Improved explicit formulation of bedload transport using a novel multi-level multi-model data-driven ensemble approach

Hossien Riahi-Madvar (h.riahi@vru.ac.ir)
Vali-e-Asr University of Rafsanjan
Mahsa Gholami
Bu-Ali Sina University
Bahram Gharabaghi
University of Guelph

Research Article

Keywords: Multi-model ensembles approach, bedload transport, function finding, equation optimization, machine learning

Posted Date: October 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2120777/v1

Improved explicit formulation of bedload transport using a novel multi-level multi-model data-driven ensemble approach

Hossien Riahi-Madvar[*1], Mahsa Gholami[2], Bahram Gharabaghi[3]

[1] Department of Water Engineering, Faculty of Agriculture, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.
[2] Department of Civil Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamedan, Iran.
3 School of Engineering, University of Guelph, Guelph, Ontario, N1G 2W1, Canada

*Corresponding Author’s email address: h.riahi@vru.ac.ir

ABSTRACT

Estimation of bedload transport in rivers is a very complex and important river engineering challenge needs substantial additional efforts in pre-processing and ensemble modeling to derive the desired level of prediction accuracy. This paper aims to develop a new framework for the formulation of bedload transport in rivers using multi-level Multi-Model Ensemble (MME) approach to derive improved explicit formulations hybridized with multiple pre-processed-based models. Three pre-processing techniques of feature selection by Gamma Test (GT), dimension reduction by principal component analysis (PCA), and data clustering by subset selection of maximum dissimilarity (SSMD) are utilized at level 0. The multi-linear regression (MLR), MLR-PCA, artificial neural network (ANN), ANN-PCA, Gene expression programming (GEP), GEP-PCA, Group method of data handling (GMDH) and GMDH-PCA are used to develop individual explicit formulations at level 1, and the inferred formulas are hybridized with the MME approach at level 2 by Pareto optimality. A newly revised discrepancy ratio (RDR) for error distributions in conjunction with several statistical and graphical indicators were used to evaluate the strategy's performance. Results of MME showed that the proposed framework acted as an efficient tool in explicit equation induction for bedload transport (i.e., 33–96% reduction of RMSE; 2-29% increase of R^2, 2-138% increase of NSE and 38-98% reduction of RAE in testing step in comparison with the best individual model) and clearly outperformed estimations made by other models. The current study highlights the importance of pre-processing and multi-modelling techniques in deep learning models to encounter the challenges of function finding for complex bedload transport estimations in multiple observed datasets.

Keywords: Multi-model ensembles approach, bedload transport, function finding, equation optimization, machine learning

1- Introduction

Sediment transport in river flows leads to several challenges for the water resources tasks and is crucial in the context of reservoir sedimentation, flood control, river morphology changes, stable channel design, fish and wild life habitat, and watershed management (Van Rijn, 1993; Bhattacharya et al., 2007; Dey, 2014; Elkurdy et al., 2021; Ahmadianfar et al., 2021). In sediment transport, the coarse-grains conveyed by higher discharges and floods immediately above the bed are known as bedload (Barry, 2007).

Sediment transport is a highly complex, stochastic phenomenon with somewhat unknown theory. It is hard to measure in the field due to the time and cost-intensive process. These features of sediment transport produce high uncertainty in predictive equations that made their applicability questionable and makes limitation on employing them (Bhattacharya et al., 2007; Riahi-Madvar and Seifi, 2018).

The predictive methods of bedload transport are generally categorized into physical and data driven models (Kitsikoudis et al., 2014; Gholami et al., 2018 & 2019). Considering challenges of phenomenon complexity, inaccuracies in the predictive equations of bedload and measuring difficulties with the physical methods, development of new data-driven-based models with an appropriate determination of effective parameters of bedload having easily accessible field variables is vital (Ghani et al., 2011; Gao, 2011; Ebtehaj et al., 2021).

With the emerging applications of machine learning (ML) models, producing effective results in formulation of complex nonlinear challenges in river engineering, researchers have endeavoured to use these new techniques to cope with the complicated nature of bedload transport in parallel with the experimental and physical-based studies (Bhattacharya et al., 2007; Safari et al., 2020).

Various ML methods were implemented for sediment transport modelling such as artificial neural network: ANN (Afan et al., 2016; Bhattacharya et al., 2007; Kitsikoudis et al., 2014), fuzzy logic and adaptive neural fuzzy inference system: ANFIS (Kitsikoudis et al., 2014; Qasem et al., 2017), support vector machine: SVM

(Roushangar and Shahnazi, 2020; Sahraei et al., 2017), genetic expression programming: GEP (Danandeh Mehr et al., 2018; Ghani and Azamathulla, 2014).

Montes, et al., (2021), Noori et al. (2010a-c, 2011) and Liu et al (2020) figured out that these techniques suffer from the generalization capabilities of the results due to inappropriate selection of training set, inaccuracy issues with limited extrapolation abilities when applied to unseen data set extensive than the data used in training phase. They suggested pre-processing techniques such as data clustering for subset selection in train and testing steps. The studies in the literature of bedload prediction, have neglected mathematical-based clustering of train and test sets, while this study considered a subset selection of maximum dissimilarity (SSMD) to overcome these challenges.

The data-driven models developed so far for bedload transport estimations are basically black-box type tools such as ANN, ANFIS and suffer from limited interpretability of physical importance of the input parameters and their interactions to the model outputs, inability to capture physical processes (Noori et al., 2010a-c, 2011; Montes et al., 2021; Seifi and Soroush, 2020; Madvar et al., 2020). Therefore, the derivation of explicit accurate equations for bedload transport in rivers based on AI models, remains challenging.

To address this problem, in the current study a hybridization of four mathematical models including ANN, GEP, group method of data handling (GMDH), and multi linear regression (MLR) are employed in deriving the explicit predictive equations for bedload using a new multi-model-based strategy.

The data-driven models are susceptible to the number of the input variable. To the best of our knowledge, few studies there are relating to the use of approach to reduce the dimension of input data space and to astutely designate appropriate input variables for prediction of bedload in a multi-model ensemble approach.

Generally, the bedload rate is chosen as a dependent parameter. The fluid properties, flow conditions, sediment properties, and channel geometry are considered independent parameters in data-driven model developments (Montes et al., 2021; Qasem et al., 2017). In conventional ML-based models usually rely on the researchers' subjective "suggesting" the input variables that will result in a poor prediction (Liu et al., 2020).

Hence, proposing a sophisticated approach such as principal component analysis (PCA) (Snieder et al., 2020) and Gamma test (GT) to reduce the dimension of the input space leading to choose proper input parameters of the model, is valuable. The studies in the literature neglected input vector manipulation and data dimension reduction for ML prediction of bedload. In contrast, the present study used PCA and

GT techniques for dimension reduction and effective variable selection. In the current study pre-processing techniques of GT and PCA as dimension reductions are used in conjunction with ANN, GEP, GMDH and MLR.

This literature review confirms that, there are three main challenges and questionable problems in the ML techniques developments for bedload rate including:

1- the input feature selection (Dehghani et al., 2019) , input dimension reduction to infer most effective variables (Snieder et al., 2020),

2- optimized subset selection of train and test data sets to avoid overfitting (Riahi-Madvar et al., 2019 & 2021), and

3- multi-model procedure to overcome the weakness of single models using ensembles modeling strategy (Khatibi et al., 2020).

4- This study aims to address these challenges and efforts to improve the estimation of bedload transport rate through considering techniques implemented in a multi-model-based approach. As powerful ML models, MLR, ANN, GMDH and GEP in conjunction with SSMD, PCA and GT techniques are utilized for modeling.

The bedload prediction challenges are improved by a successive strategy including Multiple Models (MM) in three levels as follows:

(i) Level 0: use pre-processing techniques, SSMD, GT and PCA in data manipulation, dimension reduction and input feature selection,

(ii) Level 1: developing standalone ML models as base reuse and recursion techniques, that their results are reused as inputs in the next level inputs;

(iii) Level 2: reuse and recursion of base models in a Pareto multi-gene framework by reusing the results of the previous level to the inputs of the present level and the bedload rate as a target for improved accuracy.

The main contribution of the current paper is four-fold. First, implementing the SSMD, GT, and PCA-based approaches in input vector manipulation, dimension reduction, and pre-processing of an extensive bedload transport database. Second, the utilized data set includes a wide range of low shear to high shear sediment transport observations. When combined with the pre-processing techniques, will improve the generalization issues of previous studies by dimension reduction. Third utilizing individual MLR, MLR-PCA, ANN, ANN-PCA, GEP, GEP-PCA, GMDH, GMDH-PCA models to derive explicit predictive equations for bedload. Fourth, integrating the output of individual models with the POMGGP procedure as a new multi-model strategy that utilizes individual models' power of and eliminates their weakness in bedload predictions.

To the best of the author's knowledge, the presented multi-model ensembles approach driven by the different techniques is a unique one in the literature concerning bedload rate prediction. This paper is organized as follows. Section 2 presents the material and method, including data, dimension analysis, preprocessing techniques, stand-alone, and multi-model strategy. Section 3 discusses the results of the study in three pre-defined levels. Section 4 provides summaries and conclusions.

2- Material and methods

2-1- Experimental data and dimensional analysis

Literature review revealed that bedload material properties, cross-section geometry features, and flow conditions are the main properties that affect the sediment transport in streams (Safari et al., 2020; Ghani ,1993;) and bedload transport in rivers can be defined by the following set of effective parameters in the form of unknown f_1 function

$$q_b = f_1(U, H, W, R, D_s, S, g, \rho_s, \rho_w, \mu, u_*, u_{*c}) \qquad 1$$

Where q_b is bedload transport, U is flow velocity, H is flow depth, W is river width, R is hydraulic radius, D_s is sediment size, S is bed slope, g is gravity acceleration, ρ_s and ρ_w are sediments and water mass density respectively, μ is dynamic viscosity, u* is shear velocity, and u*c is critical shear velocity. The dimensionless form of bedload transport rate can be written in unknown *f*2 functional form as

$$\emptyset = f_2\left(S, D_{gr}, \frac{R}{D_s}, \frac{U}{u_{*c}}, \frac{H}{D_S}, \frac{H}{W}, F_r, F_{rg}, R_e, R_{e*}, \theta, \frac{U}{u_*}\right) \qquad 2$$

in which the dimensionless parameters are particle mobility parameter $\emptyset$, Slope S , dimensionless grain diameter D_{gr}, relative depth $\frac{R}{D_s}$, critical velocity ratio $\frac{U}{u_{*c}}$, depth ratio $\frac{H}{D_S}$, aspect ratio $\frac{H}{W}$, Froud number F_r, densimetric Froud number F_{rg}, Reynold number R_e, densimetric Reynold number R_{e*}, shields parameter θ, velocity ratio $\frac{U}{u_*}$, defined by

$$\emptyset = \frac{q_b}{D_s\sqrt{g(s-1)D_s}}, \; D_{gr} = D_s\left[\frac{(S-1)g}{\vartheta^2}\right]^{1/3}, \; R_{e*} = \frac{u_* D_S}{\vartheta}, \; R_e = \frac{UH}{\vartheta}, \qquad 3$$

$$F_r = \frac{U}{\sqrt{gH}}, \; F_{rg} = \frac{U}{\sqrt{gD_S(s-1)}}, \; \theta = \frac{\gamma HS}{gD_S(s-1)}$$

In order to develop the models, several datasets available in the literature were extracted, pre-processed, and utilized (Cao, 1997; Meyer-Peter and Müller, 1948; Recking et al., 2004). A total of 1280 data sets are used in this current study that are provided in the paper's supplementary material. The sediment diameter ranges from 0.274 mm to 44.3 mm. The bed sloped varies from 0.01 % to 20 %, flow depth from 0.00084 m to 1.0921 m, flow velocity from 0.193 m/s to 2.88 m/s, Froud number from 0.41 to 5.19 and bed material load from 0.01 g/m^3 to 1356 g/m^3.

