Water Resour Manage (2013) 27:2541–2554
DOI 10.1007/s11269-013-0302-7

Principle Component Analysis in Conjuction with Data Driven Methods for Sediment Load Prediction

Gokmen Tayfur · Yashar Karimi · Vijay P. Singh

Received: 7 November 2012 / Accepted: 29 January 2013 /
Published online: 9 February 2013

Abstract This study investigates sediment load prediction and generalization from laboratory scale to field scale using principle component analysis (PCA) in conjunction with data driven methods of artificial neural networks (ANNs) and genetic algorithms (GAs). Five main dimensionless parameters for total load are identified by using PCA. These parameters are used in the input vector of ANN for predicting total sediment loads. In addition, nonlinear equations are constructed, based upon the same identified dimensionless parameters. The optimal values of exponents and constants of the equations are obtained by the GA method. The performance of the so-developed ANN and GA based methods is evaluated using laboratory and field data. Results show that the expert methods (ANN and GA), calibrated with laboratory data, are capable of predicting total sediment load in field, thus showing their transferability. In addition, this study shows that the expert methods are not transferable for suspended load, perhaps due to insufficient laboratory data. Yet, these methods are able to predict suspended load in field, when trained with respective field data.

Keywords Principle component analysis · Sediment load · Artificial neural network · Genetic algorithm · Transferability

1 Introduction

Since estimates of sediment loads are required in many fields of water resource engineering, considerable effort has been devoted to sedimentation engineering (Jain 2001; Tayfur 2002;

G. Tayfur (✉)
Department of Civil Engineering, Izmir Institute of Technology, Izmir, Turkey
e-mail: gokmentayfur@iyte.edu.tr

Y. Karimi
Department of Civil Engineering, Ege University, Izmir, Turkey
e-mail: yasharkarimi@hotmail.com

V. P. Singh
Department of Civil Engineering, Texas A and M University, College Station, USA
e-mail: vsingh@tamu.edu

Bhattacharya et al. 2007; Dogan et al. 2009, among others). Most of the existing models are based on combinations of several characteristics of flow and channel geometry, and sediment dynamics parameters. Zhu et al. (2006) summarizes the parameters used in several commonly employed models. Using artificial neural networks (ANNs), Bhattacharya et al. (2007) estimated sediment loads employing dimensionless parameters based mainly on studies of Yalin (1977) and van Rijn (1984). Bhattacharya et al. (2007) considered two scenarios by employing different sets of input variables to predict dimensionless total sediment transport rate. In the first scenario, they employed dimensional parameters of u (flow velocity), h (flow depth), D (particle diameter), and I (slope); and in the second scenario, they used D_* (particle parameter), T (transport stage parameter), and h/D to predict ϕ_t (dimensionless total sediment transport rate). They predicted suspended load, total load, and bed load for laboratory scale and field scale separately. They did not, however, investigate the transferability of their method from laboratory scale to field scale.

The importance of transferability is well documented in Dogan et al. (2009) who investigated it for total load from laboratory scale to field scale using the RVM (Relevance Vector Machine) method. They selected parameters, based on empirical methods, considering the ones having similar statistical distributions in laboratory and field. As a result, they employed q_* (dimensionless stream power), τ_* (Shields parameter), τ_*' (Shields parameter associated with grain or skin friction), and τ_{*c} (Shields parameter associated with incipient sediment motion) as input variables for predicting total sediment concentration *(C)*. It should be noted that in the parameter selection process for the predictive model for transferability, they considered both laboratory and field data. In doing so, they introduced a bias into their model. For the transferability study, the predictive mode should be constructed, based just upon laboratory data and this was exactly done in the present study.

The sediment load predictions and transferability were investigated using the artificial neural networks (ANNs) and genetic algorithms methods (GAs). The selection of dimensionless parameters for the input vector of the model developed in this study was based on the principal component analysis (PCA). Employing PCA for this purpose is advantageous, because it reduces the number of parameters to an optimal size, while preserving the original information as much as possible (Field 2005). Furthermore, it achieves parsimony by explaining the maximum amount of common variance in a correlation matrix using the smallest number of explanatory variables and avoids problems of multicollinearty and singularity (Field 2005).

2 Data

Brownlie (1981) composed an extended set of laboratory and field data which include 7,027 records (5,263 laboratory records and 1,764 field records). Details on each data set are available in Brownlie (1981). In line with Dogan et al. (2009), the following restrictions were imposed on the data employed in this study:

1. $\frac{B}{h}$ (where B is the channel width and h is the flow depth) is restricted to be greater than 4 to escape the sidewall effects.
2. Relative roughness, $\frac{R}{d_{50}}$ (where R is the hydraulic radius and d_{50} is the mean particle diameter) is limited to be greater than 100 to avoid extremely shallow condition.
3. Sediment size is limited to the sand range of $0.062(mm) < d_{50} < 2.0(mm)$.
4. Geometric standard (σ_g) is limited to be less than 5 to avoid extreme amounts of gravel or fine material.

5. Sediment concentration *(C)* is restricted to be greater than 10 ppm for accuracy of low concentration measurement.

Under these restrictions, 1,190 records for laboratory total load, 180 records for field total load, and 759 records for field suspended load were retained.

