Water Science & Technology

IWA PUBLISHING



Water Science & Technology Vol 87 No 7, 1791 doi: 10.2166/wst.2023.089

# The improvement of wavelet-based multilinear regression for suspended sediment load modeling by considering the physiographic characteristics of the watershed

Niloofar Nejatian[a,*], Mohsen Yavary Nia[b], Hooshyar Yousefyani[c], Fatemeh Shacheri[d] and Melika Yavari Nia[e]

[a] Department of Civil Engineering of City College, City University of New York, New York, USA
[b] Department of Civil and Coastal Engineering, University of Florida, Gainesvile, Florida, USA
[c] Department of Engineering and Technology, Central Michigan University, Mount Pleasant, Michigan, USA
[d] Department of Biological Systems Engineering, Virginia Tech University, Virginia, Blacksburg, USA
[e] Department of Civil and Environmental Engineering, Politecnico Di Milano, Milan, Italy
*Corresponding author. E-mail: nnejati000@citymail.cuny.edu

## ABSTRACT

The aim of this study is to model a relationship between the amount of the suspended sediment load by considering the physiographic characteristics of the Lake Urmia watershed. For this purpose, the information from different stations was used to develop the sediment estimation models. Ten physiographic characteristics were used as input parameters in the simulation process. The M5 model tree was used to select the most important features. The results showed that the four factors of annual discharge, average annual rainfall, form factor and the average elevation of the watershed were the most important parameters, and the multilinear regression models were created based on these factors. Furthermore, it was concluded that the annual discharge was the most influential parameter. Then, the stations were divided into two homogeneous classes based on the selected features. To improve the efficiency of the M5 model, the non-stationary rainfall and runoff signals were decomposed into sub-signals by the wavelet transform (WT). By this technique, the available trends of the main raw signals were eliminated. Finally, the models were developed by multilinear regressions. The model using all four factors had the best performance (DC = 0.93, RMSE = 0.03, ME = 0.05 and RE = 0.15).

**Key words:** feature selection, M5 model tree, physiographic characteristics, suspended sediment load, wavelet transform

## HIGHLIGHTS

- This study links the physiographic characteristics of the watershed to M5 sediment estimation.
- M5 model tree selects the most important features of the watershed.
- Wavelet transform decomposes the raw main signals into several sub-signals and improve the model performance.

## 1. INTRODUCTION

Soil erosion is a process in which soil particles are separated from their substrate and transported to another place by a transfer factor (Verheijen *et al.* 2009). Materials transported by erosive factors such as water, wind and ice, which settle in layers on the surface of the earth's crust, are referred to as sediment (Toy *et al.* 2002; Yang *et al.* 2022). The total sediments that are suspended, sliding or rolling by the streamflow, is called the sediment load. Basically, the sediment load of the entire watershed is transferred as three general forms: wash load, suspended load and bed load (Turowski *et al.* 2010; Yang *et al.* 2022). Suspended sediment load can have negative and undesirable effects on catchment areas and watersheds such as erosion increment, water quality reduction, infrastructure damage, flooding, water storage capacity reduction, etc. (Bhattacharya & Dutta 2013).

Physiographic characteristics of a watershed are referred to the set of features whose values are relatively constant for each watershed over time and indicate the appearance and morphology of the watershed. Benefiting the physiographic characteristics and the climatic conditions of the studied area can provide a relatively accurate picture of the quantitative and qualitative performances of the hydrological system of the watershed (Azizi & Nejatian 2022). The most important physiographic characteristics of a watershed are included as area, perimeter, length, main waterway, slope, form, elevation, topography and time of concentration (Ziegler *et al.* 2014; Eslami *et al.* 2022).

Estimating the suspended sediment load amount is considered very significant in a watershed study. However, direct measurement of the sediment is a very time-consuming and expensive action. So, it is easier to find the relationship between the physiographic and environmental characteristics of the watershed and obtain the suspended sediment load amount. Nowadays, the black-box models, which obtain comparable results without any cost or direct measurements, have become a popular tool (Nourani *et al.* 2019a). Multiple regression methods, cluster analysis and factor analysis are among the common methods in modeling the relationship between sediment rate and watershed characteristics. The multilinear regression model is a common method whose purpose is to express the dependent variable in the form of a mathematical function of the independent variable(s) (Nourani *et al.* 2019c).

In multiple regression, as the number of variables increases, the model becomes increasingly complex and the potential for errors to arise also increases. Therefore, implementing a method for selecting the most significant features would be highly beneficial in improving the efficiency and accuracy of the model. (Ares *et al.* 2016). Feature selection is one of the common methods, which tries to reduce the number of model input variables to a small number of important variables. In this method, a large number of variables can be reduced into a few factors and in this way, a summary of the main data can be prepared (Grabczewski & Jankowski 2005).

Among the various algorithms of a decision tree, the M5 model tree is a subset that can detect useful information from a dataset and select the important features and parameters. Through the M5 algorithm of a decision tree, the features with high scores, which were located in the upper nodes of tree, were selected in all the grid points (Quinlan 1992; Nourani *et al.* 2019d). The M5 model tree assigns a multivariate linear regression model instead of fitting a constant value to the leaf node, so it is analogous to piecewise linear functions (Khosravi *et al.* 2022). The benefits of the M5 model tree can be listed as (Quinlan 1992; Nourani *et al.* 2019c; Sayed *et al.* 2023):

- acceptable efficiency in dealing with large multi-dimensional problems and missing data;
- requires no trial and error;
- being more understandable and much simpler in the training phase than non-linear methods.

In recent years, experimental research works have been carried out in the field of studying and modeling the relationship between the watershed characteristics and the amount of suspended sediment load in the catchments. Kumar & Das (2000) utilized the multivariate regression model to estimate the daily sediment of the Ramganga River in India. It was observed that only four parameters, including the rainfall intensity at the event occurrence time and the 2 days before, the discharge 2 days before and the erosion of the previous day, were significant among all 17 variables that were introduced to the regression model step by step. Sarangi & Bhattacharya (2005) used a series of regression relationships to estimate the sediment load of watersheds in Quebec, Canada. The results showed that benefitting the physical parameters of the watershed increased the accuracy of the model so that the coefficient of determination of the model increased dramatically. Zhu *et al.* (2007) modeled suspended sediment using artificial neural networks and multivariate regression methods in China. The rainfall, temperature, rainfall intensity and discharge characteristics were used to estimate the amount of sediment in their study. The results indicate that the artificial neural network method has relatively better efficiency than the multiple regression method. Ares *et al.* (2016) experimented and analyzed the sediment concentration control factors for the Pampas region of Argentina. In this study, several rainfall events were simulated by multiple regression method and the obtained results showed that the developed linear model is able to explain 85% of sediment concentration changes. Lamb & Toniolo (2016) quantified the suspended load of three rivers in the northern region of Alaska. The study area was monitored for 3 years and suspended load sampling was done at different depths of the river, and between the amount of suspended load and the parameters of the basin, modeling was done by regression method. The results showed that in all three rivers, rainfall parameters and the shape of the basin had a great effect on the amount of suspended load in the basin.

