Use of principal component analysis in conjunction with soft computing methods for investigating total sediment load transferability from laboratory to field scale

Gokmen Tayfur and Yashar Karimi

ABSTRACT

This study quantitatively investigates the generalization from laboratory scale to field scale using the soft computing (expert) and the empirical methods. Principal component analysis is utilized to form the input vector for the expert methods. Five main dimensionless parameters are used in the input vector of artificial neural networks (ANN), calibrated with laboratory data, to predict field total sediment loads. In addition, nonlinear equations are constructed based upon the same dimensionless parameters. The optimal values of the exponents and constants of the equations are obtained by the genetic algorithm (GA) method using the laboratory data. The performance of the so-developed ANN and GA based models are compared against the field data and those of the existing empirical methods, namely Bagnold, Ackers and White, and Van Rijn. The results show that ANN outperforms the empirical methods. The results also show that the expert models, calibrated with laboratory data, are capable of predicting field total loads and thus proving their transferability capability. The transferability is also investigated by a newly proposed equation which is based on the Bagnold approach. The optimal values of the coefficients of this equation are obtained by the GA. The performance of the proposed equation is found to be very efficient.

Key words | empirical methods, expert methods, laboratory and field scale, principal component analysis, total load, transferability

Gokmen Tayfur (corresponding author)
Department of Civil Engineering,
Izmir Institute of Technology,
Urla,
Izmir,
Turkey
E-mail: *gokmentayfur@iyte.edu.tr*

Yashar Karimi
Department of Civil Engineering,
EU,
Izmir,
Turkey

INTRODUCTION

Considerable modeling research has been devoted to sediment load predictions (Jain 2001; Tayfur 2002; Dogan *et al.* 2009, among many). Most of the existing models, one way or another, are based on the combination of several flow, sediment dynamics parameters and geometric characteristics of channels. Zhu *et al.* (2006) summarize the parameters used in several commonly employed models. Bhattacharya *et al.* (2007), using artificial neural networks (ANN), estimated sediment loads employing dimensionless parameters based mainly on studies of Yalin (1977) and Van Rijn (1984a). Bhattacharya *et al.* (2007) considered two scenarios by employing different sets of input variables to predict dimensionless total load transport rate. In their first scenario, they employed dimensional parameters of u (flow velocity), h (flow depth), D (particle diameter), and I (slope) and in their 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 loads, total loads, and bed loads for laboratory scale and field scale separately. They did not investigate the transferability from laboratory to field scale.

The details of the importance of the transferability are well documented in Dogan *et al.* (2009), who investigated it from laboratory scale to field scale using a RVM (relevance vector machine) method. They selected

doi: 10.2166/nh.2013.244

parameters based on empirical methods, considering the ones having similar statistical distribution in laboratory and field data. 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 here that, in their parameter selection process for the predictive model for the transferability, they considered both the laboratory and field data. Actually, they should have considered only the laboratory data, and therefore, they had introduced a bias into their model.

For the transferability study, the predictive model, in fact, should be constructed based solely upon laboratory data and this is exactly done in this study. In forming the input vector for the expert models developed in this study, the principal component analysis (PCA) is employed. Employing PCA for this purpose is very advantageous because while preserving the original information as much as possible, it squeezes a high-dimensional data matrix into a low-dimensional matrix in which the data variability is explained by a fewer number of variables (Palau *et al.* 2012). Furthermore, it achieves parsimony by explaining the maximum amount of common variance in a correlation matrix using the smallest number of explanatory concept and avoids problems of multicollinearity and singularity (Field 2005). There are applications of the PCA in the water resource engineering, hydrology, and environmental sciences (Winter *et al.* 2000; Loska & Wiechula 2003; Ouyang 2005; Noori *et al.* 2010).

This study investigates the transferability from laboratory to field scale using PCA, ANN, and genetic algorithm (GA) methods. Also, this study investigates the transferability by a newly proposed empirical equation, which is conceptually based on the Bagnold's approach. The coefficients of the proposed empirical equation are optimized by the GA.

DATA

Brownlie (1981) composed an extended set of laboratory and field data on flow discharge, channel width, flow depth, channel bed slope, mean particle diameter, gradation, specific gravity, sediment concentration, and flow temperature. The list and details of the data were provided therein. Uniform flow conditions in straight flumes were assumed for laboratory experiments.

The field data were compiled from different rivers in Pakistan, India, Japan, Colombia, and mostly in the USA. Different sampling methods had been employed by the researchers. Bed load measurements in some US rivers were made with a Helley–Smith sampler. In other cases, bed load was sampled by using a vortex trough in the stream bed which transported the bed load material into a sampling pit adjacent to the stream. In gravel-bed rivers, the transport rates were determined with basket-type bedload samplers. Sediment discharge was measured by trapping sediment in a mesh-covered hopper and pumping it into a weighing tank. As the mixture entered the weighing tank, the sediment settled to the bottom, while excess water was allowed to overflow.

The concentration measurements were made by means of depth-integrating samplers at hydraulic structures where sufficient turbulence was present to force the total load into suspension. In some rivers, sediment concentrations were measured with the aid of Delft bottle samplers which are designed so that water is allowed to pass through the sampler while sediment coarser than 0.05 mm is trapped. Sediment particle properties, such as median diameter and gradation, were obtained from the particle-size distributions.