3 Dimensionless Parameters

Sediment transport rate is mainly a function of the following parameters (Yalin 1977; Dogan 2008):

$$C = f\left(u_*, q, d_{50}, \rho, \rho_s, h, B, \nu, \sigma_g, S, u_m, \mu, g\right) \tag{1}$$

where C = the sediment concentration (ppm); u_* = the shear velocity, (LT^{-1}), q = the unit flow discharge (L^2T^{-1}), d_{50} = the particle diameter such that 50 % (median) of particle size by weight is finer (L), ρ = the water density (ML^{-3}), ρ_s = the sediment density (ML^{-3}), h = the flow depth (L), B = the channel width (L), ν = the kinematic viscosity (L^2T^{-1}), σ_g = the sediment gradation, S = the slope, u_m = the average flow velocity (LT^{-1}), and g = the gravitational acceleration (LT^{-2}).

Performing a dimensional analysis using Buckingham's Pi theorem, Dogan (2008) first obtained 10 dimensionless parameters and then he added 8 more from the literature, as presented by Eq. (2):

$$C = f\left(\begin{array}{c} \frac{h}{d_{50}}, \frac{\rho}{\rho_s}, \frac{u_m h}{\nu}, \frac{u_* d_{50}}{\nu}, \frac{u_* h}{\nu}, \frac{hS}{(G_s-1)d_{50}}, \frac{u_m}{u_*}, \frac{B}{h}, \frac{q}{\sqrt{g\bar{h}h}}, \frac{q}{u_* d_{50}}, \frac{B}{d_{50}}, \\ \frac{\nu u_*}{g(G_s-1)d_{50}^2}, \frac{\nu^2}{g(G_s-1)d_{50}^3}, \frac{q^2}{g(G_s-1)d_{50}^3}, \frac{\rho_s u_*^2}{\gamma_s d_{50}}, \frac{u_m}{\sqrt{g(G_s-1)d_{50}}}, S, \sigma_g, \frac{R}{d_{50}} \end{array}\right) \tag{2}$$

Definitions of dimensionless parameters are given in Appendix I. We constructed the predictive models based upon these dimensionless parameters where sediment concentration (C) is considered as output variable.

In addition to the 18 parameters in Eq. (2), we further suggested $\frac{R}{d_{50}}$ dimensionless hydraulic radius in this study in order to reflect the effects of channel cross-section, flow depth and wetted perimeter by a single parameter.

In order to obtain the optimal number of parameters, we subjected the ones in Eq. (2) to the principal component analysis (PCA) for both the total load and suspended load. Note that we carried out this analysis for total load in field, suspended load in field, and total load in laboratory. As presented later, we were able to reduce 19 parameters to 6 in the case of field total sediment load and 5 in the cases of laboratory total and field suspended loads (Table 1).

4 Methods

4.1 Principal Component Analysis (PCA)

PCA is a technique for recognizing groups of variables. This method is useful for reducing the number of parameters to an optimal size, while preserving the original information as possible. By reducing a data set from a group of interrelated variables into a smaller set of variables, PCA achieves parsimony by explaining the maximum amount of common variance in a correlation matrix using the smallest number of explanatory concepts (Field

Table 1 Extracted component and loading coefficients

Laboratory	total load	$\frac{u_* h}{\nu}$	$\frac{\nu^2}{g(G_s-1)d_{50}^3}$	$\frac{R}{d_{50}}$	$\frac{q^2}{g(G_s-1)d_{50}^3}$	$\frac{\rho_s u_*^2}{\gamma_s d_{50}}$	
	PC 1	0.058	0.953	0.867	0.865	0.324	
	PC 2	0.929	−0.34	0.357	0.379	0.775	
Field	suspended load	$\frac{u_* h}{\nu}$	$\frac{u_m}{u_*}$	$\frac{R}{d_{50}}$	$\frac{q^2}{g(G_s-1)d_{50}^3}$	$\frac{u_m}{\sqrt{g(G_s-1)d_{50}}}$	
	PC 1	0.722	0.786	0.803	0.903	0.814	
	total load	$\frac{u_* d_{50}}{\nu}$	$\frac{\nu^2}{g(G_s-1)d_{50}^3}$	$\frac{R}{d_{50}}$	$\frac{q^2}{g(G_s-1)d_{50}^3}$	$\frac{\rho_s u_*^2}{\gamma_s d_{50}}$	$\frac{u_m}{\sqrt{g(G_s-1)d_{50}}}$
	PC 1	−0.725	0.753	0.758	0.776	0.838	0.916

$\frac{u_* h}{\nu}$ = Reynolds number related to shear stress, $\frac{\nu^2}{g(G_s-1)d_{50}^3}$ = dimensionless particle size, $\frac{R}{d_{50}}$ = dimensionless hydraulic radius, $\frac{q^2}{g(G_s-1)d_{50}^3}$ = dimensionless unit flow discharge, $\frac{\rho_s u_*^2}{\gamma_s d_{50}}$ = mobility number (related to particle size), $\frac{u_m}{u_*}$ = friction factor, $\frac{u_m}{\sqrt{g(G_s-1)d_{50}}}$ = Froude number related to particle size, $\frac{u_* d_{50}}{\nu}$ = Reynolds number related to particle size

2005). PCA can be carried out, based on correlation and covariance matrix of variables. Actually, these two matrices are different versions of the same thing. There are myriad applications of PCA in water resource engineering, hydrology and environmental sciences (Winter et al. 2000; Loska and Wiechula 2003; Ouyang 2005; Noori et al. 2010). Before the PCA application, one has to control 'sample size quality,' and 'data screening,' as presented below.