2-2- Level 0: Pre-Processing techniques of bedload data

2-2-1-Feature selection using Gamma test

In the current study the GT is used to select the best input variables in ML-based bedload predictions. The GT stands on the hypothesis that when two points of x' and x are close together in input space, their corresponding bedload rate in output space of φ' and φ should be close, else their difference is due to noise. In each data set of {(x_i, $\emptyset_i$) € R^m, 1≤ i ≤M} by only supposition of the functional form of bedload transport $\emptyset = \emptyset(x_1, \dots x_m) + r$, where $\emptyset$ is a smooth function, and r is a random variable that shows noise with the bounded variance of noise Var(r). In mathematics a function could be considered "smooth" if it is differentiable everywhere (hence continuous) and in the Gamma test procedure the $\emptyset$ is smooth if it has constrained first partial derivatives. For a function to be smooth, it must have continuous derivatives up to a certain order, say k. We say that function is k^{th} order smooth. Now the domain of possible predictive model is constrained to the smooth functions $\emptyset$ that have constrained first partial derivatives. The Gamma indicator Γ is an estimation of that portion of the variance of the predictions that cannot be achieved by a smooth model (Remesan et al., 2009). By calculating the Euclidean distance

δ and γ of kth nearest neighbour x_N [i, k] from x_i($1\leq i\leq M$), ($1\leq k\leq p$) the Γ is computed from least-square fit between δ and γ as: $\gamma = A\delta + \Gamma$. The slope of regression A represents the complexity of bedload transport phenomenon under investigation. In the GT, if the Γ in comparison with the variance of $\emptyset$ as V_{ratio} were high, the probability of predicting $\emptyset$ using selected inputs is low, when the V_{ratio} is small or near zero, the probability of predicting $\emptyset$ by selected inputs is high. So, using the mask tests, the most effective parameters on $\emptyset$ can be determined. Also, the GT using M-test can determine the appropriate number of data records in modelling bedload transport (Dehghani et al. ,2019). In this study the WinGamma software is sued for GT, freely available at:

http://users.cs.cf.ac.uk/O.F.Rana/Antonia.J.Jones/GammaArchive/Gamma%20Software/Mathematica/GammaTestMathematicaFiles.htm

2-2-2-Data Clustering and Subset Selection by SSMD

According to Montes et al. (2021) and Safari (2020), the range of dissimilarity in the training dataset directly influences the model generality, overfitting problem, extrapolation ability and accuracy. The SSMD is used to avoid the overfitting of data-driven models. Suppose that X is the dataset as $X = \left(x_1, x_2, \dots, x_p\right)$ and a collection of $m = 1,2,..,N$ points defined as a selected subset for the training stage. If the squared distance between ith and jth point define as D_{ij}^2, and k points have already been selected ($k<p$), then the minimum distance from applicant point of *N* to *k* points define as (Memarzadeh et al., 2020)

$$D_{ij}^2 = \left\|x_i - x_j\right\|^2 = \sum_{k=1}^{p}\left(x_{ki} - x_{kj}\right)^2 \qquad 4$$

The ($k+1$)th candidate point in train group is chosen from remaining (*N-k*) points in the dataset that has the highest distance from an existing point. In this study, the SSMD code is developed in MATLAB environment.

2-2-3-Component selection and dimension reduction using PCA

In the PCA pre-processing technique, the original input variables are converted and reduced to fewer independent principal components (PCs) through an orthogonal projection into uncorrelated PCs (Lu et al., 2003). Using this technique, combinations of the *P* primary variable, $X_1, \dots, X_p$, are used to create *P* independent components, $PC_1, \dots, PC_p$ equal to the number of original variables.

2-3- Level 1: Standalone predictive models

2-3-1-Multiple linear regression (MLR)

If we have *n* observations of the *p*-dimensional independent variable *X* and want to establish a linear relationship with the response variable $\emptyset$, we can use the following MLR model (Zounemat-Kermani et al., 2020):

$$\emptyset = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \qquad 5$$

The parameters $\beta_j,\ j = 0, 1, \ldots, p$ are called regression coefficients. The least-squares method is commonly used to estimate the regression coefficients.

2-3-2-ANN-MLP

The Multi-Layer Perceptron (MLP) models are the most popular NN tools used in most of research and literature (Seifi and Soroush, 2020). By determining the weights and biases of NN architecture, and simplifying the MLP, the predictive equation of model can be derived. The ANN is developed using the MATLAB toolbox.

2-3-3-Pareto Optimal GEP and MGEP

The innovative technique of gene expression programming (GEP) utilized with Darwinian theory of evolution by natural selection to automatically solve optimization problems based on its two main components, the chromosomes and expression trees. A new sophisticated version of GEP is the Multigene-GEP(MGEP), that the initial population is created by GP trees with different genes (a number selected from1 and G_{max}). In the MGEP approach two conflict goals are considered. The first is the selection of the bedload predictive equation with lowest complexity and the second is the highest accuracy. These two conflict objects lead to a multi objective optimization problem. Here to solve the optimization problem with two conflict goals the Pareto optimality is combined with multi-genetic programming. In the multi-model-based framework the Pareto optimization is sued in order to balance between the complexity of model and the accuracy. Suppose that X_1 and X_2 are two feasible solutions. In the dominance relationship, two solutions must satisfy the constraints of (Zhang et al., 2017): $f_d(X_1) \leq f_d(X_2), \forall d \in \{1,2,\dots,D\}$ and $f_i(X_1) \leq f_i(X_2), \exists i \in \{1,2,\dots,D\}$, In which f_d is the fitness value of d solution, and D is the number of the optimization goals.

If the feasible solution X* satisfies the above conditions and there isn't any sequence solution X while X< X*, so that the solution X* will be preserved and is called the Pareto optimal solution. A collection of entire Pareto optimal solutions is entitled as the final Pareto optimal solutions set, and a set of values of the target function that are related to the disassembly sequence is called the Pareto optimal frontier. The complexity of each multi-gene is calculated simply by summation of individual gene complexity. In individual genes, the complexity is determined by counting the nodes, the subtrees, leaves. A tradeoff between model accuracy and complexity would result in Pareto optimal selection of the best equation. The POMGGP is used in the MATLAB software with GPTIPS toolbox.

2-3-4- Group Method of Data Handling (GMDH)

GMDH is one of the meta-heuristic data-driven models based on multivariate analysis for complex systems without the need to have a special basic knowledge. The GMDH develops an analytic function using a progressive network with binomial transfer functions (Shaghaghi et al., 2018). The mathematical form of GMDH that maps inputs (x_1, x_2, x_3, …, x_n) to the predicted output ($\widehat{\emptyset}$) is written as

$$\widehat{\emptyset} = a_o + \sum_{i=1}^{n} a_1 x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} x_j a_{ijk} x_i x_j x_k + \cdots \qquad 6$$

The least-squares error rule is utilized for coefficient determination of GMDH in MATLAB environment.

2-4- Level 2: Multi-model ensembles (MME) approach

In the present study, in addition to the individual predictive models, an innovative multi-model ensembles approach is presented. This feeds the output of standalone models into the POMGGP as a multi-model technique to improve the predictive capability of models.

This new contribution in bedload rate prediction as an ensembles approach consists of two primary levels: Level 1 in which the original input variables or pre-processed (PCs) are used to estimate bedload transport rate in standalone models of MLR, MLR-PCA, ANN, ANN-PCA, GMDH, GMDH-PCA, GEP, GEP-PCA; Level 2 in which the outputs of level 1 models are used as inputs to the POMGGP along with the original bedload rate as output.

In this framework, as presented in Fig.1 the POMGGP is used to run models at level 1 based on Pareto optimality analysis. Observed values of bedload rate serve as the target output in both levels. The strength of the developed framework is learning at two levels, automatic individual model selection by natural evolution in multi-gene GEP, balancing surrogate model complexity and accuracy via Pareto optimality.

The idea behind the multi-model ensembles approach has been inspired by the hierarchical recursion of models, that teamworking of models in parallel can help achieve a more accurate prediction (Khatibi et al., 2020).

The models in level 1 and 2 are comparatively evaluated using performance metrics coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), Nash Sutcliffe efficiency (NSE), and graphical analysis including scatter plots, importance probability, Pareto front and Taylor diagrams.

Furthermore, a newly revised discrepancy ratio (RDR) for error distributions developed by the authors (Riahi et al., 2020) is used to overcome non-normality, zero or negative value predictions with a rectified linear unit (ReLu) function (Ramachandran et al., 2018). The RDR is calculated by:

$$RDR = Sign(\emptyset_{p,i} - \emptyset_{o,i}) \left| log \left| \frac{\emptyset_{p,i}}{\emptyset_{o,i}} \right| \right| \quad 7$$

In which the $\emptyset_{o,i}$ is measured value $\emptyset_{p,i}$ is the estimated model output. In the case of over-predictions by POMGGP, the value of RDR>0 and in the case of under-prediction RDR<0 and for exact predictions RDR is equal to zero. The multi-model ensemble is developed using MATLAB environment.

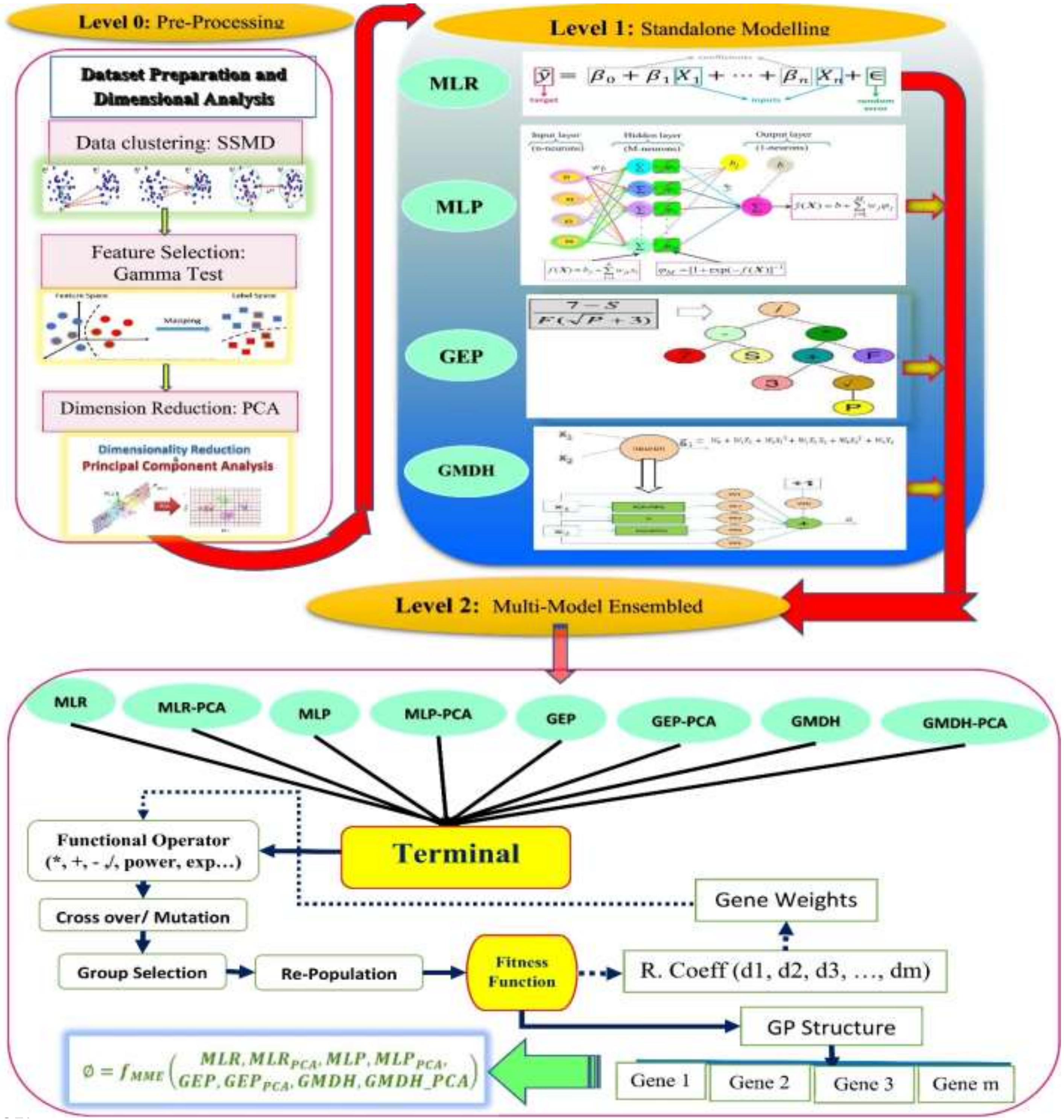

Level 0: Pre-Processing
Dataset Preparation and Dimensional Analysis
Data clustering: SSMD
Feature Selection: Gamma Test
Feature Space
Mapping
Label Space
Dimension Reduction: PCA
Dimensionality Reduction
Principal Component Analysis
Level 1: Standalone Modelling
MLR
$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon$
coefficients
target
inputs
random error
MLP
Input layer (n-neurons)
Hidden layer (M-neurons)
Output layer (1-neurons)
GEP
$\frac{7-S}{F(\sqrt{P}+3)}$
GMDH
Level 2: Multi-Model Ensembled
MLR
MLR-PCA
MLP
MLP-PCA
GEP
GEP-PCA
GMDH
GMDH-PCA
Terminal
Functional Operator (*, +, - ,/, power, exp…)
Cross over/ Mutation
Group Selection
Re-Population
Fitness Function
R. Coeff (d1, d2, d3, …, dm)
Gene Weights
GP Structure
Gene 1
Gene 2
Gene 3
Gene m
$\emptyset = f_{MME}\left(\begin{matrix} MLR, MLR_{PCA}, MLP, MLP_{PCA}, \\ GEP, GEP_{PCA}, GMDH, GMDH\_PCA \end{matrix}\right)$

Fig. 1. Flowchart of the developed multi-model ensembles approach for function finding in bedload prediction

3- Results and discussion

3-1- Level 0: Pre-processing

The results obtained using the pre-processing techniques are presented in Table 1. The train (80%) and test (20%) sets are selected using the SSMD approach. For a sizeable natural data bank like those used in this study, the SSMD expandes the envelope range of training sets, improves the applicability of developed predictive models and encompasses outlier data in the training set.

