Modelling of runoff and sediment yield using ANN, LS-SVR, REPTree and M5 models

Birendra Bharti, Ashish Pandey, S. K. Tripathi and Dheeraj Kumar

ABSTRACT

In this study, the performance evaluation of five machine learning models, namely, ANNLM, ANNSCG, least square-support vector regression (LS-SVR), reduced error pruning tree (REPTree) and M5, was carried out for predicting runoff and sediment in the Pokhariya watershed, India using hydro-meteorological variables as input. The input variables were selected using the trial-and-error procedure which represents the hydrological process in the watershed. The seven input variables to all the models comprised a combination of rainfall, average temperature, relative humidity, pan evaporation, sunshine duration, solar radiation and wind speed. The monthly runoff and sediment yield data were used to calibrate and validate all models for the years 2000 to 2008. Evaluation of models' performances were carried out using four statistical indices, i.e., Nash–Sutcliffe coefficient (NSE), coefficient of determination (R^2), percent bias (PBIAS) and RMSE-observations standard deviation ratio (RSR). Comparative analysis showed that the ANNLM model marginally outperformed the LS-SVR model and all the other models investigated during calibration and validation for runoff modelling whereas the LS-SVR model surpassed the artificial neural networks (ANN) model and other models for sediment yield modelling. Moreover, M5 model tree is better in simulating sediment yield and runoff than its near counterpart, the REPTree model, and marginally inferior when compared to LS-SVR and ANN models.

Key words | ANN, M5 model, machine learning technique, REPTree, runoff, sediment yield

Birendra Bharti (corresponding author)
Ashish Pandey
S. K. Tripathi
Department of Water Resources Development and Management,
IIT Roorkee,
Roorkee, Haridwar, 247667,
India
E-mail: *birendrabharti@gmail.com*

Dheeraj Kumar
Civil Engineering Department,
IIT Roorkee,
Roorkee, Haridwar, 247667,
India

INTRODUCTION

Soil erosion poses a serious problem for sustainable agriculture and the environment. It has become extremely serious due to inadequate consideration of nature's bearing capacity. Extensive soil erosion and its associated problems have already degraded the land and water resources of the world. The involvement of many, often, interrelated climatic and physiographic factors makes the rainfall–sediment process not only very complex to understand but also extremely difficult to simulate (Zhang & Govindaraju 2003). Precise sediment load simulation is a fundamental for sustainable water resources and environmental systems, as it plays a major role for any water availability decision-making process. The use of data-driven modelling techniques to deliver improved sediment yield rating curves has received considerable interest in recent years (Mount *et al.* 2012; Kim *et al.* 2017). Previously, hydrology researchers have developed several sediment prediction models, extending from empirical, i.e., USLE/RUSLE (Jain & Kothyari 2000) and mathematical, i.e., kinematic/diffusion wave theory (Naik *et al.* 2009) or linear programming optimization (Sarma *et al.* 2015) to physically based. The physically process-based models such as SWAT (Arnold & Allen 1999; Tripathi *et al.* 2003; Prabhanjan *et al.* 2014; Cohen Liechti *et al.* 2014a, 2014b), WEPP (Laflen *et al.*

doi: 10.2166/nh.2017.153

1991; Pandey *et al.* 2008) and many others have demonstrated better understanding in modelling sediment yield, but their need for data is often very high and even intensively monitored watersheds lack sufficient input data for these models. Therefore, it is necessary to develop substitutes for physically based models to simulate sediment yield using the available data.

More recently, techniques based on artificial neural networks (ANN) have been applied to the sedimentation engineering field (Jain 2001; Licznar & Nearing 2003; Cigizoglu 2004; Partal & Cigizoglu 2008; Cobaner *et al.* 2009; Feidas 2010; Marquez & Guevara-Pérez 2010; Kumar *et al.* 2014, 2016). However, the two drawbacks of ANNs are that the architecture has to be determined *a priori* or modified while training and ways of regularization are quite limited. Unfortunately, neural networks can get stuck in local minima while training (Smola 1996).