4.1.1 Sample Size Quality

The sample size requirement has to be investigated before PCA. The reliability of PCA depends on the sample size which is important due to the generalization of model results from laboratory to field scale. Additionally, the fluctuation in the correlation coefficient from sample to sample, particularly significant in small size samples, affects PCA. Field (2005) classified sample size 100 as poor, 300 as good, and 1000 as excellent. In our study, we had 1,190 records for laboratory total load, 180 for field total load, and 759 field suspended load. Thus, only field total load data is between poor and good range for PCA. We also carried out KMO (Kaiser-Meyer-Olkin) criterion to check the adequacy of sample sizes. KMO is a quantity of sampling adequacy that compares the value of the calculated correlation coefficients to the values of the partial correlation coefficients. It can be shown as (Pett et al. 2003):

$$KMO = \frac{\sum (correlation)^2}{\sum (correlation)^2 + \sum (partialcorrelation)^2} \tag{3}$$

KMO varies between 0 and 1. If the partial correlation is 0, then KMO will be 1, implying that the variables are measuring a common component, or vice-versa. According to Field (2005), for PCA, the minimum value of KMO is 0.5. This criterion was also satisfied for all the samples.

4.1.2 Data Screening

Data screening is the next pre-procedure before PCA to be carried out to avoid problems of multicollinearty (variables that are very highly correlated, $R>0.90$) and singularity (variables

that are perfectly correlated, $R \sim 1$) in input variables. In other words, by the data screening, one eliminates highly and perfectly correlated variables. In order to avoid the multicollinearty and singularity in the analysis, the variables should be inspected at the beginning. The correlation matrix can offer useful information about the multicollinearty, determined by the determinant of the matrix which should be greater than 1×10^{-5}.

In this study, the dimensionless variables were subjected to data screening before the PCA application. As a result, due to multicollinearty and singularity, we eliminated $\frac{h}{d_{50}}, \frac{u_m h}{\nu}, \frac{hs}{(G_s-1)d_{50}}, \frac{q}{u_* d_{50}}, \frac{B}{d_{50}}, \frac{\nu u_*}{g(G_s-1)d_{50}^2}$ for the laboratory total load, $\sigma_g, \frac{h}{d_{50}}, \frac{R}{d_{50}}, \frac{u_m h}{\nu}, \frac{hS}{(G_s-1)d_{50}}, \frac{q}{u_* d_{50}}, \frac{B}{h}, \frac{\nu u_*}{g(G_s-1)d_{50}^2}$ for the field total load, and $\frac{h}{d_{50}}, \frac{R}{d_{50}}, \frac{u_m h}{\nu}, \frac{hs}{(G_s-1)d_{50}}, \frac{q}{u_* d_{50}}$ for the field suspended load. After this elimination, the determinants of R-matrices for each data set were achieved to be 2.73×10^{-5}, 1.71×10^{-5} and 8.66×10^{-5}, respectively. After these pre-procedures were satisfied, PCA was initiated.

4.1.3 Communality

The communality is known as the proportion of common variance present in a variable (Field 2005). If it is 0, it means that the variable does not share variance with other variables. If it is equal to 1 then the variable has no particular variance (Field 2005). The solution should explain at least half of each original variance of a variable, such that the communality value for each variable should be 0.50 or higher. Due to the communality check, we eliminated $\frac{B}{h}$ for the laboratory total load, and $\frac{B}{h}$ and $\frac{\nu u_*}{g(G_s-1)d_{50}^2}$ for the field suspended load.

4.1.4 Component Rotation

The next step is to carry out the rotation during PCA. Mathematically, components are linear combinations of independent variables expressed as (Field 2005):

$$PC_i = b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + \varepsilon_i \tag{4}$$

where PC_i = the principal component, X_n = the independent variable which is loaded in the component, b_n = the component loading coefficient, presenting the relative contribution of each variable (Field 2005), and ε_i = the residual.

Figure 1 shows the component rotation for a two component case. The components can be visualized as axis and variables can be plotted on it (see Fig. 1). Hence, it is possible to calculate the degree to which variables load on to the components. Generally, variables load highly on the most important component, and load slightly on the other component (see Fig. 1). This is a qualitative characteristic and hence discrimination between components can be less than objective. Therefore, the rotation technique is employed in order to determine the importance of each variable in each component (see Fig. 1).

Table 1 summarizes the number of components and loading values for each variable for laboratory total, field total, and field suspended loads. Note that PC_s are linear combinations of dimensionless parameters, as shown by Eq. (4). Some studies use PC_s directly (Noori et al. 2010). However, in this study dimensionless parameters were loaded in PC_s as input vectors.