Table 1. Descriptive statistics of parameters in all, train and test subsets categorized by SSMD.

	Parameter	Mean	Mode	SD	Min	First quartile	Median	Third quartile	Max
All (1280 data points)	S	0.02	0.01	0.04	0.00	0.00	0.01	0.02	0.20
	Dgr	163.08	12.88	210.77	6.98	35.97	80.70	221.92	1150.14
	U/u_{*c}	11.73	11.21	3.38	3.26	9.86	12.12	14.37	18.99
	H/W	0.21	0.13	0.14	0.01	0.09	0.18	0.28	0.85
	Fr	1.19	1.13	0.60	0.41	0.79	1.09	1.35	5.19
	Re*	944.92	28.00	1834.21	21.00	55.00	199.50	1103.00	15086.00
	θ	0.20	0.05	0.30	0.01	0.05	0.07	0.29	3.70
	U/u_{*c}	11.69	11.87	3.43	3.26	9.86	12.13	14.37	18.98
	Φ	1.84	0.00	11.59	0.00	0.00	0.02	0.67	264.05
Train (1024 data points)	S	0.03	0.07	0.04	0.00	0.00	0.01	0.02	0.20
	Dgr	192.86	12.88	223.84	6.98	51.78	107.35	262.17	1150.14
	U/u_{*c}	11.08	11.21	3.36	3.26	9.00	11.25	13.42	18.99
	H/W	0.22	0.04	0.15	0.01	0.10	0.20	0.30	0.85
	Fr	1.21	1.13	0.64	0.41	0.83	1.08	1.33	5.19
	Re*	1151.67	30.00	1992.09	21.00	98.00	362.50	1265.00	15086.00
	θ	0.19	0.05	0.32	0.01	0.05	0.07	0.16	3.70
	U/u_{*c}	11.04	3.26	3.42	3.26	9.00	11.24	13.42	18.98
	Φ	1.95	0.00	12.93	0.00	0.00	0.01	0.14	264.05
Test (256 data points)	S	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.02
	Dgr	43.95	12.88	63.24	12.86	12.88	12.88	51.78	608.49
	U/u_{*c}	14.31	11.47	1.89	7.19	13.46	14.55	15.66	17.68
	H/W	0.15	0.07	0.11	0.01	0.08	0.12	0.18	0.58
	Fr	1.12	0.69	0.41	0.45	0.72	1.13	1.44	2.16
	Re*	117.91	28.00	309.44	25.00	29.00	35.00	81.00	3528.00
	θ	0.25	0.06	0.21	0.02	0.05	0.30	0.43	0.87
	U/u_{*c}	14.31	14.24	1.89	7.20	13.46	14.55	15.66	17.68
	Φ	1.42	0.00	1.79	0.00	0.00	0.94	2.32	9.07

The GT is used for feature selection and determining the proper input vector that characterizes the complex process in bedload transport. At first, the datasets are normalized [-1 1] and then GT is utilized via mask test procedure, and GT results for different input configurations are shown in Table 2. In Table 2, 12 dimensionless variables are used as the input variables with varying combinations to the GT.

In the first configuration, all 12 input parameters are used and GT indices calculated as given in the first row of Table 2. Then in the next GT run, the first input parameter is removed and masked and the GT results are recalculated, as given in the second row. Again, the removed variable is returned into the input vector and the second input variable is masked, and GT is performed in all the combinations. This method is continued for all selected variables in Table 2, one by one and in each step the Γ value is calculated.

The masking of the most influential variables in bedload prediction is associated with increases in the Γ value (V ratio) regarding the case that includes all variables (first row in Table 2). The highest Γ value indicates that the removed variable is essential and should be selected as the input variable of models.

Finally based on the results of GT in Table 2, the most important variables with the highest Γ value are S, D_{gr}, $U/u*_c$, H/W, F_r, Re_*, Q, U/u_* as shown in bold style. The input components reduced from 12 to 8 and the functional form simplified as

Table 2. The GT results on the selected 12 input masks for feature selection

	Removed	Gamma	Gradient	Standard Error	V-Ratio	Mask
0	None	0.04134632	0.24293793	0.02504585	0.165385281	111111111111
1	**S**	**0.043221076**	0.25201855	0.02499108	0.172884302	011111111111
2	**Dgr**	**0.041702514**	0.246745969	0.025434014	0.166810055	101111111111
3	R/D_s	0.041275668	0.249138526	0.025293184	0.165102672	110111111111
4	**U/u_{*c}**	**0.042657031**	0.251330176	0.025272607	0.170628125	111011111111
5	$H/\ D_s$	0.041493969	0.260073122	0.025051376	0.165975875	111101111111
6	**H/W**	**0.044526813**	0.273403925	0.025769505	0.178107253	111110111111
7	**Fr**	**0.043472253**	0.258844691	0.024823123	0.173889012	111111011111
8	Frg	0.040546081	0.271921358	0.025270117	0.162184324	111111101111
9	Re	0.040604003	0.258535669	0.024947276	0.162416013	111111110111
10	**Re_***	**0.041900604**	0.255306066	0.025551793	0.167602418	111111111011
11	**θ**	**0.042757646**	0.337648978	0.024902663	0.171030583	111111111101
12	**U/u_***	**0.042685003**	0.251320735	0.025282989	0.170740012	111111111110

$$\emptyset = f_3(S, D_{gr}, \frac{U}{u_{*c}}, \frac{H}{W}, F_r, R_{e*}, \theta, \frac{U}{u_*}) \qquad 8$$

The PCA is used as a dimension reduction technique over the GT results. According to KMO=0.624, the PCA is applicable for dimension reduction and the input variables are reduced into three principal components which are a linear combination of primitive dimensionless variables as

$$PC_1 = 0.069S_n + 0.310D_{gr,n} - 0.256\left(\frac{U}{u_{*c}}\right)_n + 0.012\left(\frac{H}{W}\right)_n - 0.042(F_r)_n$$

$$+ 0.293(R_{e*})_n - 0.151(\theta)_n - 0.253\left(\frac{U}{u_*}\right)_n$$

$$PC_2 = 0.314S_n - 0.143D_{gr,n} - 0.051\left(\frac{U}{u_{*c}}\right)_n + 0.064\left(\frac{H}{W}\right)_n + 0.39(F_r)_n -$$

$$0.081(R_{e*})_n + 0.382(\theta)_n - 0.059\left(\frac{U}{u_*}\right)_n \qquad 9$$

$$PC_3 = -0.113S_n + 0.289D_{gr,n} + 0.267\left(\frac{U}{u_{*c}}\right)_n + 0.683\left(\frac{H}{W}\right)_n + 0.037(F_r)_n$$

$$+ 0.289(R_{e*})_n + 0.266(\theta)_n + 0.264\left(\frac{VU}{u_*}\right)_n$$

Here, the n footnote indicates the normalized parameters in PCA. These three PCs explained the 85 % of total variances in the bedload transport datasets. The PCA results are given in Table 3, and the Kaiser criterion shows that three components have eigenvalues of more than 1 with a cumulative total variance of 85 %.

Therefore, the 8 bedload transport parameters can be reduced to the three uncorrelated PCs while preserving 85 % of the information of primary variables. As this table shows, the PC1 has an eigenvalue of 3.526 that explains 44.071 % of the total variance, PC2 has an eigenvalue of 2.049 with 25.614 % of total variance presented and PC with an eigenvalue of 1.16 has an eigenvalue of 14.505 %.

A scree graph of the amount of variance explained versus PCs and eigenvalues is shown in Fig.2, indicates that a break of the line occurred after PC3 and shows that only first three PCs maintain useful information. The selected PCs are rotated to determine their importance relative to each of 8 dimensionless parameters, as given in Table 4. A high value for each parameter's PC loading indicates a reasonable correlation between the parameter and corresponding PC.

Table 3. The PCA results on bedload transport data

Component	Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.526	44.071	44.071
2	2.049	25.614	69.685
3	1.160	14.505	84.190
4	0.800	10.006	94.196
5	0.297	3.712	97.908
6	0.129	1.610	99.518
7	0.037	0.462	99.980
8	0.002	0.020	100.000

Table 4. Rotated PC loading of bedload effective parameters

Parameter	Component		
	1	2	3
S	0.423	0.825	-0.159
D_{gr}	0.865	-0.171	0.345
$\frac{U}{u_{*c}}$	-0.834	-0.317	0.320
$\frac{H}{W}$	0.068	0.106	0.791
F_r	0.126	0.925	0.011
R_{e*}	0.856	-0.028	0.339
θ	-0.222	0.815	0.281
$\frac{U}{u_*}$	-0.830	-0.334	0.317

As these results show, the first component is explained by D_{gr} and R_{e*} and includes the highest level of information and describes the sediment material properties. The second PC is explained by S, ,F_r and θ that describes the flow properties, and the third PC is explained by $\frac{U}{u_{*c}}, \frac{H}{W},$ and $\frac{U}{u_*}$, which this PC describes the geometry and friction properties of the bedload transport. These three relevant PCs will be used as an input vector to the multi-models as follows

$$\emptyset = f_4(PC_1, PC_2, PC_3) \tag{10}$$

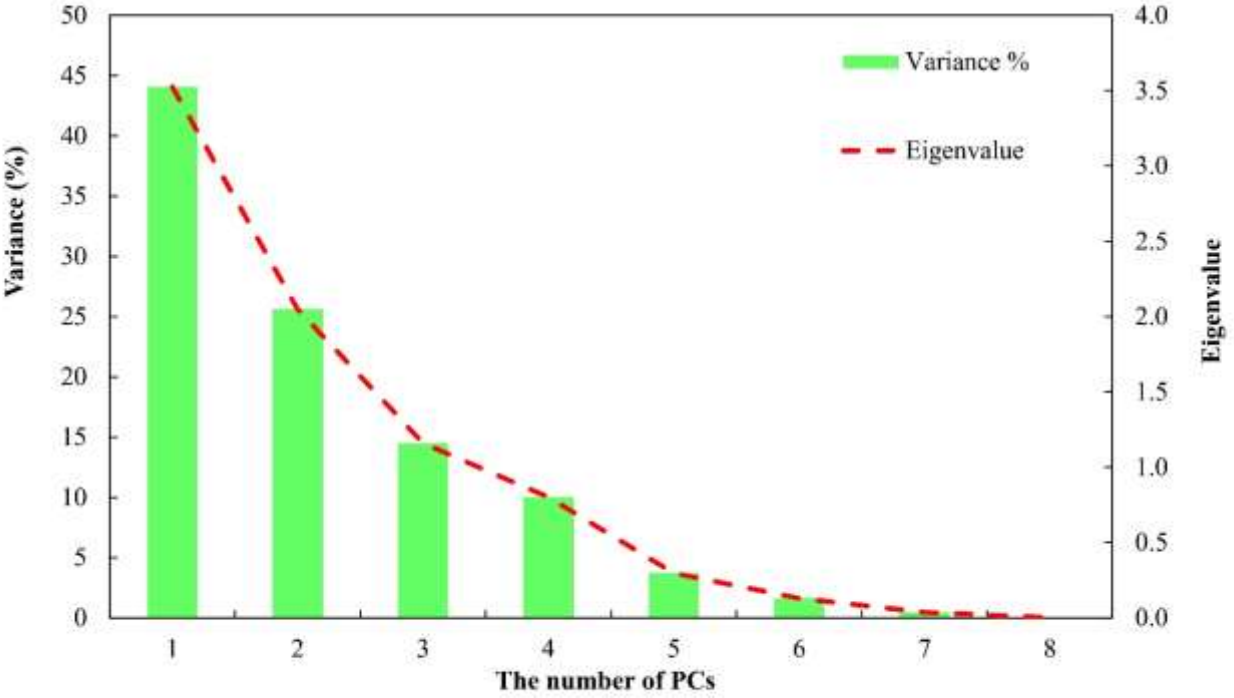


Fig. 2. Scree plot showing the variance of all components

3-2- Level 1: Performance of standalone models

The results obtained by using the presented standalone models are presented and discussed here. A comprehensive evaluation of the model predictions should include at least ‘goodness-of-fit’ such as R^2, NSE and error indices such as RMSE, or RAE.