Stream flow observations were made at gauging stations. By the measurements of flow velocity, flow depth, and the topographic surveying of cross-sections, flow discharge values were computed.

In line with Dogan *et al.* (2009), the following restrictions are carried out on the data employed in this study:

(1) B/h (where B is channel width and h is flow depth) is greater than 4 to avoid the sidewall effects.
(2) Relative roughness, R/d_{50} (where R is hydraulic radius and d_{50} is the mean particle diameter), is greater than 100 to avoid extreme shallow flow depth condition.
(3) Sediment size is the sand range of 0.062 (mm) $< d_{50} <$ 2.0 (mm).

(4) Geometric standard (σ_g) is less than 5 to avoid extreme amount of gravel or fine material.
(5) Sediment concentration (C) is greater than 10 ppm to avoid inaccuracy of low concentration measurement.

Under these restrictions, 1,190 total load records from laboratory experiments reported in Brownlie (1981) and 180 total load records from field measurements reported in Brownlie (1981) are retained in this study.

DIMENSIONLESS PARAMETERS

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

$$c = f(u^*, q, d_{50}, \rho, \rho_s, h, B, v, \sigma_g, S, u_m, \mu, g) \quad (1)$$

where c is sediment concentration (mg/L); $u_* =$ shear velocity (LT^{-1}), $q =$ unit flow discharge (L^2T^{-1}), $d_{50} =$ particle diameter such that 50% (median) of particle size by weight is finer (L), $\rho =$ water density (ML^{-3}), $\rho_s =$ sediment density (ML^{-3}), $h =$ flow depth (L), $B =$ channel width (L), $v =$ kinematic viscosity (L^2T^{-1}), $\sigma_g =$ sediment gradation, $S =$ slope, $u_m =$ average flow velocity (LT^{-1}), $\mu =$ dynamic viscosity ($ML^{-1}T^{-1}$), $g =$ gravitational acceleration (LT^{-2}).

Dogan (2008), performing a dimensional analysis using the Buckingham's Pi theorem, first obtained 10 dimensionless parameters and then added eight more from the literature. In addition, R/d_{50} dimensionless hydraulic radius is proposed in this study in order to reflect the effects of channel cross-section, flow depth, and wetted perimeter by a single parameter. Equation (2) summarizes all 19 dimensionless parameters.

Table 1 | Extracted component and loading coefficients for laboratory total load

	$\frac{u_* h}{v}$	$\frac{v^2}{g(G_s-1)d_{50}^3}$	$\frac{R}{d_{50}}$	$\frac{q^2}{g(G_s-1)d_{50}^3}$	$\frac{\rho_s u_*^2}{\gamma_s d_{50}}$
PC_1	0.058	0.953	0.867	0.865	0.324
PC_2	0.929	−0.34	0.357	0.379	0.775

$\frac{u_* h}{v}$: Reynolds number related to shear stress, $\frac{v^2}{g(G_s-1)d_{50}^3}$: 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).

where C is sediment concentration (ppm) and R is hydraulic radius. Equation (2) is commonly employed in the literature (Brownlie 1983; Nagy *et al.* 2002; Dogan *et al.* 2009; Bisantino *et al.* 2010, among many).

The PCA is used to analyze the data related to the parameters in Equation (2) for the total load. Table 1 summarizes the resulting optimal dimensionless parameters, whose definitions are given in Appendix I (available online at http://www.iwaponline.com/nh/045/144.pdf). The predictive expert models are constructed based upon these dimensionless parameters where sediment concentration (C) is the output variable.

METHODS

Principal component analysis

Field (2005) explains the aim of application of PCA as follows:

> '*Factor analysis (and PCA) is a technique for identifying groups or clusters of variables. This technique has three main uses: (1) to understand structure of a set of variables, (2) to construct a questionnaire to measure an underlying variable, (3) to reduce data to more manageable size while retaining as much of the original information as possible.*'

$$C = f\left(\frac{h}{d_{50}}, \frac{\rho}{\rho_s}, \frac{u_m h}{v}, \frac{u_* d_{50}}{v}, \frac{u_* h}{v}, \frac{hS}{(G_s-1)d_{50}}, \frac{u_m}{u_*}, \frac{B}{h}, \frac{q}{\sqrt{ghh}}, \frac{q}{u_* d_{50}}, \frac{B}{d_{50}}, \frac{vu_*}{g(G_s-1)d_{50}^2}, \frac{v^2}{g(G_s-1)d_{50}^3}, \frac{q^2}{g(G_s-1)d_{50}^3}, \frac{\rho_s u_*^2}{\gamma_s d_{50}}, \frac{u_m}{\sqrt{g(G_s-1)d_{50}}}, S, \sigma_g, \frac{R}{d_{50}}\right) \quad (2)$$

As such, PCA is a technique for recognizing groups of variables and it is useful for reducing the number of data sets to optimal size while preserving the original information as much as possible. By reducing a data set from a group of interrelated variables into a smaller set of variables, the PCA achieves parsimony by explaining the maximum amount of common variance in a correlation matrix using the smallest number of explanatory concepts (Field 2005). For this study, that means, the PCA simplifies the original set of data records related to the dimensionless parameters in Equation (2), synthesizing the most significant information into a statistical model that is able to explain most of the behavior of the sediment transport.

PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (PCs). The number of PCs is less than or equal to the number of original variables. PCs are generated in a sequential ordered manner with decreasing contributions to the variance, i.e., the first PC explains most of the variations present in the original data, and successive PCs account for decreasing proportions of the variance (Mahapatra *et al.* 2012). The generated set of PCs presents uncorrelated linear combinations of the original variables and accounts for the total variance of the original data. Note that all the PCs are generated in such a way that they are orthogonal to each other, i.e., the correlation between them is zero (Mahapatra *et al.* 2012).

Mathematically, the PCs are linear combinations of independent variables, and they can be shown as (Field 2005):

$$\mathrm{PC}_i = b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon_i \tag{3}$$

where PC_i is ith principal component. $X_1, X_2, \ldots X_n$ are independent variables, which are loaded on ith principal component. $b_1, b_2, \ldots b_n$ are ith principal component loading coefficients, presenting the relative contribution of each variable (Field 2005), and ε_i is residual.

Finding an optimal number of PCs is a concern in a PCA model. This is because reducing space dimensionality in excess may cause a significant loss of information. On the other hand, extracting too many PCs can lead to an overfitting of the model, losing its reliability and predictive capability. It is essential to extract the right number of PCs so that the system behavior can be satisfactorily explained (Palau *et al.* 2012). In general, the extraction of new PCs is terminated when adding a new variable does not significantly improve the explanatory behavior of the variable (Palau *et al.* 2012).

Before the PCA application, one has to control the 'sample size quality' and 'data screening', as presented below.

Sample size quality

The reliability of PCA strictly depends on the sample size which is important due to the generalization of model results from laboratory to field scale. Additionally, fluctuation of correlation coefficient from sample to sample, particularly significant in small size samples, affects the PCA. Field (2005) classified sample size 100 as a poor, 300 as good, and 1,000 as an excellent case. In our study, 1,190 records of data set are excellent to perform PCA.

We also carried out the KMO (Kaiser-Meyer-Olkin) criterion to check the adequacy of the sample sizes. The KMO criterion is a quantity of sampling adequacy that is expressed as (Pett *et al.* 2003):

$$\mathrm{KMO} = \frac{\Sigma(\mathrm{correlation})^2}{\Sigma(\mathrm{correlation})^2 + \Sigma(\mathrm{partial\ correlation})^2} \tag{4}$$

The KMO criterion varies between 0 and 1. The partial correlation represents how much of the variance is independent of the other variables in the data set, i.e., dependent on variables not contained in the data set. If the partial correlation is 0, then KMO criterion is 1, implying that the variables are measuring a common component, or vice versa. According to Field (2005), for the PCA, the minimum value of KMO criterion is 0.5. This criterion is satisfied for all the samples.

Data screening

The data screening is carried out to avoid problems of multicollinearity (variables that are very highly correlated, $R > 0.90$) and singularity (variables that are perfectly correlated, $R \sim 1$) in the input variables. In other words, by data screening, one eliminates highly and perfectly correlated variables. In order to avoid the multicollinearity and singularity problem in the analysis, the variables should be inspected at the beginning. The correlation matrix (R-matrix) can have

useful information about the multicollinearity. The multicollinearity is determined by the determinant of the matrix which should be greater than 1×10^{-5}.

In this study, the dimensionless variables are subjected to the data screening before the PCA application. As a result, due to the multicollinearity and singularity problem,

$$\frac{h}{d_{50}}, \frac{u_m h}{v}, \frac{hs}{(G_s-1)d_{50}}, \frac{q}{u_* d_{50}}, \frac{B}{d_{50}}, \frac{vu_*}{g(G_s-1)d_{50}^2}$$

are eliminated. After this elimination, the determinant of R-matrix is achieved as 2.73×10^{-5}.

After these pre-procedures, we are now ready to initiate the PCA, as presented below.

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. As such, due to the communality check, B/h is eliminated.

Thus, so far, seven parameters were eliminated from 19 parameters in Equation (2). In the following section, the remaining 12 dimensionless parameter data values are subjected to PC analysis whereby the number of PCs and the important parameters are decided.

Component rotation

Note that each PC (in Equation (3)) represents a cluster. There should be low similarities among samples that are associated with different clusters and high similarities among samples strongly associated with the same cluster (Mahapatra *et al*. 2012). Factor loadings ($b_1, b_2, \ldots b_n$ in Equation (3)) reflect the degree of association between each PC and the sample. The factor loadings of each member of data set on the PCs are taken into account to cluster samples into the appropriate group. The number of clusters is decided on the basis of percentage variation explained by the PCs (Mahapatra *et al*. 2012).

It is customary to use the rotation method to transform PCs to simpler and more interpretable constructs. After rotation, each variable will be related to one of the PCs and each PC will have high correlation with only a small set of variables (Mahapatra *et al*. 2012).

Figure 1 schematically presents component rotation for the case of two components. Before the rotation, the perpendicular solid lines in Figure 1 are the PCs. The components can be visualized as axis and variables can be plotted on it (the solid triangles in Figure 1). Once plotted, it may be possible to calculate to what degree variables load on to these components. Generally, variables load highly on the most important component, and load slightly on the other component. This can be seen in Figure 1 where, before the rotation, the variables highly load on PC_1. Due to this characteristic, interpretation and discrimination between components can be difficult. In such a case, the rotation technique is employed (see Figure 1). After the rotation, the perpendicular dashed lines in Figure 1 are now the PCs where some variables load on PC_1 and some on PC_2. By this technique, the importance of each variable in each component can be clearly seen. In this study, we exactly followed this viewpoint and selected important variables by considering b_i values in each component.