4.1.5 Validation of PCA

In order to validate the findings from PCA, a split-half-sample method, which randomly divides the whole sample into two parts and applies PCA to each part, was employed. The

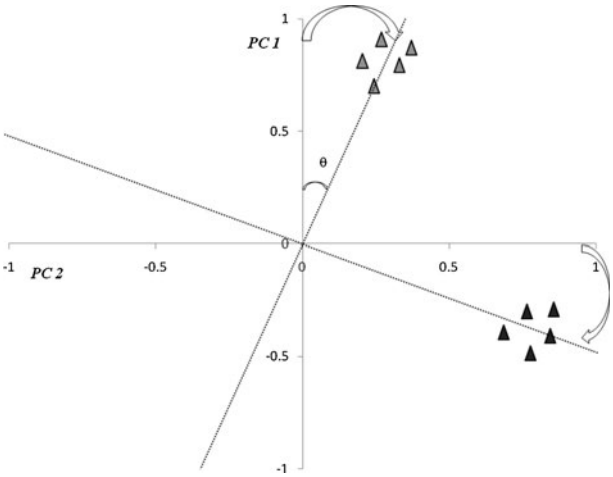

Fig. 1 Schematic illustration of component rotation

method satisfied communalities, component loading, and KMO for each part, thus verifying PCA. The validation was further carried out by employing the alpha (α) parameter method, suggested by Cronbach (1951). The α parameter measures how well the variables in a set are implicitly related and it is expressed as (Field 2005):

$$\alpha = \frac{N^2 \overline{\text{cov}}}{\sum s^2_{\text{var}} - \sum \text{cov}_{\text{var}}} \tag{5}$$

where N is the number of variables, $\overline{\text{cov}}$ is the average covariance between variables, s^2_{var} is the variable variance, and cov_{var} is the covariance. When data show a multidimensional structure, the parameter has a low value. A minimum acceptable value of α is 0.70. The computed α-values were 0.84, 0.82, and 0.87 for laboratory and field total load, and field suspended load, respectively, and thus further verified PCA.

4.1.6 Discussion

By feature selection Dogan (2008) reduced the number of parameters to 5 for laboratory total load $\left(\frac{u_m s}{w}, \frac{B}{d_{50}}, \frac{h}{d_{50}}, \frac{u_*}{w}, \frac{u_* d_{50}}{\nu}\right)$ of which $\frac{u_m s}{w}$ and $\frac{u_*}{w}$ were already suggested by Yang (1996), and 4 for field total load $\left(\frac{u_m s}{\sqrt{g(G_s-1)d_{50}}}, \frac{B}{d_{50}}, \frac{q}{u_* d_{50}}, \frac{\nu^2}{g(G_s-1)d^3_{50}}\right)$. Using the RVM method, Dogan et al. (2009) employed 4 parameters $(q_* \tau_* \tau_*' \tau_{*c})$. On the other hand, employing PCA this study reduced the number of parameters to 6 in the case of field total sediment load and to 5 in the cases of laboratory total and field suspended loads (see Table 1).

Examining the parameters employed by Dogan (2008) and Dogan et al. (2009), it is seen that our study obtained different parameters for laboratory total load and only two common parameters in the case of field total load, as shown in Table 1. The major difference is that Dogan (2008) used Reynolds number only in the laboratory total load case, while in our study Reynolds number factors in all the three cases, which is in accord with the importance of Reynolds number in sediment transport (Yalin 1977). Also, two parameters, $\frac{B}{d_{50}}$ and $\frac{h}{d_{50}}$ in Dogan (2008) merged into one parameter $\frac{R}{d_{50}}$ in our study. In addition, Froude number,

which was not employed by Dogan (2008) and Dogan et al. (2009), turned out to be an important parameter for field total load.

4.2 Artificial Neural Network

The Artificial Neural Network (ANN) is a massively parallel-distributed information-processing system that has certain performance characteristics resembling the biological neural network of the human brain (Haykin 1999). Identification of complex patterns is a specific property of ANN, which is commonly employed for solutions of nonlinear problems. ANNs are trained with a set of input and output data pairs, and tested for further analysis. There are numerous applications of ANNs in hydrology, hydraulics, and water recourses management (ASCE 2000; Tayfur 2002, 2012; Zhu et al. 2006; Tayfur and Guldal 2006; Tayfur et al. 2007, among others).

In this study, the feed forward back propagation algorithm was used to develop the sediment predictive model. In a feed forward network, the input variables provided into the input layer are multiplied by weights before reaching the hidden layer. The net information received by the hidden layer neurons are passed through an activation function to produce outputs which are, in turn, passed to the next layer as inputs. For complete details see Tayfur (2012).

The dimensionless parameters presented in Table 1 formed the input variables and sediment concentration (C) was the output variable for the constructed three layer ANN model, which had neurons between 5 and 10 in the hidden layer. The tangent hyperbolic transfer function between input and hidden layers, linear transfer function between hidden and output layers and the Levenberg-Marquardt algorithm for training were employed. The ANN model was created for each dataset. In each case, 70 % of dataset was used for training and 30 % for testing. The models were evaluated using the root mean square error (RMSE), the mean absolute relative error (MARE), and the correlation coefficient (R), as presented in Table 2 and Fig. 2.

4.3 Genetic Algorithm

The Genetic Algorithm (GA) is a nonlinear search and optimization method inspired by the biological processes of natural selection and the survival of the fittest (Tayfur and Moramarco 2008). They make relatively few assumptions and do not rely on any mathematical properties of the functions (Tayfur and Moramarco 2008). Bit, gene, chromosome, and gene pool are basic units of GA. In GA, bits create gene which is the model variable to be optimized. A collection of genes form chromosome which is a candidate for solution. Basic operations of GA are fitness evaluation, selection, cross-over, and mutation. By these operations, new generations (chromosomes) are obtained at each iteration. Complete details can be obtained from Tayfur (2012).