Owing to the multi-resolution nature of original raw suspended sediment load signals, the efficiency of models to simulate the highly non-stationary, autoregressive and seasonal suspended sediment load signals declined meaningfully. Under these conditions, benefitting an appropriate data preprocessing method, like wavelet transform (WT), maybe a suitable solution to prevail these issues. The significant temporal information and hidden frequencies of the main raw suspended sediment load signals may be extracted by WT. Hence, numerous studies have examined the capability of WT in decomposing seasonal raw suspended sediment load signals time series into sub-time series at numerous temporal scales (levels) to extract inherent properties (Shiri & Kisi 2010; Belayneh *et al.* 2014; Nourani *et al.* 2019a, 2019b).

According to the mentioned studies, the importance of estimating the amount of suspended sediment load using the physiographic characteristics of the watershed is undeniable (Yang *et al.* 2022). However, according to our knowledge in order to determine the significant and influencing parameters on the amount of sediment modeling in the studied area, a comprehensive study has not been implemented in this regard in Lake Urmia yet. In this study, it is tried to link the hydrological, environmental and physiographic characteristics of the watershed to the decomposed selected time series to model the suspended sediment load.

## 2. METHODOLOGY

### 2.1. Case study

The Lake Urmia watershed (Figure 1) is one of the smallest regional basins in Iran, which collects the water from different parts of East and West Azerbaijan provinces and flows into Lake Urmia. The Lake Urmia watershed is located in the northwest of Iran, which is geographically located between 35° 40′ to 38° and 29′ of north latitude and 44° and 13′ to 47° and 53′ of east longitude. The area of the Lake Urmia watershed is more than 52,000 km$^2$, equivalent to 21.3% of the total area of Iran. 9,000 km are involved by flat and plain areas, 35,200 km$^2$ are included in the mountainous areas, and 7,800 km$^2$ are made up of Lake Urmia and marginal marshes. In terms of the territory of this watershed, the Urmia Lake consists of the central, western and southwestern parts of the East Azerbaijan province (a relatively large part of the province is approximately 19,000 km$^2$), about half of the West Azerbaijan province (the southern half of the province is approximately 21,500 km$^2$), a part of the northern part of Kurdistan province (about 5,000 km$^2$) and a very limited part of Zanjan province. The main source of the water supply is precipitation caused by humid air currents that enter the region from the west and the

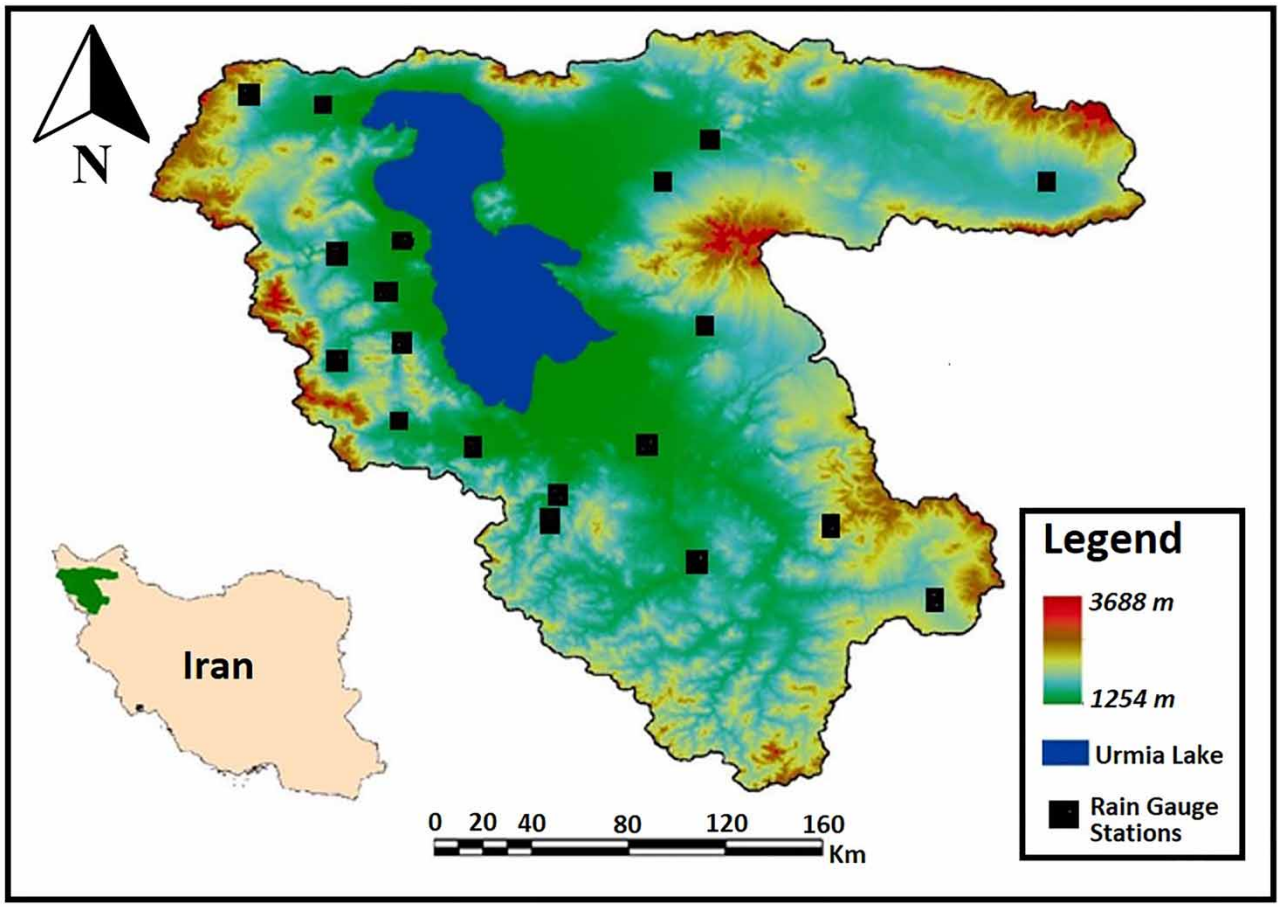


**Figure 1** | The topography of the Lake Urmia watershed.

Mediterranean. The rivers of the watershed originate from high mountains that are covered with snow most days of the year and have permanent springs and flowing water flows permanently and seasonally.