The application of the PCA on the data employed in this study, following the procedure outlined above, resulted in two PCs for laboratory total load, which explain 85% of variation ($PC_1 = 52\%$ and $PC_2 = 33\%$). Table 1

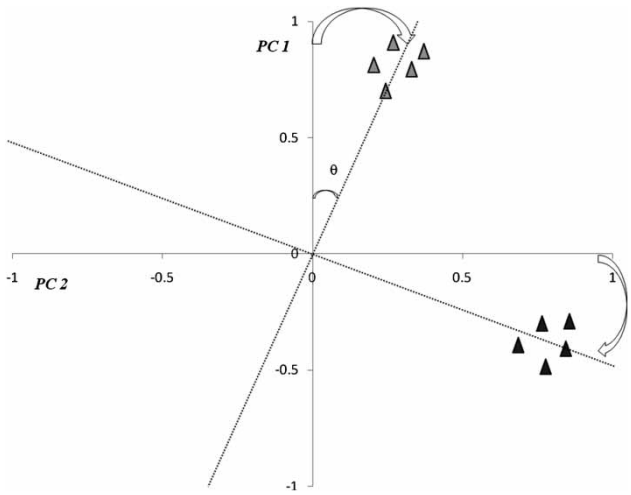

Figure 1 | Illustration of component rotation.

summarizes the number of PCs and loading factor values for each variable. As seen, the dimensionless parameters $\frac{v^2}{g(G_s-1)d_{50}^3}$, $\frac{R}{d_{50}}$, $\frac{q^2}{g(G_s-1)d_{50}^3}$ highly load on PC_1 with 0.953, 0.867, 0.865 while $u_*h/v, \rho_S u_*^2/\gamma_s d_{50}$ load on PC_2 with 0.929, 0.775 loading factors, respectively. In summary, these two PCs explain 85% of the information of the whole original data sets and therefore five parameters loaded on these PCs form the input vector for the expert models (ANN, GA) to predict total sediment loads.

It is worth noting that, for our purpose in this study, the variables which are clustered on the components are important rather than the number of components. Furthermore, in this study, we used the clustered variables as the model inputs rather than the PCs. This is because PCs, as shown by Equation (3), are a linear combination of the variables whereas the sediment transportation has a nonlinear behavior. Some studies use PC_s directly as model inputs (Noori *et al.* 2010). In this study, however, we employed the dimensionless parameters, which were loaded in PC_s, as the input vectors for the predictive models.

Validation of PCA

In order to validate the findings from the PCA, we conducted the split-half-sample method which randomly divides the whole sample into two parts and applies the PCA to each part. In the end, it satisfied communalities, component loading, and KMO criterion for each part, thus verifying the PCA. We further tested this validation by employing the alpha parameter method suggested by Cronbach (1951). The α-parameter measures how well a set of variables are implicitly related and it is expressed as (Field 2005):

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

where N is number of variables, $\overline{\text{cov}}$ is average covariance between variables, s_{var}^2 and cov_{var} are variable variance and covariance, respectively. When data show multidimensional structure, α-parameter has a low value. Minimum acceptable value for α is 0.70. The computed α-value in this study is 0.84 thus re-verifying the PCA.

Discussion

Dogan (2008), by feature selection, reduced the number of parameters to five for laboratory total load $\left(\frac{u_m S}{w}, \frac{B}{d_{50}}, \frac{h}{d_{50}}, \frac{u_*}{w}, \frac{u_* d_{50}}{v}\right)$ of which $u_m S/w$ and u_*/w had already been suggested by Yang (1996). Dogan *et al.* (2009), by the RVM method, employed four parameters (q_*, τ_*, τ_*', τ_{*c}). This study, on the other hand, by the PCA, obtained five dimensionless parameters in the case of laboratory total load (see Table 1).

When one examines the parameters employed by Dogan (2008) and Dogan *et al.* (2009) and the ones presented in Table 1, it can be seen that this study obtained different parameters for laboratory total load. Also, two parameters, $B/d_{50}, h/d_{50}$ in Dogan (2008) merged as R/d_{50} in our study.

Artificial neural network

ANN is a massively parallel-distributed information-processing system that has certain performance characteristics resembling biological neural networks of the human brain. Identification of complex patterns is a specific property of ANN, which is commonly employed in solutions of nonlinear problems. ANN are trained with a set of input and output data pairs, and tested for further analysis. There are numerous applications of ANN in hydrology, hydraulics, and water recourse management (ASCE 2000; Tayfur *et al.* 2007; Tayfur 2012, among many).

In this study, the feed forward back propagation algorithm is used to establish 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 hidden layer neurons are passed through an activation function to produce outputs which are, in turn, passed to the next layer as inputs. The details are presented elsewhere (Tayfur 2012).

The dimensionless parameters presented in Table 1 formed the input variables and the sediment concentration (C) was the output variable for the constructed three-layer ANN model, which had neurons in between five and 10 in the hidden layer. Tangent hyperbolic transfer function between input and hidden layers, linear transfer function between hidden and output layers, and Levenberg-Marquardt

as training algorithm were employed. Seventy percent of the laboratory data set is used for the training and 30% of the laboratory data for the testing. The performance of the model was 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 Figure 2(a).