GA has been extensively applied in water resource engineering (Sen and Oztopal 2001; Guan and Aral 2005; Hilton and Culver 2005; Tayfur 2009; Tayfur et al. 2009, among

Table 2 Performance of model

		ANN			GA		
		R	RMSE(m^2/s) × 10^{-4}	MARE	R	RMSE (m^2/s) × 10^{-4}	MARE
Laboratory	total load	0.97	1.682	51.8	0.89	3.944	796
Field	suspended load	0.94	11.13	59.5	0.89	14.52	164.7
	total load	0.96	1.049	35.6	0.92	1.673	58.87

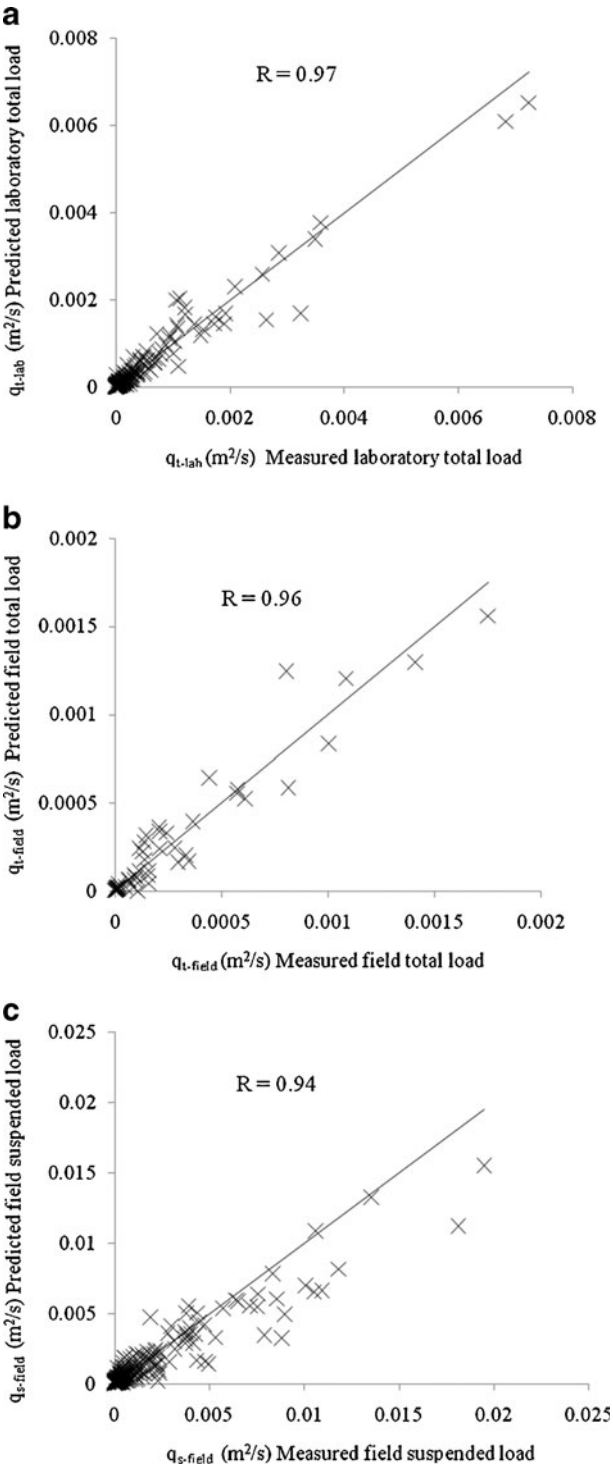

Fig. 2 Measured versus predicted ANN sediment load data (testing data) **a**) Field total load; **b**) Laboratory total load; **c**) Field suspended load

others). A few studies have applied GA in sediment transport studies. For example, Zhang et al. (2010) used GA to optimize the critical shear stress for deposition and re-suspension that are important and effective in sediment transport model. They concluded that GA can effectively improve the simulation results of a sediment transport model in coastal areas.

Sediment transport exhibits a nonlinear behavior. Hence, in this study a popular form of nonlinear equation $y = \alpha(x_1)^{\beta_1}(x_2)^{\beta_2}\ldots(x_n)^{\beta_n}$ was considered for the GA application where $(x_1, x_2, \ldots, x_n)$ constitutes the inputs, and y is the output. The dimensionless parameters in Table 1 were used as input variables, and volumetric sediment transportation rate was considered as output. For each dataset, the proposed nonlinear equations are as follows:

For laboratory total load:

$$C_{t-lab} = \alpha\left(\frac{u_* h}{\nu}\right)^{\beta_1}\left(\frac{\nu^2}{g(G_s-1)d_{50}^3}\right)^{\beta_2}\left(\frac{R}{d_{50}}\right)^{\beta_3}\left(\frac{q^2}{g(G_s-1)d_{50}^3}\right)^{\beta_4}\left(\frac{\rho_s u_*^2}{\gamma_s d_{50}}\right)^{\beta_5} \quad (6)$$

For field data total load:

$$C_{t-field} = \alpha\left(\frac{u_* d_{50}}{\nu}\right)^{\beta_1}\left(\frac{\nu^2}{g(G_s-1)d_{50}^3}\right)^{\beta_2}\left(\frac{R}{d_{50}}\right)^{\beta_3}\left(\frac{q^2}{g(G_s-1)d_{50}^3}\right)^{\beta_4}\left(\frac{\rho_s u_*^2}{\gamma_s d_{50}}\right)^{\beta_5}\left(\frac{u_m}{\sqrt{g(G_s-1)d_{50}}}\right)^{\beta_6} \quad (7)$$