The climate of the Lake Urmia watershed is often influenced by its altitude. This catchment has a semi-arid continental climate and the Mediterranean rainfall regime is the dominant climatic regime of this basin. Its average annual rainfall is 398 mm. The rainiest season is winter and early spring, so that about 75% of the total rainfall occurs in the months of December–May. The regime of rivers is caused by precipitation and snow melting. The watershed temperature varies between −20 °C and 0 in winter and up to 40 °C in summer. The Urmia Lake watershed is divided into eight sub-basins. Zarineh Rood-Simineh Rood sub-basin is the largest one. Other sub-basins of are Aji Chai, Nazlochai, Mahabad Chai, Zulachai, Shabestar, Sufi Chai and Tasouj.

The important rivers of this watershed are the Talcheh River, Zarineh River, Simine River, Barandoz Chai, Rozeh Chai and Sufi Chai. The river bed is generally steep and consists of coarse-grained materials that are transported downstream by the flood stream. It should be mentioned that the Lake Urmia watershed has small but fertile plains such as the Sarab plain, Selmas plain, Sufian plain, Tabriz plain, Naqdeh plain, Miandoab plain, etc. Lake Urmia is the center of accumulation and discharge of surface water in this watershed, which was investigated at the local and national levels due to its importance.

It should be noticed that the Lake Urmia watershed is one of the most important catchments of Iran in terms of water, energy and agricultural products. Due to the tension increments between Iran and the United States of America (USA) regarding various challenges, including Iran's nuclear issues, some countries (especially the USA) have imposed severe economic sanctions against Iran (Koruzhde 2022; Koruzhde & Popova 2022). This matter has caused a large section of the Iranian people, especially the poor stratum of the society such as the villagers and farmers, to attempt to increase the amount of their cultivation and change their cultivation pattern toward more profitable and better-selling products to improve their livelihoods, which are highly dependent on the water. This factor has resulted in the indiscriminate exploitation of groundwater and surface water to increase personal profits and improve the livelihoods. The mentioned subject is one of the most important influencing factors that resulted in intensifying the drying process of Lake Urmia. This watershed collects runoff from vast areas of different provinces and after providing water to the plains joins Lake Urmia. The lack of sufficient number of sediment measurement stations and the limitation of the number of statistical years, as well as the low number of flood sampling during river flooding shows the importance of the present study in modeling suspended sediment estimation in the Lake Urmia watershed.

In order to estimate the amount of annual suspended sediment, eight sediment measuring stations located in the Urmia Lake watershed were selected. The stations were selected and had appropriate information such as rainfall, sediment and discharge statistics. By examining the statistical period of the data, a common statistical period of 20 years (time interval between 2001 and 2022) was selected. 75% of dataset is allocated for the calibration (training) step, while others are used for the validation (verifying) step. Then topographic, waterway network, land use information and climatic data were provided. The elevation, area and perimeter of the sub-basins were obtained from topographic maps with a scale of 1:25,000 in ArcMap software. In the following, Table 1 shows the information on synoptic, hydrometric, rain gauge and sediment measurement stations and their related sub-basins. It should be mentioned that the shape of the watershed was calculated

**Table 1** | The information of Lake Urmia watershed measurement stations and their related sub-basins

| Sub-bn | A[1] (km$^2$) | P[2] | AMP[3] | AMR[4] (m$^3$/s) | MSSL[5] (Ton/day) | ME[6] | MS[7] | $T_c$[8] | C[9] | FF[10] |
|---|---|---|---|---|---|---|---|---|---|---|
| Zarineh Rood-Simineh Rood | 11,840 | 435 | 370.60 | 54.77 | 701.69 | 1,500 | 3 | 63.2 | 1.14 | 1.81 |
| Aji Chai | 9,200 | 383 | 355 | 40.52 | 396 | 1,320 | 2 | 58.1 | 1.12 | 1.52 |
| Nazlochai | 2,030 | 180 | 340 | 6.52 | 300 | 3,000 | 2 | 14.6 | 1.13 | 0.66 |
| Mahabad Chai | 811 | 113 | 318 | 2.50 | 129.26 | 1,779 | 2.5 | 25.5 | 1.11 | 0.89 |
| Zulachai | 960 | 123 | 335 | 2.43 | 312 | 1,400 | 2.1 | 29.4 | 1.10 | 0.91 |
| Shabestar | 1,293 | 143 | 285.60 | 2.50 | 276 | 1,600 | 2.3 | 13.1 | 1.05 | 1.43 |
| Sofi Chai | 1,800 | 170 | 340.02 | 3.68 | 298 | 2,450 | 3.5 | 10.7 | 1.12 | 0.48 |
| Tasuj | 30 | 21 | 260 | 2.21 | 240 | 2,200 | 2.9 | 2.5 | 1.07 | 0.59 |

*Note:* In order to avoid disordering the table, the parameters were briefly mentioned in the table: (1) A: area; (2) P: perimeter; (3) AMP: annual mean precipitation; (4) AMR: annual mean runoff; (5) MSSL: mean suspended sediment load; (6) ME: mean elevation; (7) MS: mean slope; (8) $T_c$: time of concentration; (9) C: compactness; (10) FF: form factor.

by compactness ($C$ (Equation (1))) and form factor ($FF$ (Equation (2))) to imply the physiographic characteristics of the watershed in the estimation of the suspended sediment load (Thakkar & Dhiman 2007).

$$C = 0.28 \frac{P}{\sqrt{A}} \tag{1}$$

$$FF = \frac{A}{L^2} \tag{2}$$

In the above equations, $C$ and $FF$ are the compactness and form factor of the watershed (dimensionless), $A$ represents the area of the watershed ($km^2$), $P$ shows the perimeter of the catchment (km), $L$ is the length of the watershed (km). The average slope of waterways and catchments was extracted using a DEM map in an ArcMap environment (ArcMap is the main component of Esri's ArcGIS suite of geospatial processing programs, and is used for geospatial data) and then the longitudinal profile was drawn in Excel environment and the weighted slope of the main waterway was calculated. Among the climatic parameters, the average annual precipitation (rainfall) and the average rainfall of the rainy and flood months of the year, including December, January, February, March, April and May, were considered. First, the average monthly rainfall data and the elevation of each station were collected and based on the kriging method as the most appropriate geostatistical method, the spatial distribution of rainfall curves were extracted (Table 1). Then, the average annual and monthly rainfall of 20 years was extracted for each of the sub-basins.