The support vector algorithm is a nonlinear generalization of the generalized portrait algorithm developed by Vapnik, in the early 1960s (Vapnik & Lerner 1963; Vapnik & Chervonenkis 1964). Support vector machine (SVM) utilizes the structural risk minimization principle of upper bound to the generalization error instead of minimizing the training error, which has been shown to be superior to the empirical risk minimization principle employed by ANN (Jain 2012). In hydrological studies, SVMs have been used successfully by various scientists (e.g., Sivapragasam & Muttil 2005; Khan & Coulibaly 2006; Çimen 2008; Wu *et al.* 2008; Misra *et al.* 2009; Kisi & Cimen 2011, 2012; Jain 2012; Goyal *et al.* 2014; Kumar *et al.* 2016) and this was the motivation for application in the present study to examine whether they lead to improved results. Application of SVMs to regression problems is known as support vector regression (SVR).

Recently, another machine learning technique that is gaining more attention in the hydrological community is the decision tree model, in particular, reduced error pruning tree (REPTree), which was introduced by Breiman *et al.* (1984) and is the simplest form of a decision tree (Sauquet & Catalogne 2011; Bachmair & Weiler 2012; Galelli & Castelletti 2013; Kumar *et al.* 2016). The ability to dodge finding potentially complicated parametric functions and the interpretation of tree structure as a cascade of 'if-then' rules between combinations of inputs and the output gives a better insight into the model internal structure and underlying physical processes (Iorgulescu & Beven 2004; Wei & Watkins 2011; Galelli & Castelletti 2013). The weakness of REPTree model in predicting the output is that the output is composed of discrete values and a piecewise constant function is used to reconstruct the output. The other disadvantage is comparison of all the possible combinations of input values to select the best performing partition which makes computational requirements grow rapidly with the input space dimensionality (Hyafil & Rivest 1976). The disadvantages of REPTree model were replaced by the M5 model tree which was first introduced by Quinlan (1992). The usage of averaging the tree leaves in the REPTree model was replaced by fitting a linear regression function to the data and obtaining a continuous representation of the output in the M5 model tree. Despite these clear advantages and the straightforward usage of the M5 algorithm, its use in water resource management is rather limited. Model trees have been applied in rainfall–runoff modelling (Solomatine & Dulal 2003), flood forecasting (Solomatine & Xue 2004; Singh *et al.* 2010; Kumar *et al.* 2016), statistical downscaling (Goyal & Ojha 2012; Goyal *et al.* 2012) and also the modelling of rating curves (Bhattacharya & Solomatine 2005).

In this study, an attempt has been made to investigate the performance of five machine learning models for runoff and sediment yield modelling, namely, ANNLM, ANNSCG, least square-SVR (LS-SVR), REPTree and M5 models to be used in water resources planning and management and to formulate better water management policies and remediate at local level.

MATERIAL AND METHODS

Study area

Keeping in view the objective and availability of hydrological and meteorological data, a small watershed named Pokhariya (approximately 7,168 ha) located in the Hazaribagh district of Jharkhand State which lies within the Damodar Barakar catchment has been selected as the study area for the present research work. The Pokhariya watershed lies between 86°00′ to 86°20′E longitude and 24°8′ to 24°14′N latitude (Figure 1). The topography of the study area is undulating with an elevation ranging from 229 to 388 m above