Field data suspended load:

$$C_{s-field} = \alpha\left(\frac{u_* h}{\nu}\right)^{\beta_1}\left(\frac{u_m}{u_*}\right)^{\beta_2}\left(\frac{R}{d_{50}}\right)^{\beta_3}\left(\frac{q^2}{g(G_s-1)d_{50}^3}\right)^{\beta_4}\left(\frac{u_m}{\sqrt{g(G_s-1)d_{50}}}\right)^{\beta_5} \quad (8)$$

The GA model obtained the optimal values of parameters $(a, \beta_1, \beta_2, \ldots, \beta_6)$ in Eqs. 6 to 8. The model was calibrated and tested using 70 % and 30 % of each dataset, respectively. For the three nonlinear models, optimal model parameters were obtained by minimizing the objective function of mean absolute error (MAE). At the start, parameters were randomly assigned numbers. Due to the GA algorithm requirement, the user needs to search for the values of parameters in a pre-specified range. In this study GA searched α-values in the range [−1—1], and β_1, β_2,…,β_6 in the range [−5—5]. Another range could have been employed as well. Different ranges were attempted and the model in the end converged to the same optimal values.

Evolver GA Solver for Microsoft Excel (Palisade Corporation. Evolver 5.7 2010) was employed in this study. For minimization of the objective function, the Recipe Solving method, 80 % cross-over rate, 5 % mutation rate, 200 population size, and 50,000 iterations were employed. The value of the objective function was checked at each iteration to control the trend of the error. The optimal values of the parameters are shown in Table 3. The performance of models during testing is summarized in Table 2 and Fig. 3.

Table 3 Coefficients for GA-based models

		α	β_1	β_2	β_3	β_4	β_5	β_6
Laboratory	total load	0.2481	0.3438	0.0294	−0.6568	2.2669	0.1132	
Field	suspended	0.1077	0.8027	−0.4416	0.0156	−1.2617	3.8296	
	total load	0.00075	2.5047	0.2117	1.2405	−0.3637	0.7975	0.9561

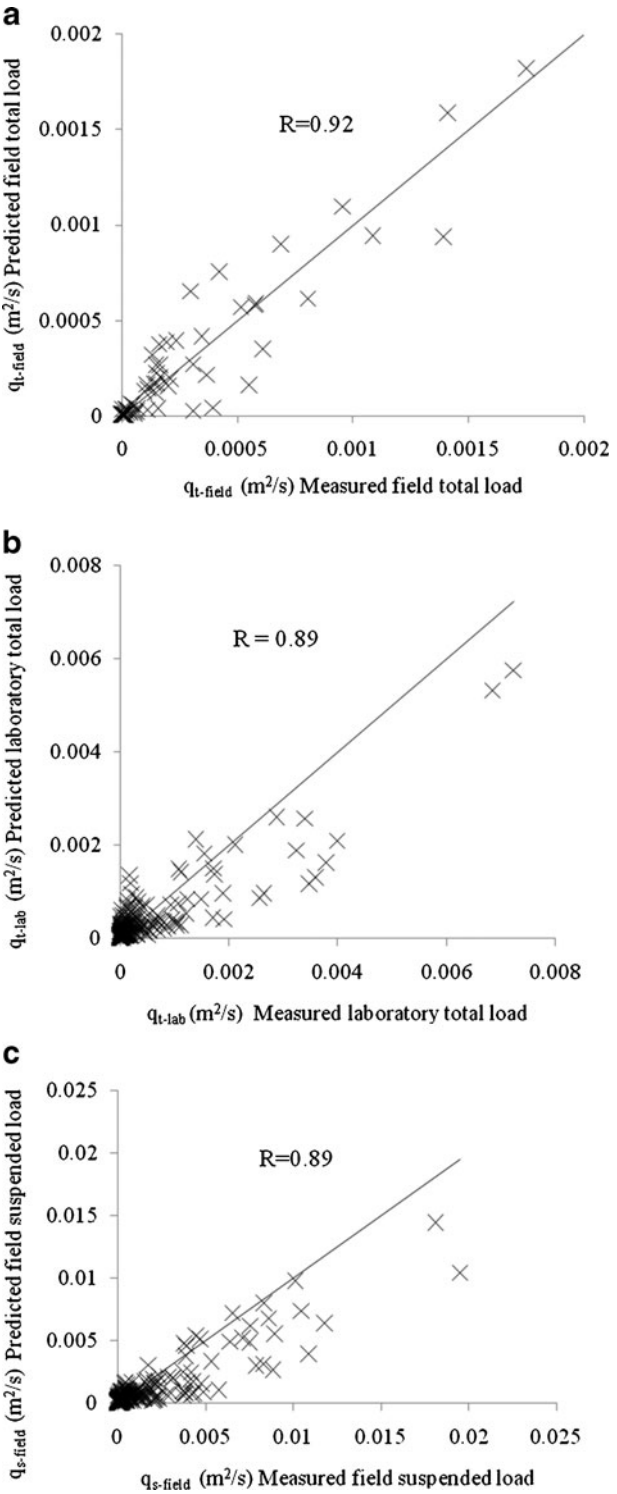

Fig. 3 Measured versus GA predicted sediment load data (testing data) **a**) Field total load; **b**) Laboratory total load; **c**) Field suspended load

5 Discussion of Results

5.1 Sediment Load Predictions

The performance of expert methods for predicting suspended and total loads for laboratory and field cases is summarized in Table 2 and Figs. 2 and 3. It is seen that ANN performed better than the other method. Figure 2 shows that the model predicted field and laboratory total load reasonable well. The measured-predicted data distribution closely followed the 1—1 line with minor deviations (Fig. 2a, b). The model however mostly underestimated field suspended loads (Fig. 2c). On the average, ANN produced a high R=0.95 and low RMSE = 4.62×10^{-4} m^2/s and MARE = 48 % (Table 2).