The comprehensive comparison of the best single model results using the selected input variables by the GT and the quantitative values of performance evaluation indices of the MLR, MLR-PCA, ANN, ANN-PCA, GEP, GEP-PCA, GMDH, GMDH-PCA are presented in Table 5.

In the training step, the ensembles ANN-PCA model showed a relatively accurate estimation of bedload with ($R^2=1\approx0.996$), RMSE=0.71 when compared with the ANN ($R^2=0.98$, RMSE=1.66), GEP-PCA ($R^2=0.95$, RMSE=2.94) and the others. Based on the classification of model performances by the R^2 metric, all models in Table 5 had an outstanding performance ($0.7>R^2>1$) in bedload predictions, except the MLR-PCA. In the test stag, the same performance trend and accuracy improvement when combined the standalone models with the PCA were declared.

The best results were comparatively obtained by the ANN-PCA, GEP-PCA and ANN models. In this regard the ANN-PCA model with $R^2=0.96$, RMSE=0.38, RAE=0.16 and NSE=0.95 have the best predictions for the bedload in the test stage. The NSE values of the GEP-PCA, ANN-PCA, GEP and ANN models in the train and testing steps confirmed excellent predictions for the bedload transport in the test stage with NSE>0.75. The best accuracy of the GEP-PCA and ANN PCA-based models confirmed their ability in the emerging non-linear system indentation when combined GT and PCA's pre-processing model-free techniques.

The hierarchical accuracy of models follows the order of ANN-PCA> ANN> GEP-PCA> GEP> GMDH-PCA> GMDH> MLR-PCA> MLR in terms of the R^2, RMSE, RAE and NSE values for the test stage, as given in Table 5. The percent of prediction improvements by utilizing the PCA as input dimension reduction in RMSE reduction was 57% and 3% in ANN-PCA, 4% and 4% in GEP-PCA, 9% and 45% in GMDH-PCA for train and testing steps, respectively. The explicit form of predictive equations based on the trained above eight models are as follows:

MLR:

$$MLR = -3.39S + 0.02D_{gr} - 2.02\frac{U}{u_{*c}} - 5.4\frac{H}{W} - 8.3F_r - 0.002R_{e*} - 46.2\theta + 1.8\,\frac{U}{u_*} + 4.81 \quad 11$$

MLR-PCA:

$$MLR - PCA = 1.95 - 2.81PC_1 + 6.35PC_2 + 3.65PC_3 \quad 12$$

GMDH:

$$G_1 = 1.9Se^{0.27(\theta+0.64)} + 0.16S^2 + 12e^{0.27(\theta+0.64)} - 2.84S - 11.44$$

$$G_2 = 411.52 * e^{0.002(G_1+1.73)} - 411.4 \quad 13$$

$$G_3 = 115671.3e^{0.000009(G_2+1.35)} - 115671.3$$

$$GMDH = 970.9e^{0.000956(G_3+1.36)} - 970.79$$

GMDH-PCA:

$$GMDH - PCA = \frac{504.35}{1+e^{4.55PC_3-17.7}} + 1.78e^{0.998+0.3008PC_3+0.55PC_2-0.225PC_1} + 0.01e^{-(3.71PC_1+10.83)} - 0.61PC_3 + 0.44PC_3 \times PC_1 - 506.39 \quad 14$$

GEP:

$$GEP = \theta \times Fr + e^{Fr-2.21} + 6.2e^{\theta} - \theta - S - 7.14 \quad 15$$

GEP-PCA:

$$GEP - PCA = 33.6e^{-\frac{(PC_2-7.81)^2}{2PC_3^2}} + 1479391.4e^{-\frac{(PC_2-24.13)^2}{2PC_3^2}} - 1.75e^{-0.37(PC_2-PC_1)^2} + 2^{(PC_2-PC_1)} + 0.73 \quad 16$$

ANN:

$$ANN = \frac{530.2}{1+e^{-2(T_1+T_2+T_3+T_4)-28.3}} - 265.1 \quad 17$$

$$T_1 = \frac{16.36}{1 + e^{0.66S - 12.94D_{gr} + 5.18\frac{U}{u_{*c}} + 0.2\frac{H}{W} + 5.44F_r + 7.66R_{e*} - 4.34\theta - 5.18\frac{U}{u_*} - 5.86}} - 8.18$$

$$T_2 = \frac{-41.46}{1 + e^{31.22S + 1.74D_{gr} + 109.78\frac{U}{u_{*c}} - 165.86\frac{H}{W} + 136.76F_r + 2.3R_{e*} + 92.44\theta + 117.63\frac{U}{u_*} - 271.4}} + 20.73$$

$$T_3 = \frac{-2.98}{1 + e^{1.26S - 12.64D_{gr} + 4.8\frac{U}{u_{*c}} + 0.1\frac{H}{W} + 5.28F_r + 7.44R_{e*} - 2.68\theta - 4.72\frac{U}{u_*} - 4.2}} + 1.49$$

$$T_4 = \frac{0.24}{1 + e^{26S - 4.78D_{gr} - 2.52\frac{U}{u_{*c}} + 9.24\frac{H}{W} + 6.06F_r - 0.3R_{e*} - 38.42\theta - 0.38\frac{U}{u_*} + 9.38}} - 0.12$$

ANN-PCA:

$$ANN - PCA = \frac{530.2}{1 + e^{-2(T_1 + T_2 + T_3) - 68.43}} - 265.1 \qquad 18$$

$$T_1 = \frac{-5.065}{1 + e^{-37PC_1 + 34.63PC_2 - 2PC_3 - 30.2}} + \frac{0.09}{1 + e^{20.2PC_1 - 16.9PC_2 + 2.75PC_3 + 9.68}} + 2.48$$

$$T_2 = \frac{68.9}{1 + e^{1025.7PC_1 + 34.7PC_2 + 195.9PC_3 + 347.4}} + \frac{5.64}{1 + e^{-36.4PC_1 + 18.4PC_2 - 8.6PC_3 - 18.6}} - 37.3$$

$$T_3 = \frac{-0.03}{1 + e^{-9.75PC_1 + 17.1PC_2 + 0.3PC_3 + 3.6}} + 0.013$$

The scatter plots of the measured bedload rate against predicted by the models are presented in Fig.3. This figure shows that MLR, MLR-PCA, GMDH and GMDH-PCA model have underestimations for the bedload rate. As the results in this figure confirmed, the GEP-PCA, ANN-PCA models are most consistent with the 1:1 line and provide superior predictions for bedload transport rate in rivers compared to the standalone models of ANN, GEP, MLR, and GMDH.

As the first main motivation and contribution of the current study was to introduce the feasibility of utilizing the pre-processing model-free techniques of SSMD, GT and PCA and their ensemble ability with standalone models for bedload transport rate prediction in rivers, these techniques show an improved generalization capacity than non-preprocessed predictions and is confirmed with high estimation accuracy obtained.

Table 5. The statistical measures of standalone models in the train and testing steps

						Train			
		MLR-		ANN-		GEP-		GMDH-	
	MLR	PCA	ANN	PCA	GEP	PCA	GMDH	PCA	MME
R^2	0.76	0.37	0.98	0.996	0.94	0.95	0.80	0.55	0.997
RMSE	6.27	10.27	1.66	0.71	3.06	2.94	12.77	9.39	0.6
RAE	0.90	1.39	0.10	0.08	0.24	0.20	2.28	1.26	0.06
NSE	0.76	0.37	0.98	0.996	0.94	0.95	0.02	0.47	0.997

						Test			
		MLR-		ANN-		GEP-		GMDH-	
	MLR	PCA	ANN	PCA	GEP	PCA	GMDH	PCA	MME
R^2	0.78	0.76	0.96	0.96	0.93	0.92	0.88	0.89	0.98
RMSE	5.66	4.75	0.37	0.36	0.56	0.54	6.16	3.37	0.24
RAE	3.06	2.83	0.18	0.16	0.34	0.25	4.21	2.16	0.1
NSE	-9.03	-6.07	0.96	0.95	0.90	0.91	-10.91	-2.57	0.98

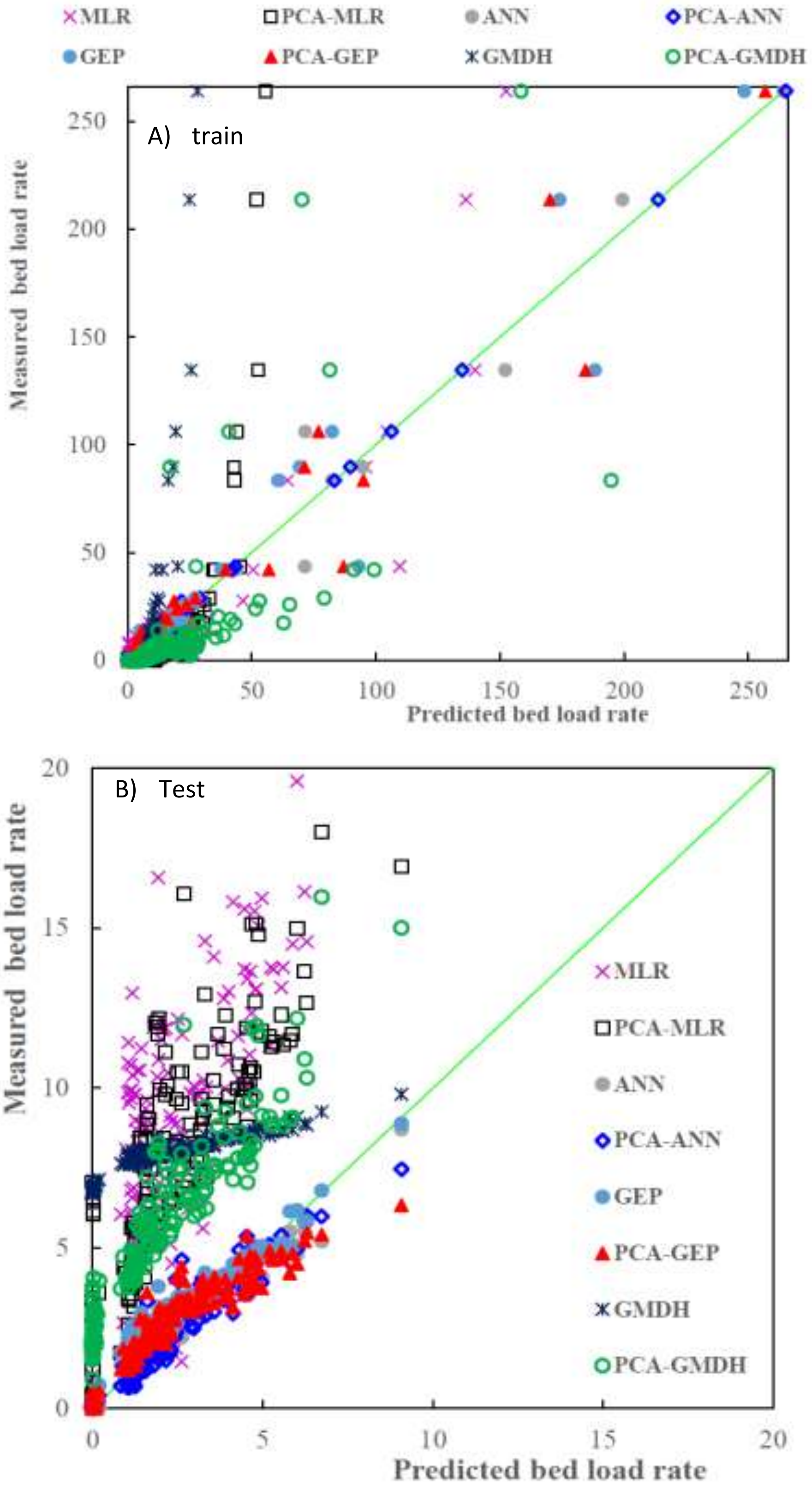


Fig. 3. Scatter plots of observed bedload rates versus prediction by standalone models in A) tarin and B) test data sets

3-3- Level 2: Performance of EMM approach: Ensembles-POMGGP

In the developed new strategy of EMM approach for bedload transport predictions, the Pareto optimality in conjunction with the multi-gene genetic programming is used to predict bedload transport by considering the output of standalone models. In this strategy the MLR, MLR-PCA, ANN, ANN-PCA, GEP, GEP-PCA, GMDH and GMDH-PCA predictions are used as the input vector to the POMGGP model and the feasible inputs are selected automatically by the geniting programming.