### 2.2. Proposed methodology

The proposed methodology consists of four steps. At first, the physiographic characteristics such as area, perimeter and slope are collected (Step 1). In the second stage, the most important variables are selected by the feature selection property of the M5 model tree algorithm (Step 2). In the third stage, the WT decomposes the main signals into several sub-signals. Each of the obtained sub-signals depicts a specific feature. There are several functions that can decompose the main signals regarding to the relation specifies a wavelet function. Based on previous studies, it can be claimed that the db4 mother wavelet is more suitable than other wavelet functions to simulate the annual discharge and SSL (Nourani *et al.* 2019c). In the fourth stage, the selected variables are classified into homogeneous classes to optimize the structure of the model (Step 4). At last, the M5 model tree tries to fit a linear regression between independent and dependent variables (Step 5).

### 2.3. Multilinear regression model construction

In this study, the physiographic characteristics information of eight sub-basins was used to estimate the suspended sediment load. The most important variables which affected the amount of suspended sediment load were identified by the M5 model tree. Unlike the other black-box algorithms, the M5 model tree can diagnose and select the most important variables among a set of variables. Then, the Lake Urmia watershed was divided into homogeneous areas by model tree classification. Finally, based on the surveyed studies, the suspended sediment load was modeled using the multilinear regression for each homogeneous area. The multilinear regression model has been widely used due to its simplicity in the implementation and interpretation of the hydrological processes, especially in estimating the suspended sediment load amount, while benefiting the physiographic characteristics of the watershed (Gellis 2013; Ziegler *et al.* 2014; Nourani *et al.* 2019c). Also, the validation was carried out by the statistics of the two remaining sub-basins, and the efficiency of the model was evaluated.

In order to improve the accuracy of the multilinear regression model, the WT was employed to eliminate the available trend in the main raw time series (rainfall and runoff). Then, the M5 model tree classified the dataset samples and finally, a suitable regression model was presented for each class. WEKA software was used to check the relations and present a tree model. WEKA software provides the implementation of different learning algorithms, and with this software, you can easily apply different algorithms to the dataset. In line with the explanation of the tools used in the current study, a brief explanation of the WT as a preprocessing tool and the M5 model tree of the decision tree has been discussed.

### 2.4. Wavelet transform

WT is one of the most efficient and effective mathematical transforms in signal processing. Mathematical transformations are used to obtain additional information from the signal, which cannot be obtained from the raw main signal itself. Similar to the Fourier analysis, which is one of the most famous mathematical transformations, wavelet analysis deals with the expansion of functions, but this expansion is done in terms of wavelets. WT is an assumed specific function with zero mean and unlike the

trigonometric polynomials, it is checked locally in space and it is provided a closer relation between some functions and their coefficients and more numerical stability in calculations. Any application based on the fast Fourier transform can be formulated using wavelets and obtain more local spatial (or temporal) information (Nourani *et al.* 2019a, 2019b; Lakshmi *et al.* 2022; Anupong *et al.* 2023).

The WT has two important characteristics: fluctuation and short term. In other words, $\psi(x)$ is a WT function if and only if its Fourier transform $\psi(\omega)$ satisfies the following condition (Nourani *et al.* 2019a, 2019b; Lakshmi *et al.* 2022):

$$\int_{-\infty}^{+\infty} \frac{|\psi(\omega)|}{|\omega|^2} d\omega < +\infty \tag{3}$$

This condition is known as the acceptance condition of the WT $\psi(x)$. The above relationship can be considered equivalent to the Equation (4) (Nourani *et al.* 2019a, 2019b; Lakshmi *et al.* 2022; Anupong *et al.* 2023):

$$\psi(0) = \int_{-\infty}^{+\infty} \psi(x) dx = 0 \tag{4}$$

This characteristic of the function with zero mean is not a limiting concern and many functions can be called WT functions based on it. $\psi(x)$ is considered as the mother WT function, which is used by the two mathematical operations of transfer and scaling to change the size and location during the analyzed signal, and finally, the WT coefficients at any point of the signal *(b)* and each value of the scale *(a)* can be calculated as (Nourani *et al.* 2019a, 2019b; Lakshmi *et al.* 2022; Anupong *et al.* 2023):

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}} \psi\left(\frac{x-b}{a}\right) \tag{5}$$

### 2.5. M5 model tree

The M5 model tree (Figure 2) is a model used to display the classifiers and regressions in data mining. As its name suggests, this tree consists of a number of nodes and branches. In the M5 model tree, the leaves represent the classes. Decisions are