Genetic algorithm

The GA is a nonlinear search and optimization method inspired by the biological processes of natural selection and the survival of the fittest (Tayfur 2012). They make relatively few assumptions and do not rely on any mathematical properties of the functions (Tayfur 2012). Bit, gene, chromosome, and gene pool are basic units of GA. In GA, bits create a gene which is the model variable to be optimized. A collection of genes form a 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. The details can be obtained elsewhere (Tayfur 2012).

The GA has extensive application in water resource engineering (Sen & Oztopal 2001; Tayfur 2012, among many). 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 models. They concluded that GA can effectively improve the simulation result of a sediment transport model in coastal areas.

Sediment transport, as is well known, exhibits nonlinear behavior. Hence, in this study, a popular form of nonlinear equation $y = \alpha(x_1)^{\beta_1}(x_2)^{\beta_2}\cdots(x_n)^{\beta_n}$ is considered for the GA application where $x_1, x_2, \ldots, x_n$ constitute inputs, α is coefficient, $\beta_1, \beta_2, \ldots, \beta_n$ are exponents, and y is output. The

Table 2 | Performance of models for laboratory total load data

	R	RMSE (m^2/h)	MARE
ANN	0.97	0.60	51.8
GA	0.89	1.56	175.0
Van Rijn	0.55	4.54	145.5
Ackers and White	0.65	4.20	66.0
Bagnold	0.93	2.79	179.0

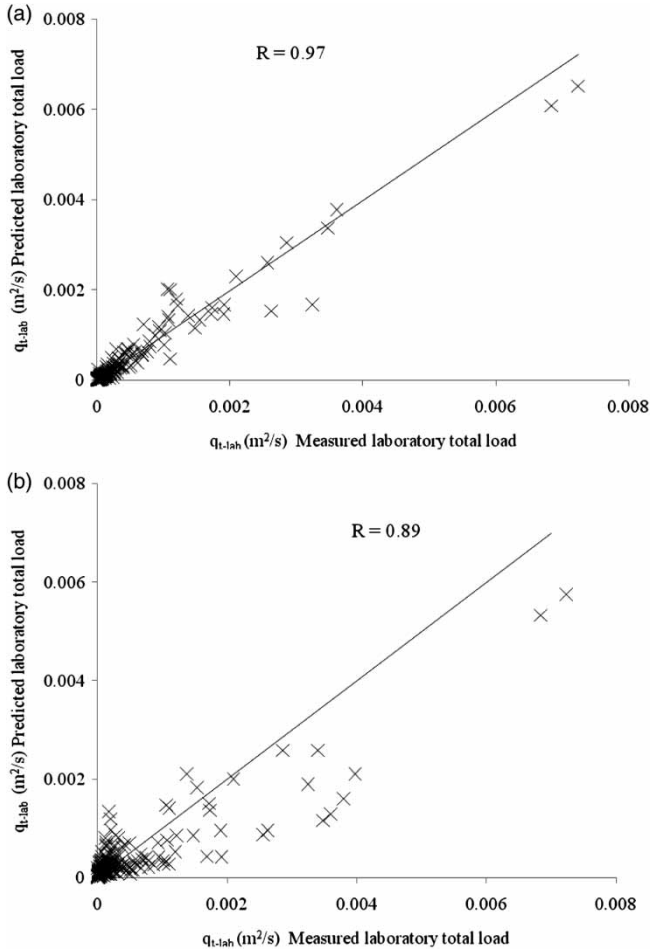

Figure 2 | Measured versus predicted sediment load data (testing data): (a) ANN, (b) GA.

dimensionless parameters in Table 1 are used as input variables, and volumetric sediment transportation rate is considered as output. The proposed nonlinear equation for laboratory total load is expressed as follows:

$$C_{t-\mathrm{lab}} = \alpha\left(\frac{u_* h}{v}\right)^{\beta_1}\left(\frac{v^2}{g(G_s-1)d_{50}^3}\right)^{\beta 2}\left(\frac{R}{d_{50}}\right)^{\beta 3}\left(\frac{q^2}{g(G_s-1)d_{50}^3}\right)^{\beta 4}\left(\frac{\rho_s u_*^2}{\gamma_s d_{50}}\right)^{\beta 5} \quad (6)$$

The GA model obtains the optimal values of the model parameters ($a, \beta_1, \beta_2, \ldots, \beta_5$) in Equation (6). The model calibration and testing for the laboratory data were performed by using 70 and 30% of each data set, respectively. For the nonlinear model, optimal model parameters were obtained

by minimizing the objective function of mean absolute error. At the start, parameters were randomly assigned numbers. The user, due to the GA algorithm requirement, needs to search the values of parameters in a pre-specified range. GA searched α-values in [−1 to +1], β_1, β_2, … , β_5 in [−5 to +5] in this study. Another range could have been employed as well. We tried different ranges and the model in the end converges to the same optimal values.

Evolver GA Solver for Microsoft Excel (Palisade Corporation 2010) was employed in this study. In minimization, the objective function, the Recipe Solving method, 80% crossover rate, 5% mutation rate, 200 population size, and 50,000 iterations were employed. The value of the objective function is checked at each iteration to control the trend of the error. The optimal values of the parameters are presented in Table 3. The performance of the model for the testing case is summarized in Table 2 and Figure 2(b).