The Multi-Model input variable importance is shown in Fig. 4. As this figure shows, the most important sub-model is the ANN with an importance probability of 0.301, followed by the ANN-PCA sub-model with an importance probability of 0.286, and the MLR model with an importance probability of 0.225. Less important sub-models in the ensembles multi-model for predicting the dimensionless bedload transport rate follows the order of GMDH (probability=0.075)> GEP (probability=0.071)> GMDH-PCA (probability=0.029)> GEP-PCA (probability=0.014)> and MLR-PCA (probability=0.0).

The importance probability graph of sub-models in Fig. 4 shows that using the results of ANN, ANN-PCA, MLR, GMDH and GEP models, we are able to derive a predictive equation with an importance probability of 95.8%. So, in order to reduce the complexity of the final multi-model, and increase the application feasibility of the results, the Pareto optimality is used to derive the equation of final multi-model.

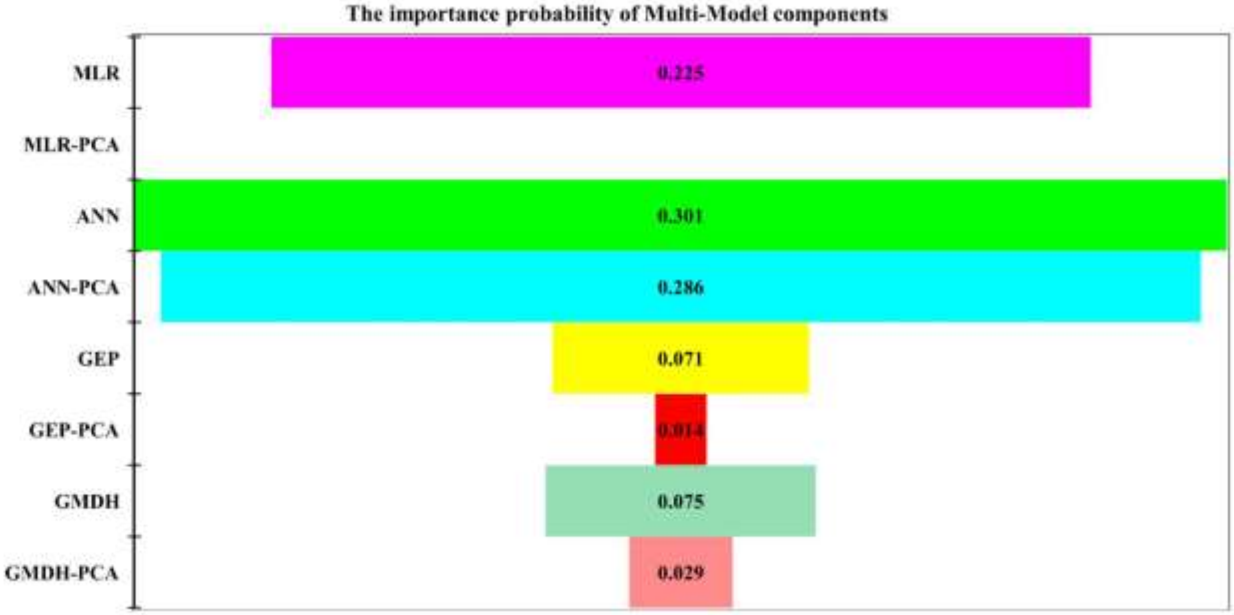


Fig. 4. The Multi-Model input variable (standalone models) importance

The parameters in Table 6 are determined by trial and error and using those suggested in the literature. The multi-gene genetic programming is trained and optimized by the least square error, the RMSE as the fitness function, basic math operators in function set, and Pareto optimality as the selection criteria. The Pareto graph of the evolved multi-models for bedload predictions using all sub-models as inputs, i.e: MLR, MLR-PCA, ANN, ANN-PCA, GEP, GEP-PCA, GMDH, and GMDH-PCA are shown in Fig. 5.

The Pareto-optimal solution of different multi-models on the Pareto front are chosen not more than 10% decrease occurred in model accuracy neither in the train nor at testing step. In this figure, the Pareto front is demonstrated with green circles, and the best final multi-model as the optimal solution is displayed by a circle with red perimeter and green color filled. The structural properties of the final multi-model include the overall complexity of 367, with 89 nodes in the selected symbolic expression, 4 individual genes, depth value 6 and -7.77 as the bias term, with MLR, ANN, ANN-PCA, GEP-PCA, GMDH as selected optimum input sub-models in agreement with probability importance graph in Fig. 4.

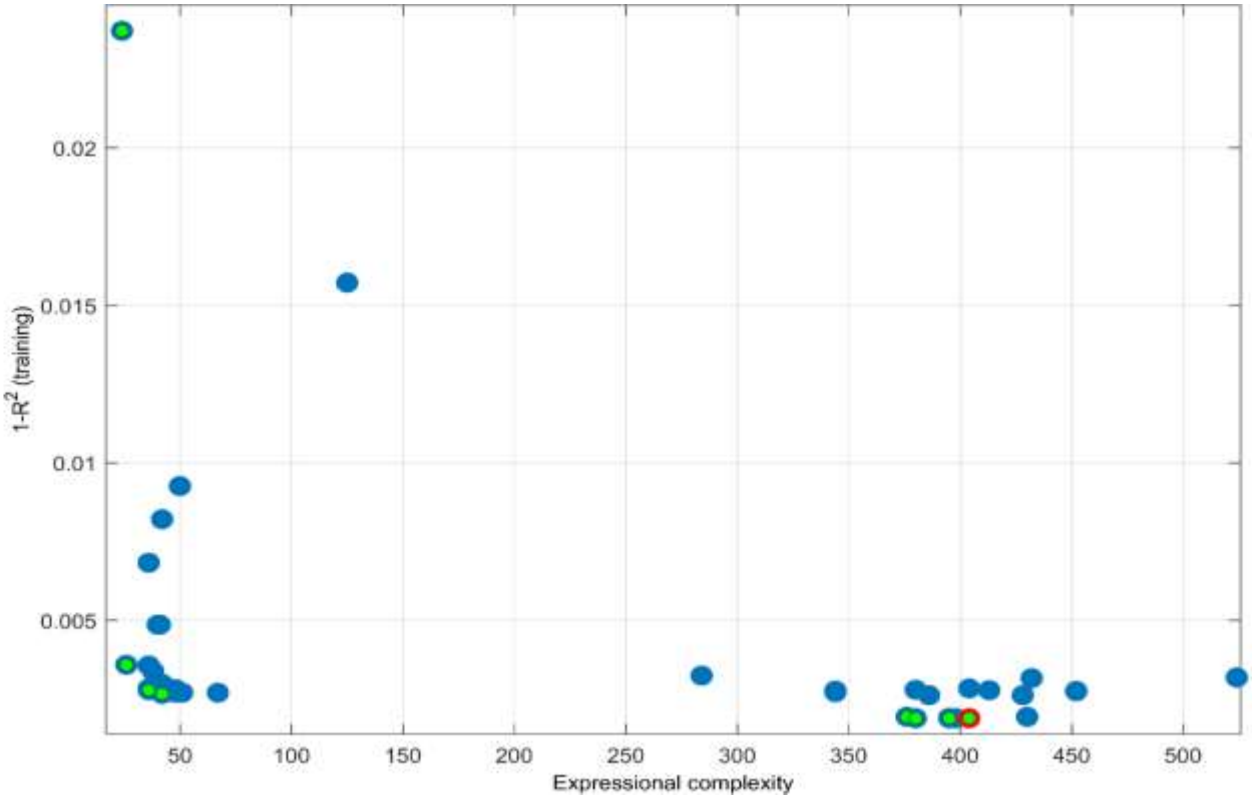


Fig. 5. Pareto graph of the best evolved multi-models

The final parse tree of the Pareto selected multi-model is presented in Fig. 6. This figure presents the symbolic expression of each gene in the multi-gene model. The corresponding equation and simplified expression of each gene, the individual gene weights, the number of nodes, individual complexities and depths are presented in Table 7.

As shown in Table 7, The bias term with weight=-7.77, Gene 2 that includes ANN and ANN-PCA with weight=7.6, Gene 4 with weight= -3.86 added the MLR, GMDH followed by Gene 1, has the highest weight and importance in the multi-gene model solution. To evaluate the statistical significance of individual genes the p-value of each gene calculated and the p-values in all of the genes were smaller than 0.00001, confirms and indicates the statistical importance of individual genes in the multigene model. Finally, by applying the coefficients of individual genes and simplifying the final Pareto solution of multi-gene expression, the final explicit predictive equation for dimensionless bedload rate based on the developed multi-model strategy with its effective sub-models is derived as

$$\emptyset = 1.13\mathrm{ANN}_{\mathrm{PCA}} - 0.079ANN - 0.073GEP_{PCA} + 0.027MLR + 7.6e^{\left(Exp\left(-3.34Exp\left(ANN_{PCA}{}^{2}\right)\right)\right)} + \frac{3.93ANN}{GMDH} - \frac{4ANN_{PCA}}{GMDH} + \frac{0.073MLR}{GMDH} - \frac{ANN_{PCA}}{13.74GMDH - 3.82ANN_{PCA} + \frac{13.74MLR}{ANN}} - 7.77 \qquad 19$$

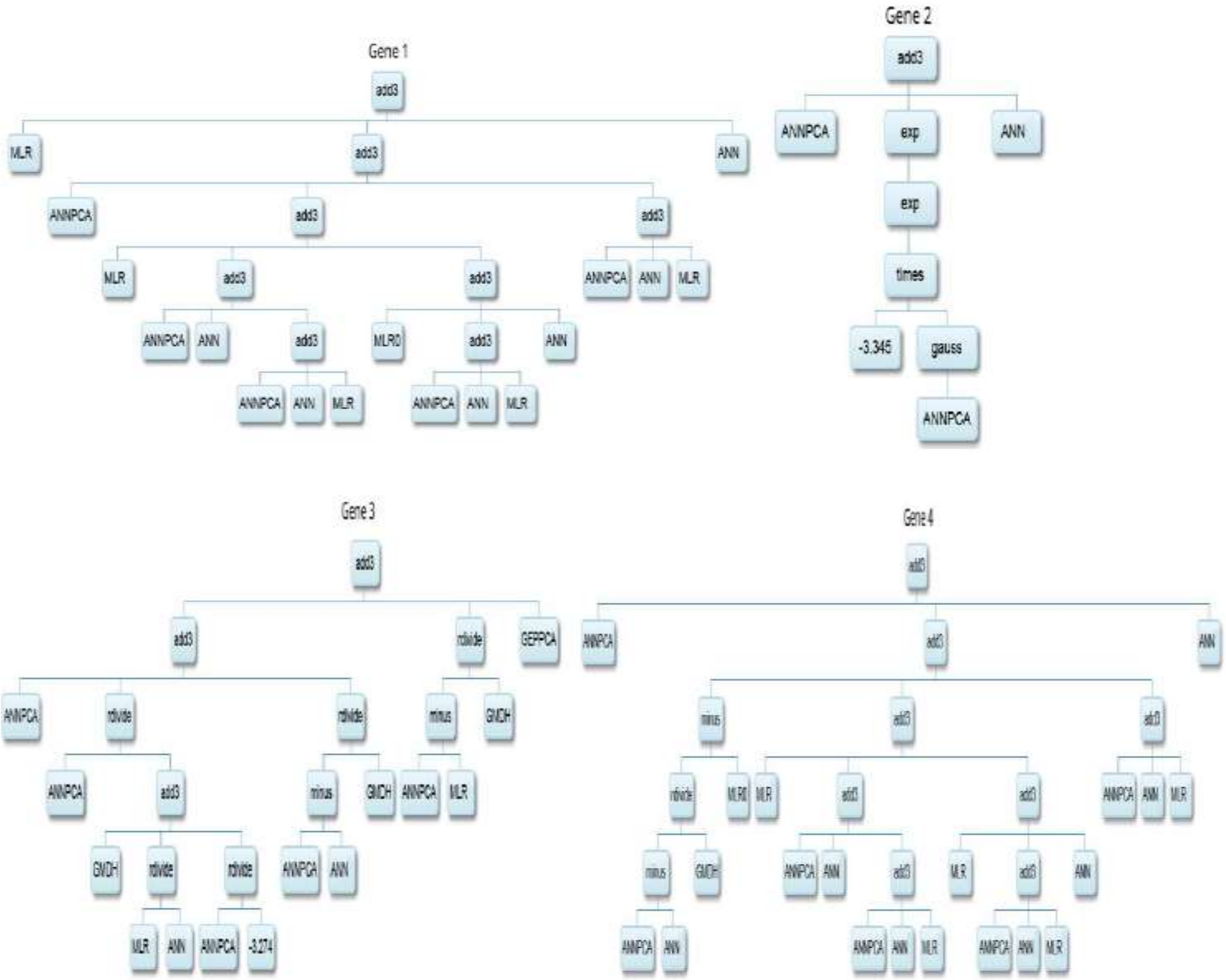


Fig. 6. The final parse tree of the Pareto selected multi-model

Performance indices of the final MME predictions for bedload are compared in Table 5 in train and test stages. Graphically, the results of Eq. 34 as the final multi-model solution are presented in Figs. 7 and 8 for the test stage and in Figs. 9 and 10 for the train stage. The scatter plots and series plots show the multi-model is accurately capable of capturing low and high values of bedload with different conditions in input observations. This is one remarkable aspect of our multi-model in mimicking low and high flows.