Figure 3 presents GA produced prediction results. Figure 3a shows that predicted versus measured data follows the 1—1 line with minor deviations, implying satisfactory performance for field total load. The 1—1 lines in Fig. 3b and c however show that the GA mostly underpredicted the measured data for laboratory total load and field suspended load. On average, GA produced a high R=0.90 and RMSE = 6.71×10^{-4} m^2/s (Table 2).

5.2 Transferability from Laboratory Scale to Field Scale

5.2.1 ANN Model

The variables obtained by PCA (see Table 1) for the laboratory total load formed the input vector of the ANN model. The trained model was then tested against field total load data. Figure 4a presents the prediction results and 1—1 line. The ANN model produced reasonable values of R=0.85, and RMSE = 2.438×10^{-4} m^2/s. The 1—1 line in Fig. 4a, however, shows that the model overall under-predicted the measured data.

5.2.2 GA Model

We obtained the optimal values of the parameters of Eq. 6 by the GA using laboratory total load data and presented the parameter values in Table 3. We then tested the GA-based equation against the field total load data. Figure 4b shows the model predicted results and 1—1 line. GA produced results similar to ANN (see Fig. 4a and b), with R=0.85, and RMSE = 5.468×10^{-4} m^2/s. The 1—1 line in Fig. 4b shows that, as opposed to ANN, GA overall over-predicted the measured field data.

6 Conclusions

Following conclusions are drawn from this study:

(1) The principal component analysis (PCA) is applied to identify the effective variables in sediment transport. The predictive models are developed, based upon the outcomes of PCA. Results show that PCA is beneficial in such studies.
(2) The ANN and GA methods better predict total loads than suspended loads.
(3) The ANN and GA methods are employed to investigate the transferability from laboratory to field scale for sediment transport. The transferability from laboratory to field scale can be carried out for the suspended loads, provided that there are sufficient data.

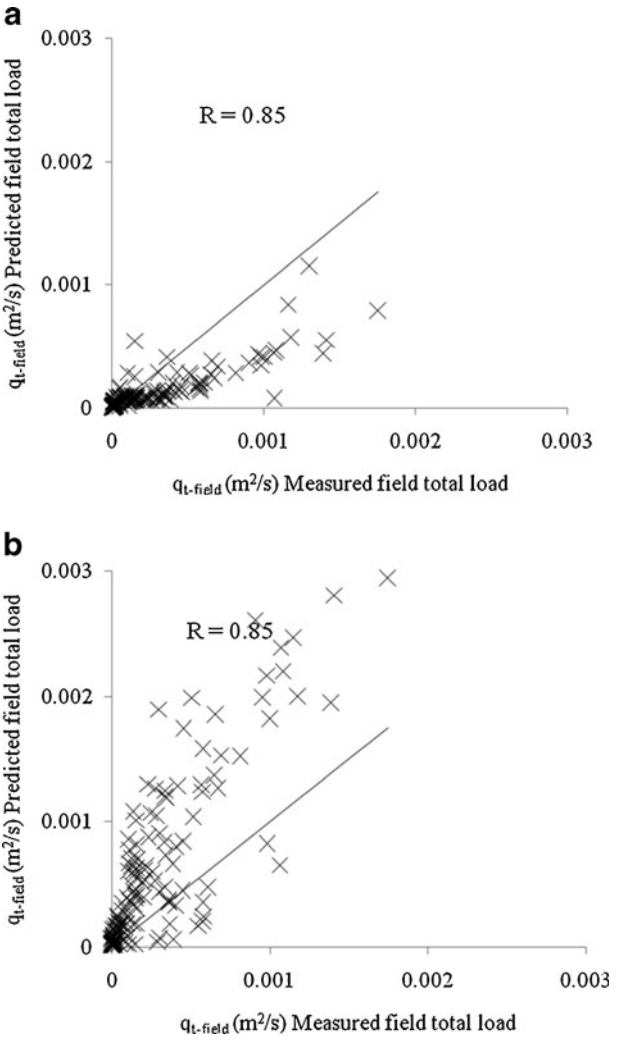

Fig. 4 Transferability of laboratory to field scale **a**) ANN and **b**) GA

(4) This study shows that these methods can be employed to predict field loads in ungauged basins which are common in under-developed and developing countries. Planning and operating hydraulic structures may require establishment and maintenance of gauging stations. Since such stations would be an economical burden in under-developed countries, the methods developed in this study can be utilized.

(5) The data used in this study are from natural channels in plains. Hence, the results presented in this study may not be applicable to mountains rivers. In such a case, the models may have to be re-calibrated and re-tested.

(6) As future work, the transferability can be also carried out for other modes of sediment transport, provided that there is sufficient data. This also implies that these methods are data-driven. Hence, limited data restricts their applicability.