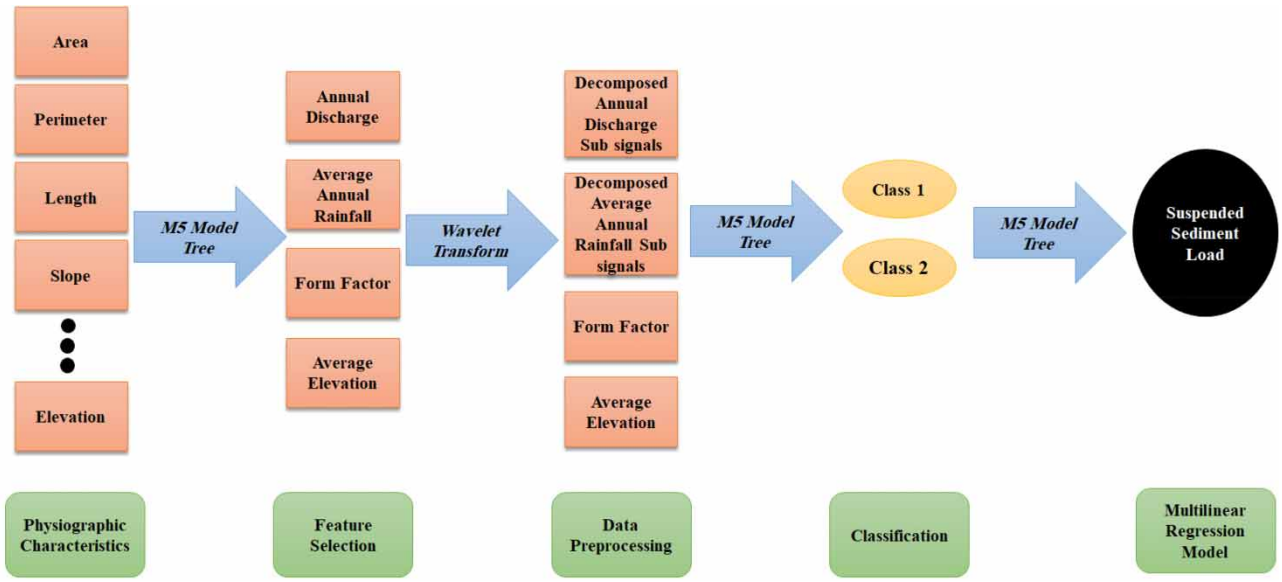


**Figure 2** | The schematic of the proposed methodology.

made in each of the other nodes (non-leaf nodes) according to one or more special attributes. The M5 model tree is a popular technique in data mining due to its simplicity and comprehensibility. In other words, the decision M5 model tree describes all the content and there is no need for an expert to interpret the output. In fact, this is a graphical method and therefore its interpretation is perhaps easier than other classification techniques. Obviously, having a large number of nodes can make the graphical representation difficult (Quinlan 1992; Nourani *et al.* 2019b).

The first step to create the M5 model tree is to use a branching criterion that is performed by one of the variables or features. The branching criterion is based on the function of the standard deviation of the values of each class or cluster, which is obtained in each node. This method is the basis of the classification method, called entropy. Entropy can be interpreted as a measure of the disorder of a system. The branching criterion indicates the amount of error in the node, and the model calculates the minimum expected error as the result of testing each attribute in that node. The model error is generally measured by the accuracy of the target prediction values of unseen cases. The standard deviation reduction (SDR) is calculated by (Quinlan 1992; Nourani *et al.* 2019d):

$$SDR = Sd(T_i) - \sum_{i=1}^{N} \frac{|T_i|}{|T|} Sd(T_i) \tag{6}$$

where $T$ is a set of input samples in each node. $T_i$ represents a subset of samples that have the *ith* potential test result. $Sd$ indicates the standard deviation. $i$ and $N$ show the data number (Figure 3).

### 2.6. Efficiency criteria

The accuracy of the models was assessed by the determination coefficient ($DC$), root mean square error ($RMSE$), mean error ($ME$) and relative error ($RE$) as (Xianzhao & Jiazhu 2008; Nourani *et al.* 2019c):

$$DC = 1 - \frac{\sum_{i=1}^{N} (SSL_i - \widehat{SSL_i})^2}{\sum_{i=1}^{N} (Q_i - \overline{Q_i})^2} \tag{7}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (SSL_i - \widehat{SSL_i})^2}{N}} \tag{8}$$

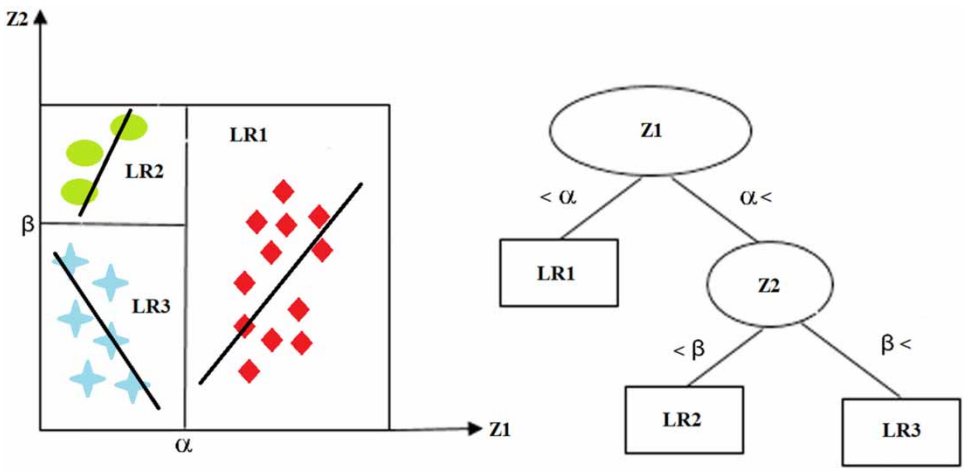


**Figure 3** | The schematic of the M5 model tree.

$$ME = \frac{\sum_{i=1}^{N} (SSL_i - \widehat{SSL_i})}{N} \quad (9)$$

$$RE = \frac{SSL_i - \widehat{SSL_i}}{SSL_i} \quad (10)$$

where $DC$, $RMSE$, $ME$, $RE$, $N$, $SSL_i$, $\widehat{SSL_i}$, $\overline{SSL_i}$ are determination coefficient, root mean square error, mean error, relative error, number of observations, observed suspended sediment load data, calculated suspended sediment load values and mean of observed suspended sediment load data, respectively.

## 3. RESULTS AND DISCUSSION

At first, 10 independent variables, including physiographic, climatic and hydrological characteristics, which are considered effective in suspended sediment load production, were identified and extracted.

After selecting the effective variables by the M5 model tree in order to reduce the calculation value and avoid the error growth, it can be seen that the four factors of annual runoff (discharge), average annual rainfall, FF and the average elevation of the watershed are able to explain the variances. Table 2 demonstrates the cumulative variance percentage and the specific values of the influential factors.

The average annual discharge has the highest weight on the first factor and explains more than half of the change in the main data. The second factor is the average annual rainfall, which has the highest weight. The third factor is the FF and finally, the fourth factor is the average elevation of the watershed. Totally, these four factors contain 90.28% of the variance or change in the original data and are selected for classification.

As mentioned earlier, based on the results of the previous studies, it can be claimed that the db4 wavelet function benefits the adequate essential features to decompose the runoff time series. Also, there are several jumps in the runoff time series due to the sudden start and cessation of rainfall over the catchment. Consequently, because of the formation of db4 wavelet that is similar to the runoff time series, it could capture the signal characteristic, especially peak points, efficiently and led to comparatively good results. Because of the proportional relation between the amount of runoff and SSL, these signals were supposed to have the same seasonality level and both time series were decomposed by the same wavelet function. In some previous studies, also db4 mother wavelet showed reliable outcomes to decompose runoff and SSL time series (Nourani *et al.* 2019c).

In general, relatively larger sub-basins were placed in homogeneous class 1. Homogeneous class 2 included smaller sub-basins. The multilinear regression models of the suspended sediment load estimation of the homogeneous classes 1 and 2 are presented in Table 3 for the annual scale by using four factors of the annual discharge, average elevation, FF and average rainfall of the watershed.