These results revealed that the multi-model approach improved the generalization capacity of single standalone single models, as confirmed with better estimation accuracy obtained in this extensive dataset (Train: R^2=0.997, RMSE=0.6, RAE=0.06, NSE=0.997, and in test Train: R^2=0.98, RMSE=0.1, RAE=0.24, NSE=0.98. Comparing the results of the training period of multi-model with the greatest improvement, about 16% in RMSE, and 25% in RAE was obtained compared to the best standalone model, ANN-PCA.

Based on the results in training step, the Multi-model had a decrease of 90% (in RMSE, RAE) and an increase of 31% (in NSE, R^2) compared to MLR. Multi-Model also showed a decrease of 95% (in RMSE, RAE) and an increase of 169% (in NSE, R^2) compared to MLR-PCA; a decrease of 63% and 40%% (in RMSE, RAE) and an increase of 2% (in NSE, R^2) compared to ANN, a decrease of 80% and 75%% (in RMSE, RAE) and an increase of 6% (in NSE, R^2) compared to GEP; a decrease of 79% and 70% (in RMSE, RAE) and an increase of 5% (in NSE, R^2) compared to GEP-PCA in train stages.

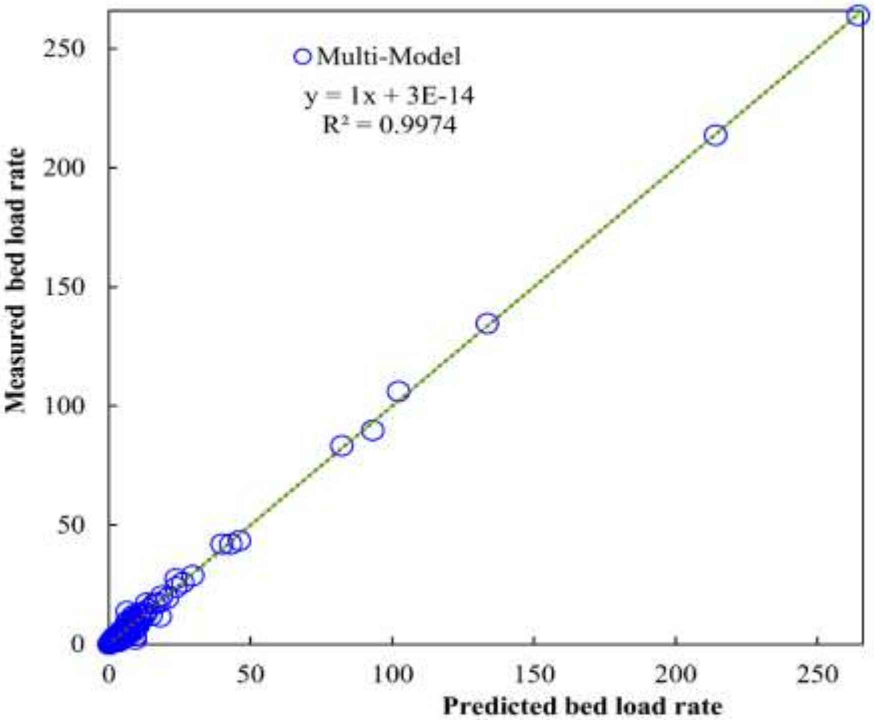


Fig. 7. Scatter plot of multi-model in training step

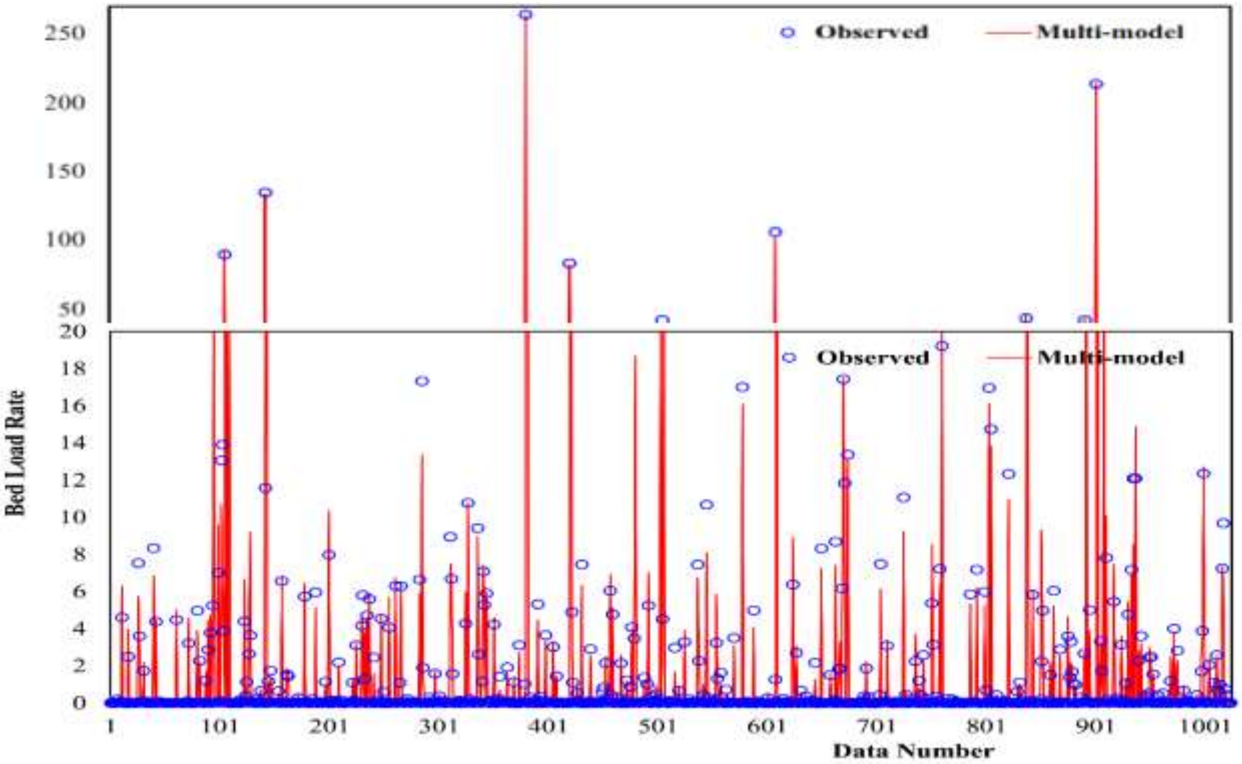


Fig. 8. Comparison of observed versus predicted bedload transport in training step

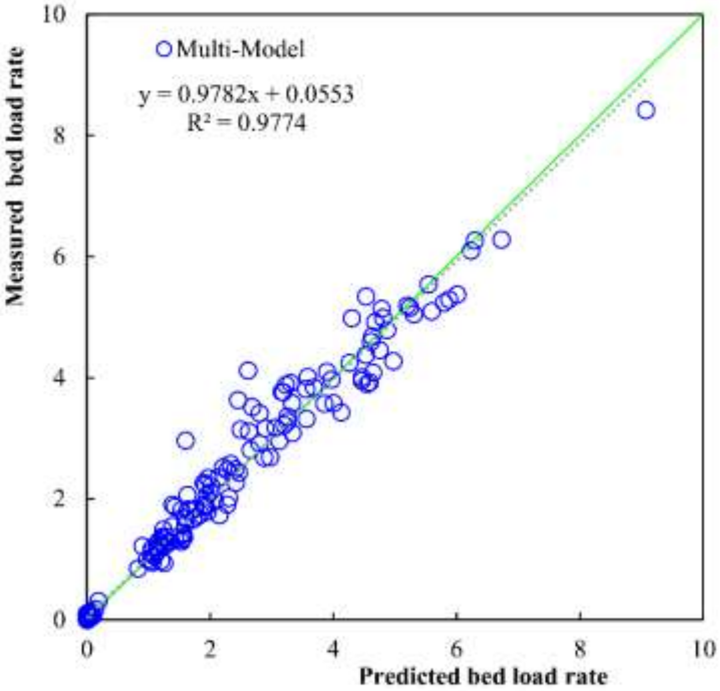


Fig. 9. Scatter plot of multi-model in testing step

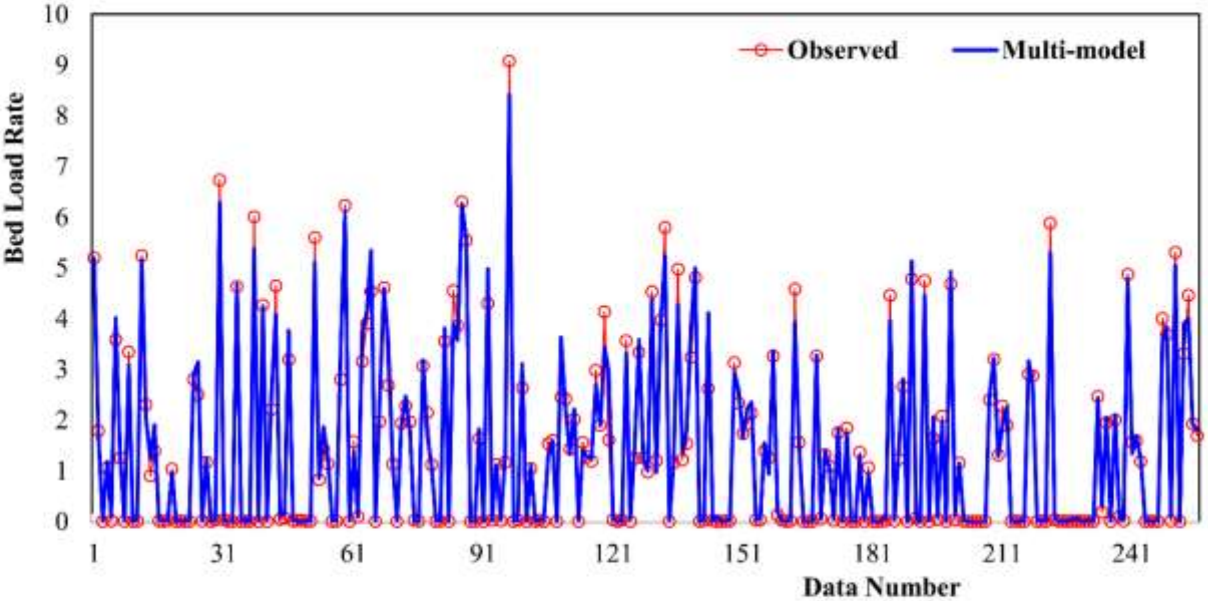


Fig. 10. Comparison of observed versus predicted bedload transport in testing step

The percentage of improvements in the test stage of multi-Model when these results are compared with other standalone predictions, are presented in Table 8. The improvement percentages in this table indicates that utilizing the multi-model strategy the RMSE and RAE values, as major error indices are decreased from 33% in ANN-PCA up to 96% in MLR and GMDH. In the R^2 measure the improvement varies from 2% up to 29%, and in the NSE, the gain varies from 2 up to 138%. These values confirm the superiority of the developed strategy in the generalization of bedload prediction.

To compare the underestimation or overestimation of the multi-model with the other models, in Figs. 11 and 12 the standardized error distribution of prediction in terms of RDR versus probability and the Taylor diagram of all models in train and test stages are shown. As these figures show the MLR, MLR-PCA, ANN and GEP models have underestimated and the GMDH-PCA, GMDH, ANN-PCA and GEP-PCA models overestimated for the bedload, while the multi-Model have reasonable estimates in training step. In the testing step, the RDR graph in Fig. 10, declares that the multi-Model strategy provides more generalities in the predictions and the RDR distribution is accurately around the 0, while the ANN, GEP, MLR and MLR-PCA have considerable underestimates and ANN-PCA, GEP-PCA, GMDH-PCA and GMDH have high overestimate in bedload. Reasonable accuracy and generality and parsimonious structure, endorse the developed multi-model approach for bedload transport estimation in practice. The leading cause behind the improvement in MME originates from the inherent multi-process nature and different patterns of bedload transport in the extremely low flows up to high flows is that the sediment transport is a mixture of a laminar, turbulent, linear and nonlinear phenomenon in rivers that would be taken into account by integration of linear and nonlinear models.