Appendix I

Notation

$\frac{h}{d_{50}}$	dimensionless flow depth
$\frac{u_m h}{v}$	flow Reynolds number (average velocity)
$\frac{u_* d_{50}}{v}$	Reynolds number related to particle size
$\frac{u_* h}{v}$	flow Reynolds number (shear velocity)
$\tau_* = \frac{hs}{(G_s-1)d_{50}}$	dimensionless shear stress
$\frac{u_m}{u_*}$	friction factor
$\frac{B}{h}$	width to depth ratio
$\frac{q}{\sqrt{ghh}}$	Froud number
$\frac{q}{u_* d_{50}}$	dimensionless flow unit discharge
$\frac{B}{d_{50}}$	dimensionless width
$\frac{vu_*}{g(G_s-1)d_{50}^2}$	dimensionless shear velocity
$\frac{1}{d_*^3} = \frac{v^2}{g(G_s-1)d_{50}^3}$	dimensionless particle size
$\frac{q^2}{g(G_s-1)d_{50}^3}$	dimensionless flow unit discharge
$\frac{\rho_s u_*^2}{\gamma_s d_{50}}$	mobility number (related to particle size)
$\frac{u_m}{\sqrt{g(G_s-1)d_{50}}}$	Froude number (related to particle size)

References

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000) Artificial neural network in hydrology. I: preliminary concepts. J Hydrol Eng 5(2):115–123

Bhattacharya B, Price RK, Solomatine DP (2007) Machine learning approach to modeling sediment transport. J Hydraul Eng 133(4):440–450

Brownlie WR (1981) "Compilation of alluvial channel data: laboratory and field." W.M Keck laboratory of Hydraulics and Water Resources, Division of Engineering and Applied Science, California Institute of Technology, Pasadena, Calif

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334

Dogan E (2008) "Prediction of total sediment load in open channel with ANN." Ph.D. Thesis, Dept. of Civil Eng., Sakarya Uinv., Turkey (in Turkish)

Dogan E, Tripathi S, Lyn DA, Govindaraju RS (2009) From flume to rivers: can sediment transport in natural alluvial channels be predicted from observations at the laboratory scale? Water Resour Res 45(8):1–16

Field A (2005) Discovering statistics using SPSS. SAGE Publications

Guan J, Aral MM (2005) Remediation system design with multiple uncertain parameters using fuzzy sets and genetic algorithm. J Hydrol Eng 10(5):194–386

Haykin S (1999) Neural networks: a comprehensive foundation. Prentice-Hall, Englewood Cliffs

Hilton ABC, Culver TB (2005) Groundwater remediation design under uncertainty using genetic algorithms. J Water Resour Plan Manag 131(1):25–34

Jain KS (2001) Development of integrated sediment rating curves using ANNs. J Hydraul Eng 127(1):30–37

Loska K, Wiechula D (2003) Application of principal component analysis for the estimation of source of heavy metal contamination in surface sediments from the Rybnik Reservoir. J Chemosphere 51(8):723–733

Noori R, Khakpour A, Omidvar B, Farokhnia A (2010) Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic. J Eswa 37(8):5856–5862

Ouyang Y (2005) Evaluation of river water quality monitoring stations by principal component analysis. J Waters 39(12):2621–2635
Palisade Corporation. Evolver 5.7 (2010) The genetic algorithm solver for Microsoft Excels
Pett MA, Lackey NR, Sulivan JJ (2003) Making sense of factor analysis. Sage publications
Sen Z, Oztopal A (2001) Genetic algorithm for the classification and prediction of precipitation occurrence. Hydrol Sci J 46(2):255–268
Tayfur G (2002) Artificial neural networks for sheet sediment transport. Hydrol Sci J 47(6):879–892
Tayfur G (2009) GA-optimized model predicts dispersion coefficient in natural channels. J Hydrol Res 40(1):65–78
Tayfur G (2012) Soft computing in water resources engineering. WIT Press, Southampton
Tayfur G, Guldal V (2006) Artificial neural networks for estimating daily total suspended sediment in natural stream. J Nord Hydrol 37:69–79
Tayfur G, Moramarco T (2008) Predicting hourly-based flow discharge hydrographs from level data using genetic algorithms. J Hydrol 352(1–2):77–93
Tayfur G, Moramarco T, Singh VP (2007) Predicting and forecasting flow discharge at sites receiving significant lateral inflow. Hydrol Process 21(14):1848–1859
Tayfur G, Barbetta S, Moramarco T (2009) Genetic algorithm-based discharge estimation at sites receiving lateral inflows. J Hydrol Eng 14(5):463–474
Van Rijn LC (1984) Sediment transport. Part I: bed load transport. J Hydraul Eng 110(10):1431–1456
Winter TC, Mallory SE, Allen TR, Rosenberry DO (2000) The use of principal component analysis for interpreting ground water hydrographs. J Ground Water 38(2):234–246
Yalin MS (1977) Mechanics of sediment transport. Pergamon, Oxford
Yang CT (1996) Sediment transport: theory and practice. McGraw-Hill international edition
Zhang FX, Wai OWH, Jiang YW (2010) Prediction of sediment transportation in deep bay (Hong Kong) using genetic algorithm. J Hydrodyn 22(5):599–604
Zhu YM, Lu XX, Zhou Y (2006) Suspended sediment flux modeling with artificial neural network: an example of the Longchuanjiang River in the Upper Yangtze. J Geomorphol 84(1–2):111–125