All the homogeneous class 1 models have a higher $R^2$ than the homogeneous class 2 models. Due to the more involved parameters in class 1, it is expected $R^2$ would be increased, although the amount of error may also increase to some extent. As it is shown in Table 3, the annual discharge variable is more significant in both classes and in class 1, where all the effective factors are involved, $R^2$ is higher and closer to 1.

The WT was taken to the input signals and then, the M5 model tree was employed to the sub-signals obtained through the WT decomposition. The performance of the wavelet-M5 model tree is presented in Table 4 for calibration and validation

**Table 2** | Cumulative variance percentage and specific values of different factors

| Component | Total | Variance (%) | Cumulative (%) |
|---|---|---|---|
| Annual discharge | 8.11 | 52.77 | 52.77 |
| Annual mean rainfall | 2.59 | 16.49 | 69.26 |
| Form factor | 1.80 | 12.01 | 81.27 |
| Mean elevation | 1.33 | 9.01 | 90.28 |

**Table 3** | The results of the multilinear regression model of the regional analysis of suspended sediment load in homogeneous classes 1 and 2

| Eq. No. | Independent variable | Eq. | $R^2$ |
|---|---|---|---|
| Homogeneous Class 1 | | | |
| (11) | $Q^1$, $P^2$, $F^3$ and $H^4$ | $SSL_t = 2569.3(Q) + 956.8(P) + 1031.2(F) - 387.6(H) - 37416.8$ | 0.93 |
| (12) | $Q$, $P$ and $F$ | $SSL_t = 32217.8(Q) + 1098.4(P) + 253596(F) - 28546.1$ | 0.90 |
| (13) | $Q$, $P$ and $H$ | $SSL_t = 15250.3(Q) + 955.4(P) + 346.3(H) + 17288.3$ | 0.89 |
| (14) | $P$, $F$ and $H$ | $SSL_t = 10584.4(P) + 18547.3(F) - 846.6(H) + 31298.5$ | 0.88 |
| Homogeneous Class 2 | | | |
| (15) | $Q$ and $P$ | $SSL_t = 34219.1(Q) + 054.4(P) + 39456.1$ | 0.61 |
| (16) | $Q$ | $SSL_t = 35214.7(Q) + 18547.5$ | 0.58 |

*Note:* (1) $Q$: annual discharge; (2) $P$: annual mean rainfall; (3) $F$: form factor; (4) $H$: mean elevation.

**Table 4** | The results of the wavelet-M5 model tree for Lake Urmia sub-basins in calibration and validation steps

| | DC | | RMSE | | ME | | RE | |
|---|---|---|---|---|---|---|---|---|
| Sub-basin | Calibration | Validation | Calibration | Validation | Calibration | Validation | Calibration | Validation |
| Zarineh Rood | 0.93 | 0.78 | 0.03 | 0.01 | 0.05 | 0.07 | 0.15 | 0.18 |
| Aji Chai | 0.91 | 0.79 | 0.02 | 0.01 | −0.12 | −0.13 | 0.12 | 0.14 |
| Nazlochai | 0.90 | 0.76 | 0.04 | 0.03 | 0.02 | −0.05 | 0.26 | 0.19 |
| Mahabad Chai | 0.88 | 0.71 | 0.05 | 0.04 | 0.07 | 0.11 | 0.32 | 0.29 |
| Zulachai | 0.87 | 0.70 | 0.02 | 0.03 | 0.10 | 0.09 | 0.11 | 0.19 |
| Shabestar | 0.88 | 0.69 | 0.06 | 0.07 | −0.11 | −0.15 | 0.23 | 0.31 |
| Sofi Chai | 0.85 | 0.68 | 0.04 | 0.05 | 0.04 | 0.05 | 0.16 | 0.24 |
| Tasuj | 0.84 | 0.61 | 0.06 | 0.07 | −0.19 | 0.17 | 0.29 | 0.30 |

*Note:* RMSE is normalized.

steps. As it can be observed in Tables 4 and 5, the utilization of WT can improve the efficiency of the multilinear regression model significantly due to its ability in overcoming the non-stationary precipitation and streamflow signals.

The proximity of the DC in the verification and training phases is another point of view. Wavelet-M5 is not dependent on the number of data and is suitable for the processes in that a lot of historical data are not available (Table 4).

**Table 5** | The results of the M5 model tree for Lake Urmia sub-basins in calibration and validation steps

| | DC | | RMSE | | ME | | RE | |
|---|---|---|---|---|---|---|---|---|
| Sub-basin | Calibration | Validation | Calibration | Validation | Calibration | Validation | Calibration | Validation |
| Zarineh Rood | 0.72 | 0.63 | 0.05 | 0.03 | 0.07 | 0.17 | 0.25 | 0.28 |
| Aji Chai | 0.60 | 0.49 | 0.04 | 0.04 | −0.15 | −0.16 | 0.32 | 0.24 |
| Nazlochai | 0.59 | 0.42 | 0.06 | 0.06 | 0.09 | −0.09 | 0.16 | 0.19 |
| Mahabad Chai | 0.67 | 0.49 | 0.07 | 0.05 | 0.17 | 0.11 | 0.22 | 0.19 |
| Zulachai | 0.76 | 0.61 | 0.04 | 0.06 | 0.10 | 0.13 | 0.31 | 0.29 |
| Shabestar | 0.67 | 0.57 | 0.08 | 0.09 | −0.12 | −0.16 | 0.33 | 0.21 |
| Sofi Chai | 0.64 | 0.55 | 0.06 | 0.07 | 0.14 | 0.06 | 0.46 | 0.34 |
| Tasuj | 0.73 | 0.60 | 0.09 | 0.09 | −0.12 | 0.13 | 0.19 | 0.35 |

*Note:* RMSE is normalized.

The computed wavelet-M5 model tree versus the observed suspended sediment load signal and the scatter plot for Zarineh Rood-Simineh Rood, which contain all the effective variables and have the highest $R^2$, presented at Figures 4 and 5, respectively.

Figure 4 shows that the wavelet-M5 model tree overcomes the non-stationary features of the suspended sediment load signals, because of benefits of the WT as a preprocessing tool. Also, the WT can handle the signal features, especially the peak values, and acquire comparatively high efficiency according to its structure.