Empirical methods

Extensive studies have been carried out for the determination of sediment transport in alluvial channels. In the literature, there are many empirical sediment predictive methods, which are mainly developed using laboratory flume experimental data. They are however used for the estimation of field sediment loads, despite the fact that the applicability and accuracy of laboratory data to field conditions is still controversial. In this study, Bagnold, Ackers and White, and Van Rijn empirical methods are used for the comparative analysis. These methods are briefly summarized in Appendix II (available online at http://www.iwaponline.com/nh/045/144.pdf) and details can be obtained from the literature, including Yang (1996). The results of the empirical methods for the laboratory data are presented in Table 2.

Discussion of results

The performance of the expert and the empirical methods for the laboratory data are summarized in Table 2. As seen, ANN performs better than the other methods. 1–1 line in Figure 2(a) is also presented. According to Figure 2(a), the model predicts the total load reasonably well. The measured–predicted data distribution closely follows the 1–1 line with minor deviation (Figure 2(a)). ANN produced, for the results presented in Figure 2(a), high $R = 0.97$ and low RMSE = 0.60 m^2/h and MARE = 51.8% (Table 2).

Figure 2(b) presents the prediction results that the GA produced. The 1–1 line in Figure 2(b) shows that the GA mostly underpredicts the measured data. For the results presented in Figure 2(b), GA produced $R = 0.89$, RMSE = 1.56 m^2/h, and MARE = 175% (Table 2).

The empirical methods tested here, on the other hand, showed poor performance (Table 2), compared to ANN. Among them, the Bagnold method produced better results with $R = 0.93$, RMSE = 2.79 m^2/h, and MARE = 179% (Table 2), as good as the GA. Bagnold was followed by Ackers and White, with $R = 0.65$, RMSE = 4.20 m^2/h, and MARE = 66% (Table 2). Van Rijn shows a poor performance, with R = 0.55, RMSE = 4.54 m^2/h, and MARE = 145.5% (Table 2).

Table 3 | Coefficients for GA-based model

α	β_1	β_2	β_3	β_4	β_5
0.248	0.344	0.029	−0.657	2.267	0.113

GENERALIZATION FROM LABORATORY SCALE TO FIELD SCALE

ANN model

The variables obtained by the PCA (see Table 1) for the laboratory total load formed the input vector of the ANN model. The trained model was then tested against the field total load data. Figure 3(a) presents the prediction results and 1–1 line.

GA model

We obtained the optimal values of the parameters of Equation (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 3(b) shows the model predicted results and 1–1 line.

Proposed empirical method

According to Bagnold (1966), the total load and transport of bed material particles can be achieved by summation of the

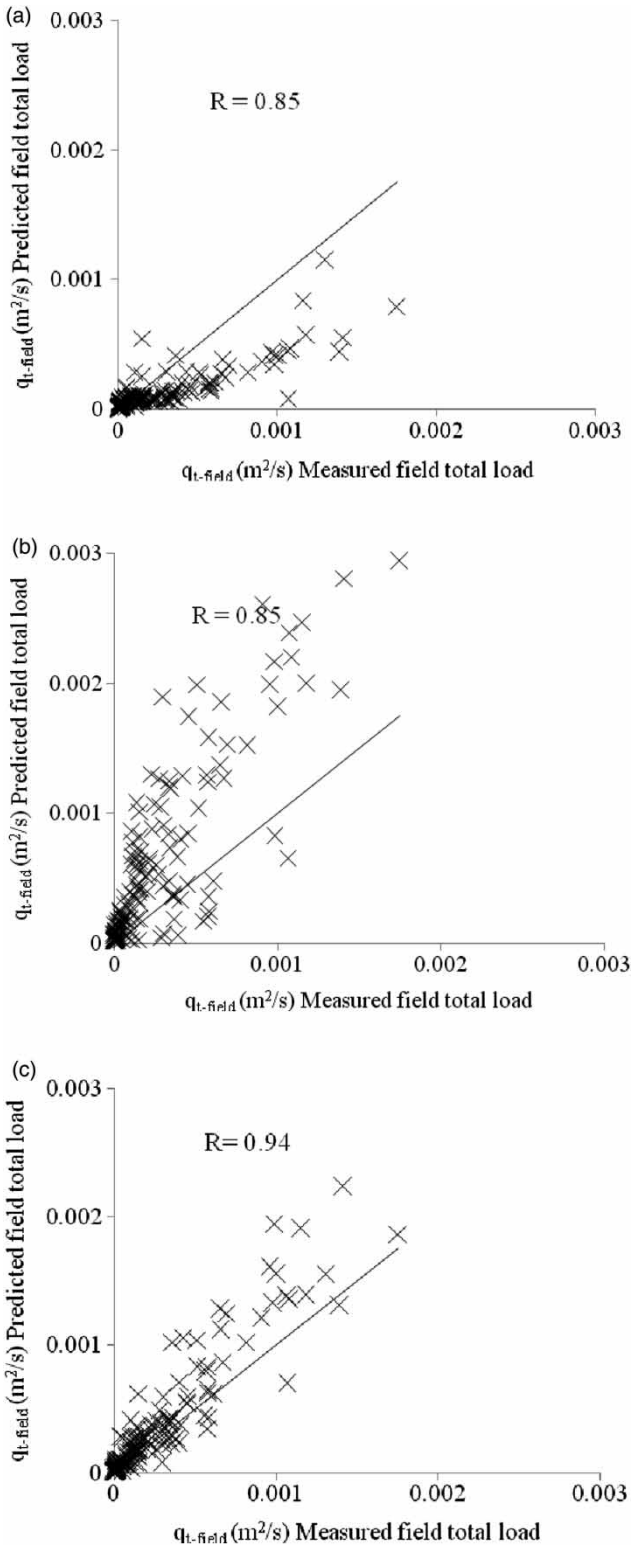

Figure 3 | Transferability of laboratory to field scale: (a) ANN, (b) GA, (c) Equation (8).