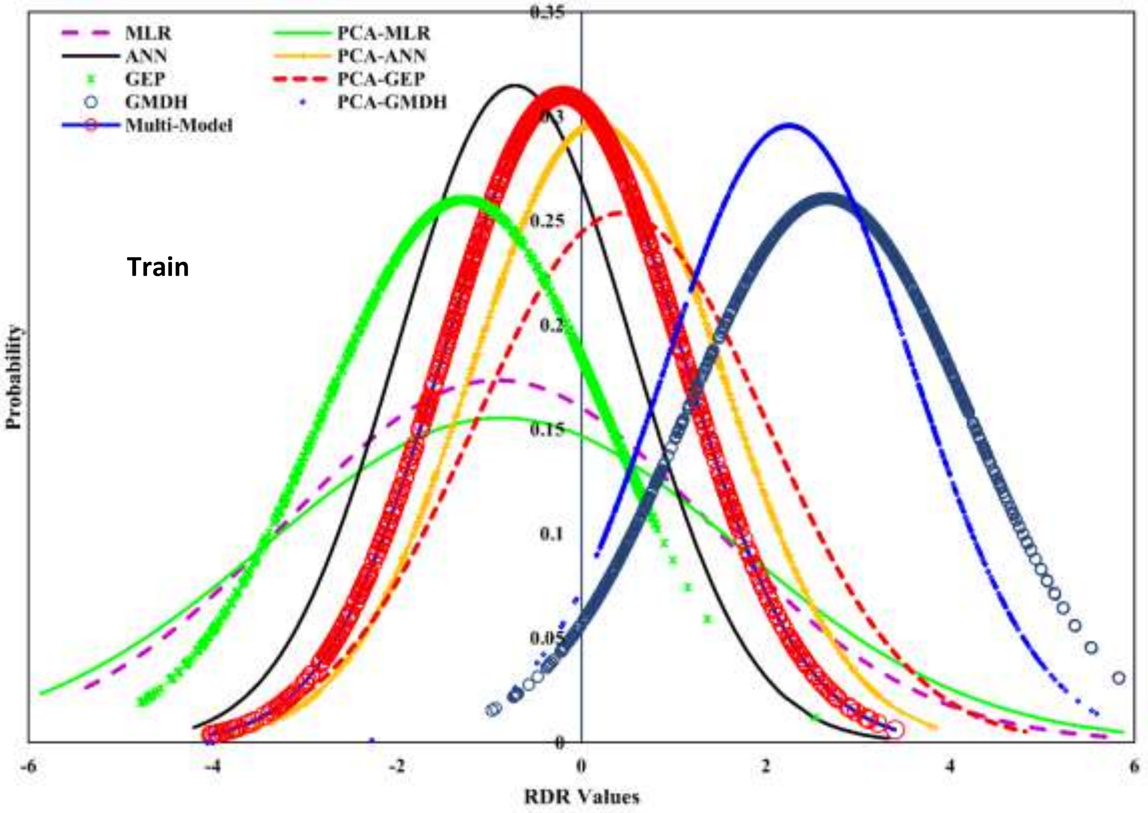
Table 6. Parameter setting for the MME development.

Run parameter	Value	Run parameter	Value
Population size	100	Gaussian perturbation of constant	0.05
Max. generations	500	Max. genes	4
Generations elapsed	500	Max. tree depth	6
Input variables	8	Max. total nodes	Inf
Training instances	1024	ERC probability	0.3
Tournament size	50	Crossover probability	0.84
High level Crossover	0.2	Low level Crossover	0.8
Elite fraction	0.75	Mutation probabilities	0.14
Sub-tree mutation	0.9	Input Mutation probabilities	0.05
Lexicographic selection pressure	On	Complexity measure	Expressional
Function set	*, -, +, /, ^,√ , exp, ln, multi3, cub, gauss, add3, square,		

Table 7. The Multi-gene results of Pareto solution in MME.

Term	Value	Gene weights	Nodes	Depth	Complexity
Bias	-7.77	-7.71	-	-	-
Gene 1	15.4 ANN + 12.9 ANN_{PCA} + 15.4 MLR	2.57	34	6	151
Gene 2	7.6 ANN + 7.6 ANN_{PCA} + 7.6 Exp(Exp(-3.34 gauss(ANN_{PCA})))	7.6	30	6	125
Gene 3	(0.0728 ANN)/GMDH - 0.0728 GEP_{PCA} - (ANN_{PCA})/(13.74 GMDH – 4.2 ANN_{PCA} + (13.74 MLR)/ANN) - 0.0728 ANN_{PCA} - (0.146 ANN_{PCA})/GMDH + (0.0728 MLR)/GMDH	-0.0728	9	6	31

Term	Value	Gene weights	Nodes	Depth	Complexity
Gene 4	-(3.86 (ANN_{PCA} -ANN + 6.0 ANN× GMDH + 5.0 ANN_{PCA} ×GMDH + 4.0GMDH ×MLR))/GMDH	-3.86	23	6	97
Overall Structure of Multi-Model: Genes:4; Nodes:89; Complexity: 367; Depth:6; Inputs selected: MLR, ANN, ANN-PCA, GEP-PCA, GMDH					



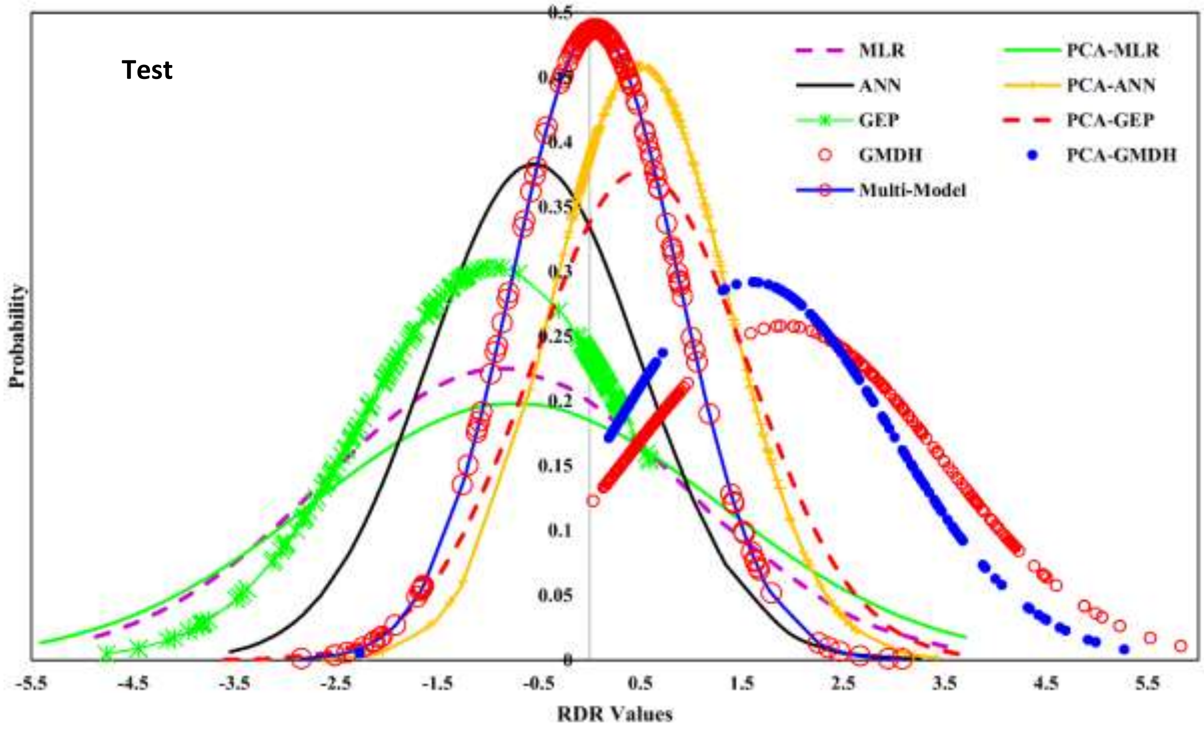
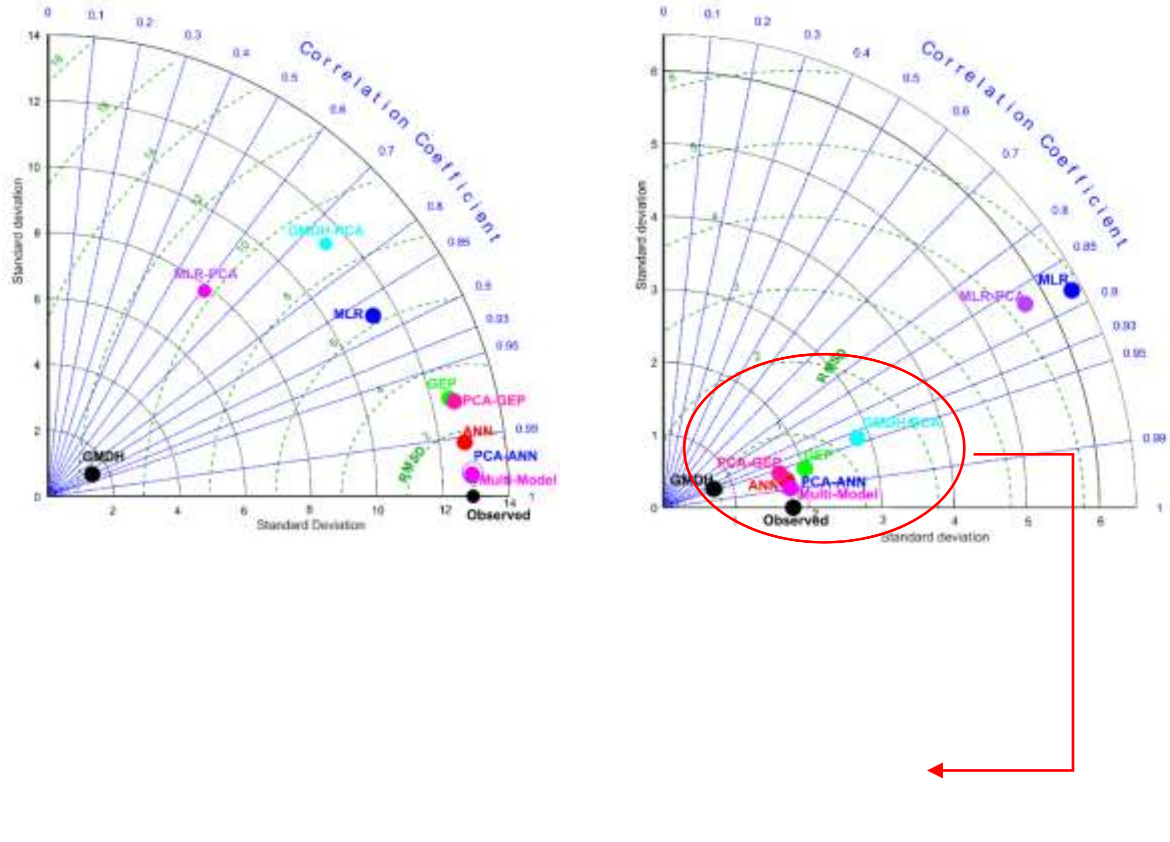


Fig. 11. The RDR graph in train and test stages of the MME



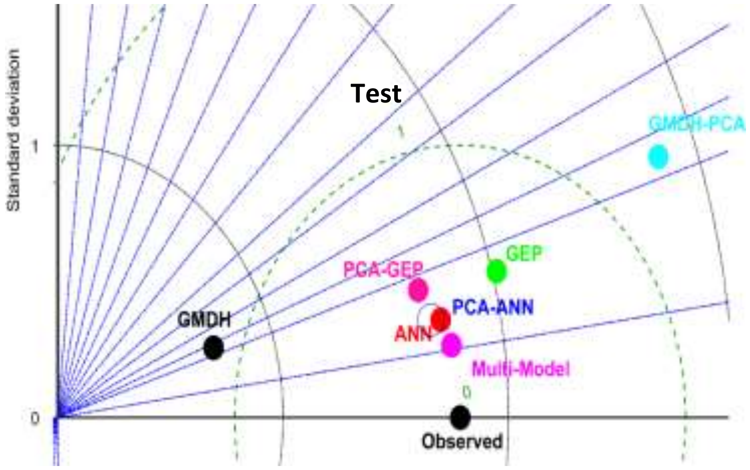


Fig. 12. The Taylor diagram in train and test stages

Table 8. The percentage of improvements in bedload rate prediction with multi-Model strategy in testing step

	MLR	MLR-PCA	ANN	ANN-PCA	GEP	GEP-PCA	GMDH	GMDH-PCA
R^2	26	29	2	2	5	7	11	10
RMSE	-96	-95	-35	-33	-57	-56	-96	-93
RAE	-97	-96	-44	-38	-71	-60	-98	-95
NSE	111	116	2	3	9	8	109	138

4- Conclusions

In this study, a new multi-Model strategy integrated with pre-processing techniques of SSMD, GT, and PCA is developed to derive an explicit predictive equation for the bedload transport in rivers with extensive dataset. The framework of enhanced multi-modelling improves the accuracy and heuristic capability to learn tendencies within residuals of individual model results and gain an insight into the nature of bedload transport in three-level strategy.