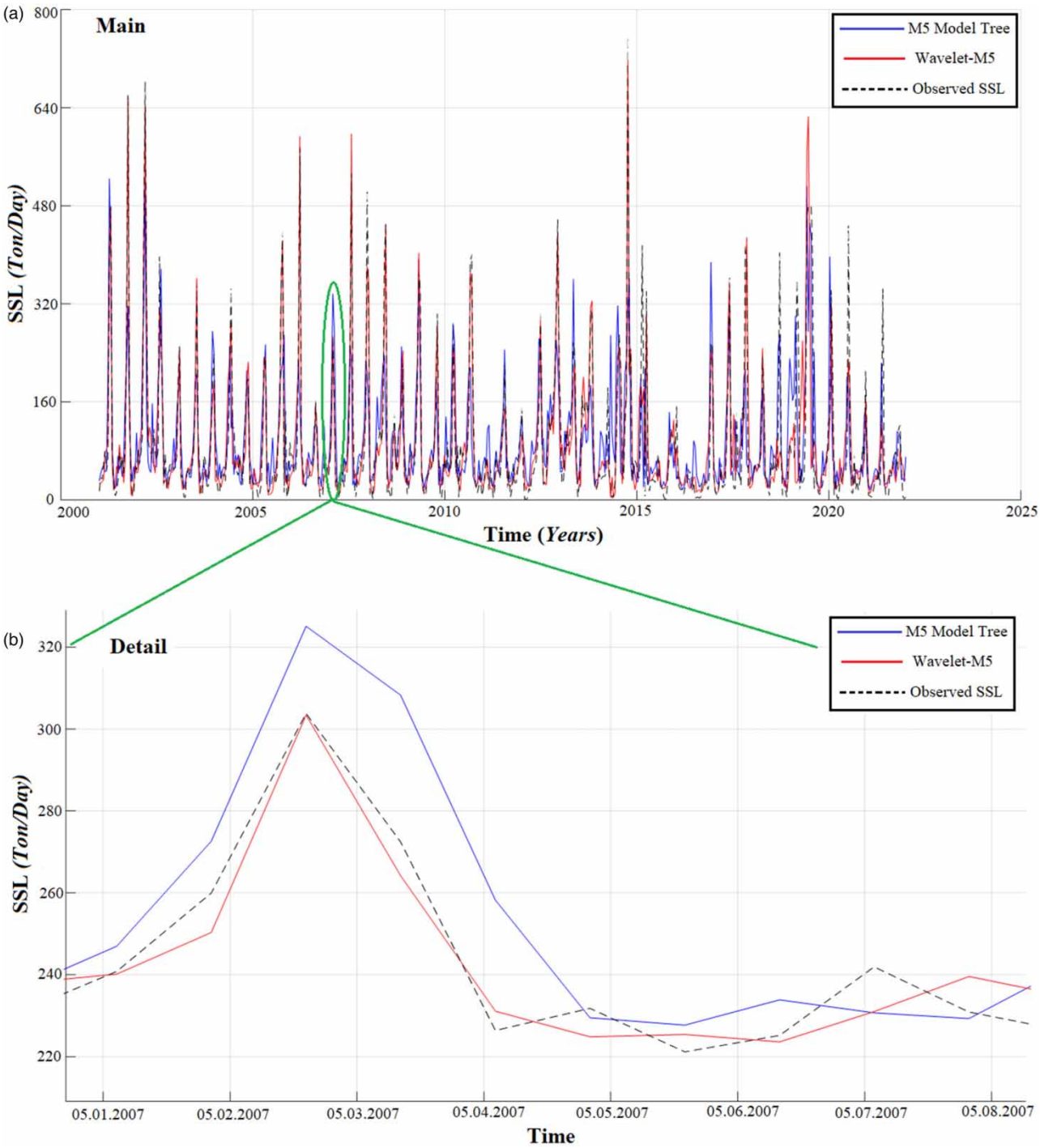


**Figure 4** | The wavelet-M5 model tree vs. the observed SSL for Zarineh Rood-Simineh Rood.

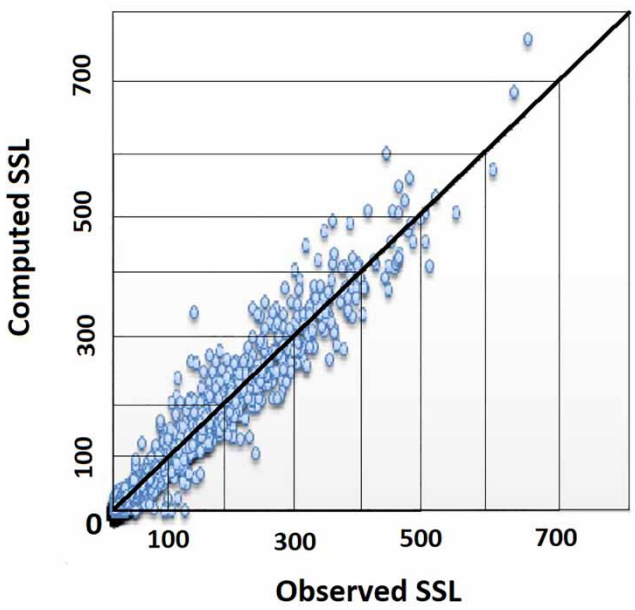


**Figure 5** | The scatter plot of the wavelet-M5 model tree vs. the observed SSL for Zarineh Rood-Simineh Rood.

## 4. CONCLUSION

Regional analysis of the rivers' suspended sediment load and its relation to the characteristics of the watersheds is considered significant in estimating the amount of erosion and sedimentation, especially in arid and semi-arid regions. It is possible to estimate the correct amount of the suspended sediment load by exploring and modeling the relationship between the physiographic and environmental characteristics of the watershed. The purpose of the current study was to model the relationship between the environmental characteristics of the Lake Urmia watershed and the amount of the suspended sediment load using a multilinear regression model. The obtained results indicated that the four factors of the annual discharge, average elevation, FF and average rainfall of the watershed were the most important factors in estimating the amount of the suspended sediment load based on the feature selection of the M5 model tree (see Table 2). The results also showed that the multilinear regression model obtained from all four factors has the highest $R^2$ (see Tables 3 and 4). Furthermore, benefiting the WT as a preprocessing tool resulted in acceptable criteria efficiency (see Table 4). It can be concluded that the combined use of feature selection and multilinear regression model has a suitable and acceptable performance in estimating the suspended sediment load. It is recommended to considered more characteristics of the watershed in the future studies. Also, it is suggested to compare the performance of the model with other black-box and physical-based models.

## DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

## CONFLICT OF INTEREST

The authors declare there is no conflict.

## REFERENCES

Anupong, W., Jweeg, M. J., Alani, S., Al-Kharsan, I. H., Alviz-Meza, A. & Cárdenas-Escrocia, Y. 2023 Comparison of wavelet artificial neural network, wavelet support vector machine, and adaptive neuro-fuzzy inference system methods in estimating total solar radiation in Iraq. *Energies* **16** (2), 985.