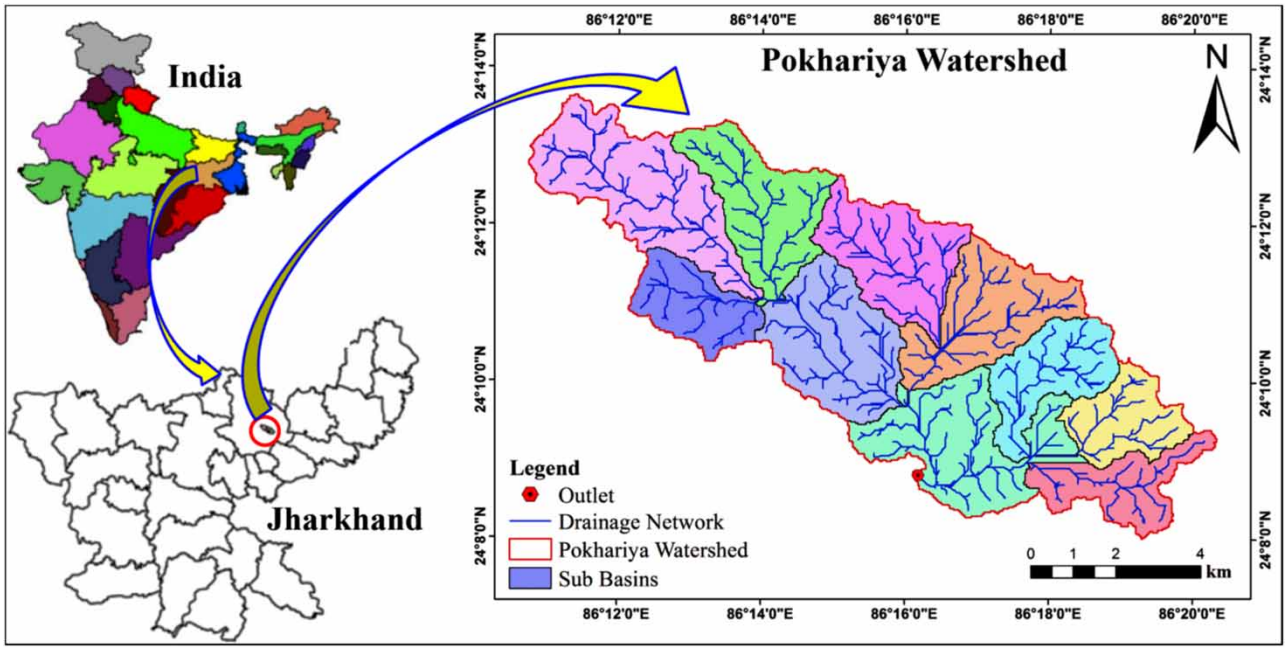

Figure 1 | Location map of the Pokhariya.

MSL. Two cropping seasons, Kharif (monsoon season) extending from June to September and Rabi (non-monsoon season) extending from October to January, are mainly followed. The Soil Conservation Department of Damodar Valley Corporation, Hazaribagh and Indo-German Bilateral Project (IGBP) on 'Watershed Management', New Delhi, India monitor the hydrological data in some of the watersheds of Damodar Valley, the Pokhariya watershed being one of them. The hydro-meteorological data, namely, rainfall, temperature, relative humidity, pan evaporation, sunshine hours, solar radiation and wind speed were collected from the concerned department and used in the prediction of runoff and sediment yield.

Input vector selection and normalization

In this study, input variables are selected by the trial-and-error method (Fernando & Jayawardena 1998) in identifying the appropriate input vector that best represents the hydrological process in the watershed. Based on the appraisal of trial-and-error procedure, the following input vectors and their combination were selected for runoff and sediment yield modelling employing a machine learning approach:

Combination 1:

$$S_t \text{ or } D_t = f(R_t) \quad (1)$$

Combination 2:

$$S_t \text{ or } D_t = f(R_t, T_t) \quad (2)$$

Combination 3:

$$S_t \text{ or } D_t = f(R_t, T_t, RH_t) \quad (3)$$

Combination 4:

$$S_t \text{ or } D_t = f(R_t, T_t, RH_t, E_t) \quad (4)$$

Combination 5:

$$S_t \text{ or } D_t = f(R_t, T_t, RH_t, E_t, \text{Sun}_t) \quad (5)$$

Combination 6:

$$S_t \text{ or } D_t = f(R_t, T_t, RH_t, E_t, \text{Sun}_t, \text{Solar}_t) \quad (6)$$

Combination 7:

$$S_t \text{ or } D_t = f(R_t, T_t, RH_t, E_t, \text{Sun}_t, \text{Solar}_t, W_t) \quad (7)$$

where S_t is sediment yield (t/ha) at time t, D_t is runoff (m^3/sec), R_t rainfall values (mm), T_t is average temperature (°C), RH_t is relative humidity (%), E_t is pan evaporation (mm), Sun_t is sunshine duration (hrs), Solar_t is solar radiation (MJ/m^2) and W_t is wind speed (m/sec).

Prior to calibration of all the machine learning models investigated in the present study, standardization/normalization procedure was applied to all the datasets. The main goal of data standardization/normalization is to scale the data within a certain range to minimize bias and to ensure that they receive equal attention within the neural networks (Maier & Dandy 2000). In the study, the following formula has been employed for the min–max normalization method:

$$\xi_{\text{norm}} = \frac{\xi_{\text{ori}} - \xi_{\text{min}}}{\xi_{\text{max}} - \xi_{\text{min}}} \quad (8)$$

where ξ_{norm} and ξ_{ori} represent the normalized and original data. ξ_{max} and ξ_{min} represent the maximum and minimum values among original data. Hydro-meteorological datasets were normalized within a range of 0.1–0.9.

