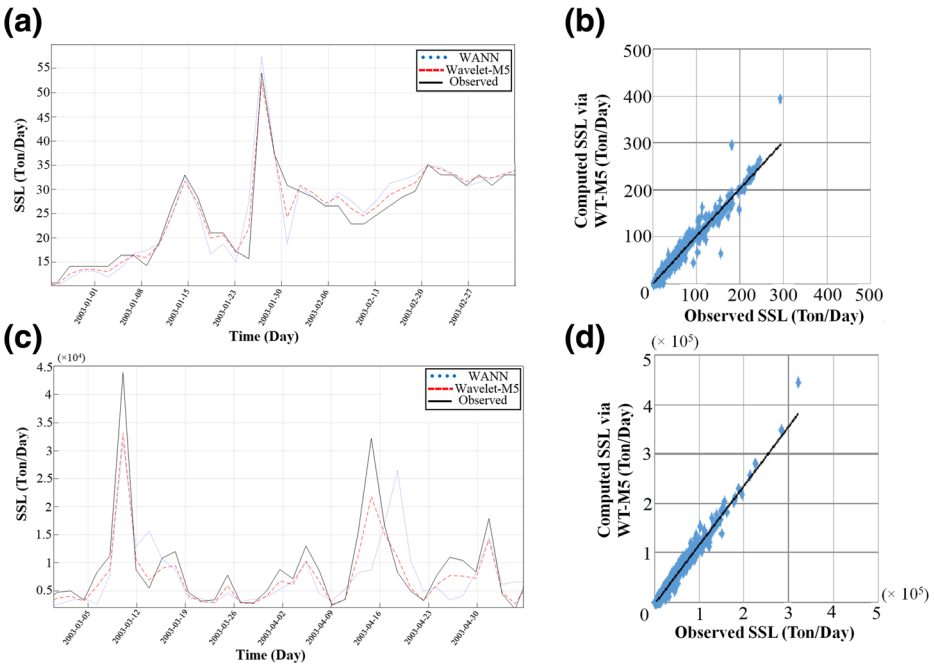


Fig. 3 Observed versus computed SSL (a detail) and scatter plot at daily time scale for a,b) Lighvanchai River and c,d) Upper Rio Grande River

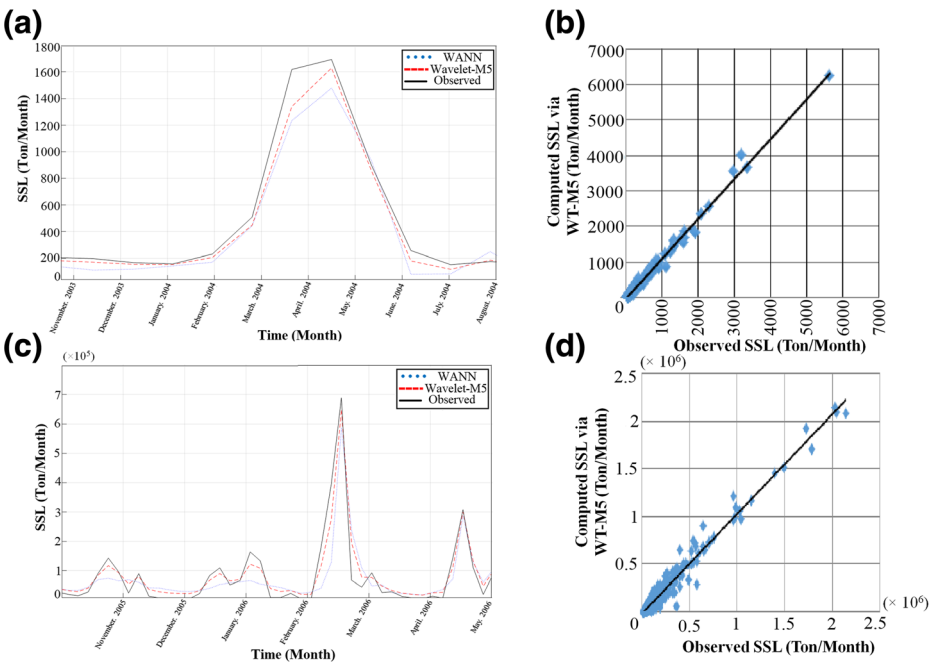


Fig. 4 Observed versus computed SSL (a detail) and scatter plot at monthly time scale for a,b) Lighvanchai River and c,d) Upper Rio Grande River