Ares, M. G., Varni, M. & Chagas, C. 2016 Suspended sediment concentration controlling factors: an analysis for the Argentine Pampas region. *Hydrological Sciences Journal* **61**, 2237–2248.

Azizi, H. & Nejatian, N. 2022 Evaluation of the climate change impact on the intensity and return period for drought indices of SPI and SPEI (study area: Varamin plain). *Water Supply* **22**, 4373–4386.

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. *Journal of Hydrology* **508**, 418–429.

Bhattacharya, A. K. & Dutta, R. K. 2013 Sediment yield estimation and its impact on reservoir sedimentation: a review. *Journal of Hydro-Environment Research* **7**, 171–184. doi:10.1016/j.jher.2012.11.005.

Eslami, H., Yousefyani, H., Yavary Nia, M. & Radice, A. 2022 On how defining and measuring a channel bed elevation impacts key quantities in sediment overloading with supercritical flow. *Acta Geophysica* **70**, 2511–2528.

Gellis, A. C. 2013 Factors influencing storm-generated suspended-sediment concentrations and loads in four basins of contrasting land use, humid-tropical Puerto Rico. *Catena* **104**, 39–57.

Grabczewski, K. & Jankowski, N. 2005 Feature selection with decision tree criterion. In: *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*. Rio de Janeiro, Brazil, 2005, IEEE, p. 6. doi: 10.1109/ICHIS.2005.43.

Khosravi, M., Afshar, A. & Molajou, A. 2022 Decision tree-based conditional operation rules for optimal conjunctive use of surface and groundwater. *Water Resources Management* **36**, 2013–2025.

Koruzhde, M. 2022 The Iranian crisis of the 1970s-1980s and the formation of the transnational investment bloc. *Class, Race and Corporate Power* **10** (2), 1–17.

Koruzhde, M. & Popova, V. 2022 Americans still held hostage: a generational analysis of American public opinion about the Iran nuclear deal. *Political Science Quarterly* **137**, 511–537.

Kumar, A. & Das, G. 2000 Dynamic model of daily rainfall, runoff and sediment yield for a Himalayan watershed. *Journal of Agricultural Engineering Research* **75**, 189–193.

Lakshmi, P. J., Apaza, R. A., Alkhayyat, A., Marhoon, H. A. & Alameri, A. A. 2022 Hybrid wavelet-gene expression programming and wavelet-support vector machine models for rainfall-runoff modeling. *Water Science & Technology* **86**, 3205–3222.

Lamb, E. & Toniolo, H. 2016 Initial quantification of suspended sediment loads for three Alaska north slope rivers. *Water* **8**, 419.

Nourani, V., Davanlou Tajbakhsh, A., Molajou, A. & Gokcekus, H. 2019a Hybrid wavelet-M5 model tree for rainfall-runoff modeling. *Journal of Hydrologic Engineering* **24**, 04019012.

Nourani, V., Tajbakhsh, A. D. & Molajou, A. 2019b Data mining based on wavelet and decision tree for rainfall-runoff simulation. *Hydrology Research* **50**, 75–84.

Nourani, V., Molajou, A., Tajbakhsh, A. D. & Najafi, H. 2019c A wavelet based data mining technique for suspended sediment load modeling. *Water Resources Management* **33**, 1769–1784.

Nourani, V., Razzaghzadeh, Z., Baghanam, A. H. & Molajou, A. 2019d ANN-based statistical downscaling of climatic parameters using decision tree predictor screening method. *Theoretical and Applied Climatology* **137**, 1729–1746.

Quinlan, J. R. 1992 Learning with continuous classes. In: *Paper Presented at the 5th Australian Joint Conference on Artificial Intelligence*. Vol. 92, pp. 343–348.

Sarangi, A. & Bhattacharya, A. K. 2005 Comparison of artificial neural network and regression models for sediment loss prediction from Banha watershed in India. *Agricultural Water Management* **78**, 195–208.

Sayed, B. T., Al-Mohair, H. K., Alkhayyat, A., Ramírez-Coronel, A. A. & Elsahabi, M. 2023 Comparing machine-learning-based black box techniques and white box models to predict rainfall-runoff in a northern area of Iraq, the Little Khabur River. *Water Science and Technology* **87** (3), 812–822.

Shiri, J. & Kisi, O. 2010 Short-term and long-term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model. *Journal of Hydrology* **394**, 486–493.

Thakkar, A. K. & Dhiman, S. D. 2007 Morphometric analysis and prioritization of miniwatersheds in Mohr watershed, Gujarat using remote sensing and GIS techniques. *Journal of the Indian Society of Remote Sensing* **35**, 313–321.

Toy, T. J., Foster, G. R. & Renard, K. G. 2002 *Soil Erosion: Processes, Prediction, Measurement, and Control*. John Wiley & Sons, New York, United States, 352 Pages. ISBN: 978-0-471-38369-7.

Turowski, J. M., Rickenmann, D. & Dadson, S. J. 2010 The partitioning of the total sediment load of a river into suspended load and bedload: a review of empirical data. *Sedimentology* **57**, 1126–1146.

Verheijen, F. G., Jones, R. J., Rickson, R. J. & Smith, C. J. 2009 Tolerable versus actual soil erosion rates in Europe. *Earth-Science Reviews* **94**, 23–38.

Xianzhao, L. & Jiazhu, L. 2008 Application of SCS model in estimation of runoff from small watershed in Loess Plateau of China. *Chinese Geographical Science* **18** (3), 235–241.

Yang, Y., Huang, B. & Zhu, D. Z. 2022 Experimental study of sediment washout from stormwater sumps. *Water Science & Technology* **86**, 2454–2464.

Zhu, Y. M., Lu, X. X. & Zhou, Y. 2007 Suspended sediment flux modeling with artificial neural network: an example of the Longchuanjiang River in the Upper Yangtze Catchment, China. *Geomorphology* **84**, 111–125.

Ziegler, A. D., Benner, S. G., Tantasirin, C., Wood, S. H., Sutherland, R. A., Sidle, R. C., Jachowski, N., Nullet, M. A., Xi, L. X., Snidvongs, A., Giambelluca, T. W. & Fox, J. M. 2014 Turbidity-based sediment monitoring in northern Thailand: hysteresis, variability, and uncertainty. *Journal of Hydrology* **519**, 2020–2039.

First received 6 February 2023; accepted in revised form 13 March 2023. Available online 24 March 2023