bed load and suspended load. The Bagnold equation for total load is given in Appendix II. As seen, the method uses five variables (γ, γ_s, τ, $\bar{u}$ w_s) and three coefficients ($tan\alpha$, e_b, 0.01). This study, however, for the transferability, proposed a new equation which is considered to be simpler and more compact by using three variables (τ_b, $\bar{u}$, w_s) and three coefficients (α, β_1, β_2), as presented by Equation (7).

$$q_t = \frac{\alpha(\tau_b \bar{u})^{\beta_1}}{(w_s/\bar{u})^{\beta_2}} \tag{7}$$

where τ_b = overall bed shear stress ($ML^{-1}T^{-2}$), $\bar{u}$ = depth-averaged velocity (LT^{-1}), w_s = fall velocity of sediment (LT^{-1}), and α, β_1 and β_2 are the coefficients.

The optimal values of the coefficients of the proposed equation are obtained by GA. The transferability of this method was investigated for total load. The coefficients were optimized by the GA method employing the laboratory total load data. The so-obtained optimal values are $\alpha = 0.0156$, $\beta_1 = 1$, and $\beta_2 = 0.659$. The method was then tested against field total data (Figure 3(c)). Thus, the proposed equation is as follows:

$$q_t = 0.0156 \frac{(\tau_b \bar{u})}{(w_s/\bar{u})^{0.659}} \tag{8}$$

Note that the transferability of the empirical methods cannot be easily performed. This may be because they are very complicated (see Appendix II).

Discussion of results

Figure 3 and Table 4 present the transferability results. Figure 3(a) shows the results for the ANN model. The

Table 4 | Performance of models for field total load data

	R	RMSE (m²/h)	MARE (%)
ANN	0.85	0.88	44.2
GA	0.85	1.07	83.7
Equation (8)	0.94	0.72	37.8
Bagnold	0.86	5.28	80.0

ANN model in Figure 3(a) produced reasonable values of $R = 0.85$ and RMSE $= 0.88$ m^2/h. The 1–1 line in Figure 3(a), however, implies that the model overall underpredicted the measured data.

Figure 3(b) shows results for the GA model (Equation (6)). GA produced similar results as ANN (see Figures 3(a) and 3(b)), with $R = 0.85$ and RMSE $= 1.97$ m^2/h. The 1–1 line in Figure 3(b) shows that, as opposed to ANN, GA overall overpredicted the measured field data.

Figure 3(c) presents the results for the GA-based Bagnold method (Equation (8)). As seen in Figure 3(c), it produced satisfactory results with the highest $R = 0.94$ and lowest RMSE $= 0.72$ m^2/h. The 1–1 line in Figure 3(c) shows that measured versus predicted data followed the line closely, implying that the model did not, overall, under- and overpredict the measured data. It fairly captured the measured field data, including the low and high values.

Table 4 also presents the error measures for the predictions of the field sediment total loads by the Bagnold method, given in Appendix II. As seen in Table 4, although the Bagnold method performs almost as well as the expert methods, the proposed Equation (8) outperforms all of them.

CONCLUSIONS

This study employed laboratory and field total load data, compiled from the literature by Brownlie (1981) to investigate performance of expert (ANN, GA) and empirical (Bagnold, Ackers and White, and Van Rijn) methods for predicting total loads. Following the restrictions to avoid extreme shallow flow depth conditions, side wall effects, extreme amount of gravel and/or fine material, and inaccuracies in low concentration measurements, 1,190 laboratory total load and 180 field total load records were used.

The number of dimensionless parameters which formed the input vector for the expert methods were obtained using PCA which involved several operations such as sample size quality, data screening, communality, and component rotation. Five dimensionless parameters (Reynolds number related to shear stress, dimensionless particle size, dimensionless hydraulic radius, dimensionless unit flow discharge, and mobility number related to particle size) formed the input variables for the expert methods.

The expert methods were first trained (calibrated) using 70% of the laboratory data and then applied to predict the remaing 30% of the laboratory total load data. The performance of these models were tested aginst the empirical methods for the laboratory data. Then, the generalization capability of the expert methods were investigated. For this purpose, the models were trained using only laboratory data and then tested against the field total load data.

This study also proposed an empirical formula based on Bagnold's concept for the generalization purpose. The coefficient of the proposed formula was found by the GA using only the laboratory data. The performance of the proposed formula was tested against the field loads as well as those of the expert methods.