ANN – an overview

An ANN is a black box model that has been applied in several diverse hydrological problems, the results of which have been encouraging. The important characteristics of ANNs include their adaptive nature and learning by examples (Obradovic & Deco 1996; Mattera & Haykin 1999). ANNs have been applied in hydrological studies for rainfall–runoff modelling (French *et al.* 1992; Shamseldin 1997; Anmala *et al.* 2000; Agarwal & Singh 2004; Chiang *et al.* 2004; Lin & Chen 2004; De Vos & Rientjes 2005), flood forecasting (Fernando & Jayawardena 1998; Babel *et al.* 2017), groundwater modelling (Yang *et al.* 1997; Krishna *et al.* 2008; Gorgij *et al.* 2017) and sediment yield estimation (Agarwal *et al.* 2005; Raghuwanshi *et al.* 2006).

The Levenberg–Marquardt (LM) back propagation neural network is a confidence neighbourhood-based method with a hyper-spherical confidence region. The LM algorithm uses this approximation to the Hessian matrix in the following Newton-like update:

$$X_{k+1} = X_k - [J_k^T J + \mu I]^{-1} J^T e \quad (9)$$

where J is the Jacobian matrix which contains first derivatives of the network errors, e is the vector of network errors and I is the identity matrix. Details of multi-layer feed forward network and its theorem can be found in Hornik *et al.* (1989).

Scaled conjugate gradient (SCG) method is performed along conjugate directions, which produces generally faster convergence than steepest descent directions. In steepest descent search, a new direction is perpendicular to the old direction (Hagan *et al.* 1996). The general method to determine the new search direction is to combine the new steepest descent direction with the previous search direction that are conjugated as governed by the subsequent equations:

$$\omega_{k+1} = \omega_k + \alpha_k p_k \quad (10)$$

$$p_k = E'(\omega) + \alpha_k p_{k+1} \quad (11)$$

where p_k and p_{k+1} are the conjugate directions in successive iterations. α_k and ω_k are calculated in each iteration. SCG needs to calculate the Hessian matrix which is approximated by:

$$E''_{(\omega k)P_k} = \frac{E'(\omega_k + \sigma_k p_k) - E'(\omega_k)}{\sigma_k} + \lambda_k p_k \quad (12)$$

where E' and E'' are the first and second derivative of E'', p_k, σ_k and λ_k are the search direction, parameter controlling the second derivation approximation and parameter regulating indefiniteness of the Hessian matrix. Considering the machine precision, the value of σ should be as small as possible ($\sigma \leq 10^{-4}$) (Møller 1993).

Support vector regression

The idea of SVMs, which are known as the classification and regression procedures, was developed by Vapnik (1995). LS-SVM is a least square version of SVM, and are a set of related supervised learning methods that analyse

data and recognize patterns, and which are used for classification and regression analysis. In this method, the solution can be found by solving a set of linear equations instead of a convex quadratic programming (QP) problem for classical SVMs. LS-SVM classifiers were proposed by Suykens & Vandewalle (1999). The goal is to construct a linear function, which represents the dependence of the output y on the input x and can be represented as:

$$y = \mathbf{w}^T \phi(x) + b \tag{13}$$

where $\mathbf{w}$ is known as weight vector and b as bias. $\phi(x)$ represents the nonlinear transformation function defined to convert a nonlinear problem to a linear problem.