**Table 3** The performance evaluation of the hybrid models compared to ANN and M5 model tree

Models	River	Daily		Monthly	
		Training	Verification	Training	Verification
WANN vs. ANN	Lighvanchai	5%	13%	59%	62%
	Upper Rio Grande	44%	58%	1.11 times	1.56 times
Wavelet-M5 vs. ANN	Lighvanchai	4%	19%	54%	87%
	Upper Rio Grande	52%	68%	1.05 times	1.87 times
Wavelet-M5 vs. M5	Lighvanchai	5%	10%	21%	30%
	Upper Rio Grande	22%	19%	32%	23%

wavelet decomposition. The obtained results via Wavelet-M5 are presented in Table 2 for daily and monthly SSL and water discharge time series of the Lighvanchai and Upper Rio Grande Rivers. The results show that Wavelet-M5 could lead to the better performance than the sole M5 (see Figs. 3 and 4). For example, Wavelet-M5 shows about 52% and 22% improvements with regard to ANN and M5 models in the training step for Upper Rio Grande River at the daily scale (Table 3). As it can be seen in Table 3, the application of wavelet-based data pre-processing by Wavelet-M5 model could drastically improve the forecasting efficiency, so that the performance of the multi-linear Wavelet-M5 model is quite similar and even better than the non-linear WANN model (See Table 2).

The computed time series via WANN and Wavelet-M5 models versus the observed time series for Lighvanchai and Upper Rio Grande rivers at the daily and monthly scales are respectively presented at Figs. 3 and 4.

As Figs. 3 and 4 show, Wavelet-M5, due to benefiting wavelet analysis, can handle both autoregressive and seasonal characteristics of the SSL modeling. The proximity of the NSE in the training and verification steps is another point of view. Wavelet-M5 is not dependent on the number of data and is suitable for the processes that a lot of historical data are not available. Since M5 model tree is the basis of Wavelet-M5 model, all the positive features that were discussed about M5 model tree, such as benefiting an understandable insight from its structure, prevention of the error magnification, the applicability of the superposition principle and similar performance in the training and verification steps are still true. These features could help the model to use large number of input parameters without any change in the model accuracy, unlike ANN/ WANN models.

## 4 Conclusions

The prediction of suspended sediment load carried by a river is an important task in all studies on environmental modeling, river and dam engineering, as well as water resources management. The methods for the estimation of sediment load based on the properties of flow and sediment have limitations attributed to the simplification of important parameters and boundary conditions. Under such circumstances, soft computing approaches have proven to be efficient tools in modelling the sediment load. In this paper, the hybrid Wavelet-M5 model was utilized to simulate the sediment phenomenon of two different watersheds at both daily and monthly scales. To assess the efficiency of the proposed hybrid method (Wavelet-M5), it was compared with the WANN, ANN and M5 models.

It is noticeable that the effect of data pre-processing on the performance of both non-linear and multi-linear models to handle non-stationary nature of the SSL time series is quite clear. The application of wavelet transform could improve the performance of ANN model up to 13% at the daily and 62% at the monthly time scales for Lighvanchai (in verification phase). Relative variations in the performance of WANN model were 58% at the daily and 1.56 times at the monthly scale for Upper Rio Grande (see Table 3). The effect of the wavelet transform on the performance of M5 tree model was 10% at the daily scale and 30% at the monthly scale for Lighvanchai (in verification phase). Also, relative changes in the performance of Wavelet-M5 model were 19% and 23% at the daily and monthly time scales for Upper Rio Grande, respectively. In general, it could be concluded that the performance of the proposed hybrid multi-linear Wavelet-M5 model is desirable and better than WANN model. The results also showed that the hybrid multi-linear Wavelet-M5 model can be efficiently calibrated by the training data while faced with diverse catchments behaviors.

In addition to the sediment phenomenon which is an important input in the water resources planning and management, the ability of the proposed Wavelet-M5 model may be examined for the simulation of the other hydro-environmental processes. It is also suggested that the capability of the proposed hybrid Wavelet-M5 model is further compared with some conceptual and other black box models such as adaptive neuro-fuzzy inference system (ANFIS) and Wavelet-ANFIS.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