The following conclusions are drawn from this study:

(1) The PCA was applied, for the first time, to identify the effective variables in sediment transport. The predictive models were created based upon the outcomes of the PCA. The results proved that the PCA is beneficial in such studies.
(2) The ANN outperformed the empirical methods in predicting the laboratory total loads.
(3) GA and Bagnold methods showed comparable performance in predicting the laboratory total loads, outperforming the other empirical methods.
(4) The ANN and GA methods were employed to investigate the transferability from laboratory to field scale for sediment transport. ANN and GA can be calibrated with laboratory sediment data and then applied to predict field sediment data.
(5) The transferability was investigated using the proposed Equation (8). It produced satisfactory results. It performed better than the ANN, GA, and Bagnold methods. Hence, it can be employed for predicting field total sediment loads.
(6) The implication of this study is that these procedures can be employed to predict field loads in ungauged basins which are common in underdeveloped and developing countries. Planning and operating hydraulic structures may require establishment and maintenance of gauging stations. Since such stations would bring about an

economic burden especially in underdeveloped countries, the methods developed in this study can be utilized.

(7) The field data used in this study are from natural channels. Hence, the results presented in this study may not be applicable to mountain rivers. In such a case, the models may have to be recalibrated and retested.

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

REFERENCES

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 2000 Artificial neural network in hydrology. II: hydrologic application. *J. Hydrol. Eng.* **5** (2), 124–137.

Bagnold, R. A. 1966 An approach to the sediment transport problem from general physics. U.S. Geological Survey Professional Paper 422-J.

Bhattacharya, B., Price, R. K. & Solomatine, D. P. 2007 Machine learning approach to modeling sediment transport. *J. Hydraul. Eng.* **133** (4), 440–450.

Bisantino, T., Gentile, F., Milella, P. & Liuzzi, G. T. 2010 Effect of time scale on the performance of different sediment transport formulas in a semiarid region. *J. Hydraul. Eng.* **136** (1), 56–61.

Brownlie, W. R. 1981 Compilation of alluvial channel data: Laboratory and field. Report No. KH-R-43B, W. M. Keck Laboratory of Hydraulics and Water Resources, Division of Engineering and Applied Science, California Institute of Technology, Pasadena, California. November 1981.

Brownlie, W. R. 1983 Flow depth in sand-bed channels. *J. Hydraul. Eng.* **109** (7), 959–990.

Cronbach, L. J. 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. PhD Thesis, Department of Civil Engineering, Sakarya University, Sakarya, Turkey (in Turkish).

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

Field, A. 2005 *Discovering Statistics Using SPSS*. SAGE Publications, London.

Jain, K. S. 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. *Chemosphere* **51** (8), 723–733.

Mahapatra, S. S., Sahu, M., Pater, R. K. & Panda, B. N. 2012 Prediction of water quality using principle component analysis. *Water Qual. Expo. Health* **4**, 93–104.

Nagy, H. M., Watanabe, B. & Hirano, M. 2002 Prediction of sediment load concentration in rivers using artificial neural network model. *J. Hydraul. Eng.* **128** (6), 588–595.

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. *Exp. Syst. Appl.* **37** (8), 5856–5862.

Ouyang, Y. 2005 Evaluation of river water quality monitoring stations by principal component analysis. *Water Res.* **39** (12), 2621–2635.

Palau, C. V., Arrequi, F. J. & Carlos, M. 2012 Burst detection in water networks using principal component analysis. *J. Water Resour. Plan. Manage.* **138** (1), 47–54.

Palisade Corporation 2010 Evolver, The Genetic Algorithm Solver for Microsoft Excel. West Drayton, UK.

Pett, M. A., Lackey, N. R. & Sulivan, J. J. 2003 *Making Sense of Factor Analysis*. SAGE Publications, London.

Sen, Z. & Oztopal, A. 2001 Genetic algorithm for the classification and prediction of precipitation occurrence. *J. Hydrol. Sci.* **46** (2), 255–268.

Tayfur, G. 2002 Artificial neural networks for sheet sediment transport. *J. Hydrol. Sci.* **47** (6), 879–892.

Tayfur, G. 2012 *Soft Computing in Water Resources Engineering*. WIT Press, Southampton, UK.

Tayfur, G., Moramarco, T. & Singh, V. P. 2007 Predicting and forecasting flow discharge at sites receiving significant lateral inflow. *Hydrol. Process.* **21** (14), 1848–1859.

Van Rijn, L. C. 1984a Sediment transport. Part I: bed load transport. *J. Hydraul. Eng.* **110** (10), 1431–1456.

Winter, T. C., Mallory, S. E., Allen, T. R. & Rosenberry, D. O. 2000 The use of principal component analysis for interpreting ground water hydrographs. *J. Ground Water* **38** (2), 234–246.

Yalin, M. S. 1977 *Mechanics of Sediment Transport*. Pergamon, Oxford, UK.

Yang, C. T. 1996 *Sediment Transport: Theory and Practice*. McGraw-Hill, New York.

Zhang, F. X., Wai, O. W. H. & Jiang, Y. W. 2010 Prediction of sediment transportation in deep bay (Hong Kong) using genetic algorithm. *J. Hydrodyn.* **22** (5), 599–604.

Zhu, Y. M., Lu, X. X. & Zhou, Y. 2006 Suspended sediment flux modeling with artificial neural network: an example of the Longchuanjiang River in the Upper Yangtze. *J. Geomorph.* **84** (1–2), 111–125.

First received 12 September 2012; accepted in revised form 15 August 2013. Available online 8 October 2013