At level 0, the pre-processing, input selection and dimension reduction are carried out by SSMD, GT, PCA. At level 1, the standalone models of MLR, MLR-PCA, ANN, ANN-PCA, GEP, GEP-PCA, GMDH, GMDH-PCA are compared to derive explicit predictive equations. At level 2, the EMM is developed by utilizing the output of individual models as an external input to the multigene genetic programming with Pareto optimality. The main conclusions of this ensemble modeling are as follows:

1- The hierarchical accuracy of models follows the order of ANN-PCA> ANN> GEP-PCA> GEP> GMDH-PCA> GMDH> MLR-PCA> MLR in terms of the R^2, RMSE, RAE and NSE values for the test stage.

2- The percent of prediction improvements by utilizing the PCA as input dimension reduction in terms of RMSE reduction was 57% and 3% in ANN-PCA, 4% and 4% in GEP-PCA, 9% and 45% in GMDH-PCA for training and testing steps respectively.

3- The MME had a decrease of 90% (in RMSE, RAE) and an increase of 31% (in NSE, R^2) compared to MLR, a reduction of 95% (in RMSE, RAE) and an increase of 169% (in NSE, R^2) compared to MLR-PCA; a decrease of 63% and 40%% (in RMSE, RAE) and an rise of 2% (in NSE, R^2) compared to ANN, a decrease of 80% and 75%% (in RMSE, RAE) and an increase of 6%

(in NSE, R^2) compared to GEP; a reduction of 79% and 70% (in RMSE, RAE) and an increase of 5% (in NSE, R^2) compared to GEP-PCA.

4- The explicit predictive equation based on EMM approach has resulted in the gaining of a robust system with significant predictive accuracy improvement, (i.e., 33–96% in terms of RMSE; 2-29% in terms of R^2, 2-138% in terms of NSE and 38-98% in terms of RAE in testing step).

Finally, the authors would like to acknowledge the not always subtle differences in the previous studies' data measurement/collection methods. These differences constitute a limitation of the current research and a potential source of error when compiling the data set for machine learning. However, most of the sources used for compiling the comprehensive data set needed for the training and testing of the machine learning models have followed similar data measurement methods and standard data analysis, and reporting protocols to serve a truly global international community of researchers in this field.

References

Afan, H. A., El-shafie, A., Mohtar, W. H. M. W., & Yaseen, Z. M. (2016). Past, present and prospect of an Artificial Intelligence (AI) based model for sediment transport prediction. Journal of Hydrology, 541, 902-913.

Ahmadianfar, I., Kheyrandish, A., Jamei, M., & Gharabaghi, B. (2021). Optimizing operating rules for multi-reservoir hydropower generation systems: An adaptive hybrid differential evolution algorithm. Renewable Energy, 167, 774-790.

Barry, J. J. (2007). Bed load transport in gravel-bed rivers. Boise, ID: University of Idaho. 164 p. Dissertation, USA.

Bhattacharya, B., Price, R. K., & Solomatine, D. P. (2007). Machine learning approach to modeling sediment transport. Journal of Hydraulic Engineering, 133(4), 440-450.

Cao, Z.(1997). Turbulent Bursting-based sediment entrainment fluctuation. J. Hydraul. Eng, 123(3), 233–236.

Dehghani, M., Seifi, A., & Riahi-Madvar, H. (2019). Novel forecasting models for immediate-short-term to long-term influent flow prediction by combining ANFIS and grey wolf optimization. Journal of Hydrology, 576, 698-725.

Dey, S. (2014). Fluvial hydrodynamics: Hydrodynamic and sediment transport phenomena. Berlin Heidel berg: Springer-Verlag, Berlin.

Ebtehaj, I., Bonakdari, H., Zaji, A. H., & Gharabaghi, B. (2021). Evolutionary optimization of neural network to predict sediment transport without sedimentation. Complex & Intelligent Systems, 7(1), 401-416.

Elkurdy, M., Binns, A. D., Bonakdari, H., Gharabaghi, B., & McBean, E. (2021). Early detection of riverine flooding events using the group method of data handling for the Bow River, Alberta, Canada. International Journal of River Basin Management, 1-12.

Gao, P. (2011). An equation for bed-load transport capacities in gravel-bed rivers. Journal of Hydrology, 402(3-4), 297-305.

Ghani, A. A., & Azamathulla, H. M. (2014). Development of GEP-based functional relationship for sediment transport in tropical rivers. Neural Computing and Applications, 24(2), 271-276.

Gholami, A., Bonakdari, H., Zeynoddin, M., Ebtehaj, I., Gharabaghi, B., & Khodashenas, S. R. (2019). Reliable method of determining stable threshold channel shape using experimental and gene expression programming techniques. Neural Computing and Applications, 31(10), 5799-5817.

Gholami, A., Bonakdari, H., Ebtehaj, I., Gharabaghi, B., Khodashenas, S. R., Talesh, S. H. A., & Jamali, A. (2018). A methodological approach of predicting threshold channel bank profile by multi-objective evolutionary optimization of ANFIS. Engineering Geology, 239, 298-309.

Khatibi, R., Ghorbani, M. A., Naghshara, S., Aydin, H. A. R. U. N., & Karimi, V. (2020). A framework for 'Inclusive Multiple Modelling'with critical views on modelling practices– Applications to modelling water levels of Caspian Sea and Lakes Urmia and Van. Journal of Hydrology, 587, 124923.

Kitsikoudis, V., Sidiropoulos, E., & Hrissanthou, V. (2014). Machine learning utilization for bed load transport in gravel-bed rivers. Water resources management, 28(11), 3727-3743.

Liu, M. Y., Huai, W. X., Yang, Z. H., & Zeng, Y. H. (2020). A genetic programming-based model for drag coefficient of emergent vegetation in open channel flows. Adv. Water Resour. 140, 103582.

Lu, C., Zhang, T., Zhang, R., & Zhang, C. (2003, April). Adaptive robust kernel PCA algorithm. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). (Vol. 6, pp. VI-621). IEEE.

Madvar, H. R., Dehghani, M., Memarzadeh, R., Salwana, E., Mosavi, A., & Shahab, S. (2020). Derivation of optimized equations for estimation of dispersion coefficient in natural streams using hybridized ANN with PSO and CSO algorithms. IEEE Access, 8, 156582-156599.

Memarzadeh, R., Zadeh, H. G., Dehghani, M., Riahi-Madvar, H., Seifi, A., & Mortazavi, S. M. (2020). A novel equation for longitudinal dispersion coefficient prediction based on the hybrid of SSMD and whale optimization algorithm. Science of The Total Environment, 716, 137007.

Meyer-Peter, E., Müller, R. 1948. Formulas for bed-load transport. In IAHSR 2nd meeting, Stockholm, appendix 2. IAHR.

Montes, C., Kapelan, Z., Saldarriaga, J. 2021. Predicting non-deposition sediment transport in sewer pipes using Random forest. Water Research, 189, 116639.

Noori, R., Karbassi, A., & Sabahi, M. S. (2010a). Evaluation of PCA and Gamma test techniques on ANN operation for weekly solid waste prediction. Journal of environmental management, 91(3), 767-771.

Noori, R., Khakpour, A., Omidvar, B., & Farokhnia, A. (2010b). Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic. Expert Systems with Applications, 37(8), 5856-5862.

Noori, R., Sabahi, M. S., Karbassi, A. R., Baghvand, A., & Zadeh, H. T. (2010c). Multivariate statistical analysis of surface water quality based on correlations and variations in the data set. Desalination, 260(1-3), 129-136.

Noori, R., Karbassi, A. R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M. H., Farokhnia, A., & Gousheh, M. G. (2011). Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. Journal of hydrology, 401(3-4), 177-189.

Qasem, S. N., Ebtehaj, I., Riahi Madavar, H., 2017. Optimizing ANFIS for sediment transport in open channels using different evolutionary algorithms. J. Appl. Res. Water Wastewater, 4(1), 290-298.

Ramachandran, P., Zoph, B., Le, Q. V. 2017. Searching for activation functions. arXiv preprint arXiv:1710.05941.

Recking, A., Boucinha, V., Frey, P. 2004. Experimental study of bed-load grain size sorting near incipient motion on steep slopes. River flow, Napple, 253–258.

Reid, I., Laronne, J.B., 1995. Bedload sediment transport in an ephemeral stream and a comparison with seasonal and perennial counterparts. Water Resour. Res. 31 (3), 773–781.

Remesan, R., Shamim, M. A., Han, D., Mathew, J. 2009. Runoff prediction using an integrated hybrid modelling scheme. J. Hydro., 372(1-4), 48-60.

Riahi-Madvar, H., Dehghani, M., Memarzadeh, R., & Gharabaghi, B. (2021). Short to long-term forecasting of river flows by heuristic optimization algorithms hybridized with ANFIS. Water Resources Management, 35(4), 1149-1166.

Riahi-Madvar, H., Seifi, A. 2018. Uncertainty analysis in bedload transport prediction of gravel bed rivers by ANN and ANFIS. Ara. J. Geosci. 11(21), 1-20.

Riahi-Madvar, H., Dehghani, M., Parmar, K. S., Nabipour, N., Shamshirband, S. 2020. Improvements in the explicit estimation of pollutant dispersion coefficient in rivers by subset selection of maximum dissimilarity hybridized with ANFIS-firefly algorithm (FFA). IEEE Access, 8, 60314-60337.

Riahi-Madvar, H., Dehghani, M., Seifi, A., Singh, V. P. 2019. Pareto optimal multigene genetic programming for prediction of longitudinal dispersion coefficient. Water resour. Manag. 33(3), 905-921.

Roushangar, K., Mehrabani, F. V., Shiri, J. 2014. Modeling river total bed material load discharge using artificial intelligence approaches (based on conceptual inputs). J. hydrol. 514, 114-122.

Roushangar, K., Shahnazi, S., 2020. Prediction of sediment transport rates in gravel- bed rivers using Gaussian process regression. J. Hydroinf. 22 (2), 249-262.

Safari, M. J. S., Mohammadi, B., Kargar, K. 2020. Invasive weed optimization-based adaptive neuro-fuzzy inference system hybrid model for sediment transport with a bed deposit. J.Clean. Prod. 276, 124267.

Sahraei, S., Alizadeh, M.R., Talebbeydokhti, N., Dehghani, M., 2017. Bed material load estimation in channels using machine learning and meta-heuristic methods. J. Hydroinf. 20, 100–116.

Searson, D. P. 2015. GPTIPS 2: an open-source software platform for symbolic data mining. In Handbook of genetic programming applications (pp. 551-573). Springer, Cham.

Seifi, A., Soroush, F. 2020. Pan evaporation estimation and derivation of explicit optimized equations by novel hybrid meta-heuristic ANN based methods in different climates of Iran. Comp. Elec. Agri. 173, 105418.

Shaghaghi, S., Bonakdari, H., Gholami, A., Kisi, O., Shiri, J., Binns, A. D., Gharabaghi, B. 2018. Stable alluvial channel design using evolutionary neural networks. J. Hydro. 566, 770-782.

Smith, L.I. 2002) A tutorial on principal components analysis. Cornell Univ. USA2002,51, 52.

Snieder, E., Shakir, R., Khan, U. T. 2020. A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models. J. Hydrol. 583, 124299.

Van Rijn, L.C., 1993. Principles of Sediment Transport in Rivers, Estuaries and Coastal Areas. Aqua Publications, Amsterdam, The Netherlands.

Zhang, Z., Wang, K., Zhu, L., Wang, Y. 2017. A Pareto improved artificial fish swarm algorithm for solving a multi-objective fuzzy disassembly line balancing problem. Exp. Sys. App. 86, 165-176.

Zounemat-Kermani, M., Mahdavi-Meymand, A., Alizamir, M., Adarsh, S., Yaseen, Z. M. 2020. On the complexities of sediment load modeling using integrative machine learning: Application of the great river of Loíza in Puerto Rico. J. Hydrol. 585, 124759.