- Addison PS, Murrariy KB, Watson JN (2001) Wavelet transform analysis of open channel wake flows. *J Eng Mech* 127(1):58–70
- ALP, M., Cigizoglu & H.K (2007) Suspended sediment load simulation by two artificial neural network methods using hydrometeorological data. *Environ Model Softw* 22:2–13
- Azad A, Karami H, Farzin S, Saeedian A, Kashi H, Sayyahi F (2018) Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (case study: Gorganrood River). *KSCE J Civ Eng* 22(7): 2206–2213
- Bhattacharya B, Solomatine DP (2006) Machine learning in sedimentation modelling. *Neural Netw* 19(2):208–214
- Belayneh A, Adamowski J, Khalil B, Ozga-Zielinski B (2014) Long-term SPI drought forecasting in the Awash River basin in Ethiopia using wavelet neural networks and wavelet support vector regression models. *J Hydrol* 508:418–429
- Goyal MK (2014) Modeling of sediment yield prediction using M5 model tree algorithm and wavelet regression. *Water Resour Manag* 28(7):1991–2003
- Kuo CC, Gan TY, Yu PS (2010) Wavelet analysis on the variability, teleconnectivity and predictability of the seasonal rainfall of Taiwan. *Mon Weather Rev* 138(1):162–175
- Mustafa MR, Rezaur RB, Saiedi S, Isa MH (2012) River suspended sediment prediction using various multilayer perceptron neural network training algorithms—a case study in Malaysia. *Water Resour Manag* 26:1879–1897
- Nourani V, Davanlou Tajbakhsh A, Molajou A (2018a) Data mining based on wavelet and decision tree for rainfall-runoff simulation. *Hydrol Res*. <https://doi.org/10.2166/nh.2018.049>
- Nourani V, Razzaghzadeh Z, Hosseini Baghanam A, Molajou A (2018b) ANN-based statistical downscaling of climatic parameters using decision tree predictor screening method. *Theor Appl Climatol*. <https://doi.org/10.1007/s00704-018-2686-z>
- Partal T, Cigizoglu HK (2008) Estimation and forecasting of daily suspended sediment data using wavelet–neural networks. *J Hydrol* 358:317–331
- Quinlan JR (1992) Learning with continuous classes. *Proceedings of Australian joint conference on Artif Intell*, 343–348

- Salas JD, Delleur JW, Yevjevich VM, Lane WL (1980) Applied modeling of hydrologic time series, Water Resource Publications. Water Resources Publication, Littleton
- Salimi A, Karami H, Farzin S, Hassanvand M, Azad A, Kisi O (2018) Design of water supply system from rivers using artificial intelligence to model water hammer. *ISH J Hydraul Eng.* <https://doi.org/10.1080/09715010.2018.1465366>
- Senthil Kumar A, Ojha C, Goyal MK, Singh R, Swamee P (2011) Modeling of suspended sediment concentration at Kasol in India using ANN, fuzzy logic, and decision tree algorithms. *J Hydrol Eng* 17(3):394–404
- Sharghi E, Nourani V, Molajou A, Najafi H (2018a) Conjunction of emotional ann (eann) and wavelet transform for rainfall-runoff modeling. *J Hydroinf.* <https://doi.org/10.2166/hydro.2018.054>
- Sharghi E, Nourani V, Najafi H, Molajou A (2018b) Emotional ANN (EANN) and Wavelet-ANN (WANN) approaches for markovian and seasonal based modeling of rainfall-runoff process. *Water Resour Manag* 32(10):3441–3456
- Shiri J, Kisi O (2010) Short-term and long-term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model. *J Hydrol* 394:486–493
- Singer MB, Dunne T (2001) Identifying eroding and depositional reaches of valley by analysis of suspended sediment transport in the Sacramento River, California. *Water Resour Res* 37:3371–3381
- Sivakumar B, Wallender WW (2005) Predictability of river flow and suspended sediment transport in the Mississippi River basin: a non-linear deterministic approach. *Earth Surf Process Landf* 30:665–677. <https://doi.org/10.1002/esp.1167>
- Solomatine DP, Xue Y (2004) M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. *J Hydrol Eng* 9(6):491–501. [https://doi.org/10.1061/\(ASCE\)1084-0699\(2004\)9:6\(491\)](https://doi.org/10.1061/(ASCE)1084-0699(2004)9:6(491))
- Yang CT, Marsooli R, Aalami MT (2009) Evaluation of total load sediment transport formulas using ANN. *Int J Sedim Res* 24(3):274–286

## Affiliations

Vahid Nourani<sup>1,2</sup> · Amir Molajou<sup>3</sup> · Ali Davanlou Tajbakhsh<sup>1</sup> · Hessam Najafi<sup>1</sup>

<sup>1</sup> Department of Water Resources Engineering, Faculty of Civil Engineering, University of Tabriz, Tabriz, Iran

<sup>2</sup> Faculty of Civil & Environmental Engineering, Near East University, Lefkosa, North Cyprus, via Mersin 10, Turkey

<sup>3</sup> Department of Water Resources Engineering, Faculty of Civil Engineering, Iran University of Science and Technology, Tehran, Iran