The training phase of the learning machine involves adjusting the parameter $\mathbf{w}$. The parameters are estimated by minimizing the cost function $J(\mathbf{w}, e)$. The LS-SVM optimization problem for function estimation is formulated by minimizing the cost function:

$$J(\mathbf{w}, e) = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \gamma\tfrac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{14}$$

Subject to the equality constraint

$$y_i = \mathbf{w}^T\varphi(x_i) + b + e_i \quad i = 1, \ldots, N$$

where e_i is the random error and γ is a positive real constant. The first and second term of the cost function represent weight decay function and penalty function. The objective is to find the optimal parameters that minimize the prediction error of the regression model. The solution of the optimization problem is obtained by the Lagrangian function as:

$$L(w, b, e; \alpha) = J(\mathbf{w}, e) - \sum_{i=1}^{N} \alpha_i \{\mathbf{w}^T\varphi(x_i) + b + e_i - y_i\} \tag{15}$$

where α_i are Lagrange multipliers and b is the bias term. Differentiating Equation (15) with respect to w, b, e_i and α_i, i.e.,

$$\frac{\partial L}{\partial w} = 0 \rightarrow \mathbf{w} = \sum_{i=1}^{N} \alpha_i \varphi(x_i) \tag{16}$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0 \tag{17}$$

$$\frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i, \quad i = 1, \ldots, N \tag{18}$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow \mathbf{w}^T\varphi(x_i) + b + e_i - y_i = 0, \quad i = 1, \ldots, N \tag{19}$$

From Equations (16)–(19), $\mathbf{w}$ and e can be eliminated which will yield a linear system instead of a QP problem. Replacing $\mathbf{w}$ in Equation (13) from Equation (16), the kernel matrix may be obtained from application of Mercer's theorem.

$$K(x, x_i) = \varphi(x_i)^T\varphi(x) \tag{20}$$

where $\phi(x)$ represents the nonlinear transformation function defined to convert a nonlinear problem to a linear problem.

Thus, the resulting LS-SVR model can be expressed as:

$$\hat{y} = \sum_{i=1}^{N} \hat{\alpha}_i K(x, x_i) + \hat{b} \tag{21}$$

where $K(x, x_i)$ is a kernel function, $\hat{\alpha}_i$ and $\hat{b}$ are the estimated values of α_i and b which can be obtained by solving the linear system.

The nonlinear radial basis function (RBF) kernel which demonstrated more favourable performance than the other kernel functions and has been suggested in many studies (Dibike *et al.* 2001; Liong & Sivapragasam 2002; Choy & Chan 2003; Han & Cluckie 2004; Yu & Liong 2007) is defined as:

$$K(x, x_i) = \exp\left(-\frac{1}{2\sigma^2}\|x - x_i\|^2\right) \tag{22}$$

where σ is the kernel function parameter of the RBF kernel. In the context of sediment prediction, x_i is the new vector hydro-climatic input, based on which, sediment prediction $\hat{y}_i$ is made. The model performances are assessed by comparing the observed sediment yield (y_i) and computed sediment yield ($\hat{y}_i$).

Due to being more compact and able to shorten the computational training process and improve the generalization performance of LS-SVR (Suykens & Vandewalle 1999), the RBF kernel has been selected in this study.

Reduced error pruning tree

REPTree algorithm is a fast-learning method. It builds a decision/regression tree by using information gain/variance and also prunes it by using reduced error with back-fitting. The algorithm only considers values for numeric attributes once. It is primarily a method of constructing a set of decision rules on the predictor variables (Breiman *et al.* 1984; Verbyla 1987; Clark & Pregibon 1992).

M5 model tree

Tree-based regression models were also studied within the machine learning community. One of the contributions of the work carried out within this community is the possibility of using different models in the leaves of the trees (Quinlan 1992; Torgo 1997). Quinlan (1992) pioneered techniques for dealing with continuous-class learning problems by introducing 'model trees' and the M5 learning algorithm. They have a conventional decision tree structure, but the leaves consist of linear functions instead of discrete class labels.

During model prediction, a smoothing procedure can be applied to compensate the discontinuities between adjacent linear models. This process uses the leaf model to compute the predicted value, and then filters that value along the path back to the root, smoothing it at each node by combining it with the value predicted by the linear model for that node. The procedure is described in Quinlan (1992). The standard deviation reduction is computed by Quinlan (1992) as:

$$SDR = \mathrm{sd}(T) - \sum_i \frac{|T_i|}{|T|} \times \mathrm{sd}(T_i) \tag{23}$$

where T are sets of instances that reach the node, 'sd' represents standard deviation and T_i are the sets resulting from splitting the node according to a given attribute and split value.

Model training

The whole dataset was divided into two sets, i.e., training and validation of all the machine learning techniques discussed above. The monthly data of rainfall, average temperature, relative humidity, pan evaporation, sunshine, solar radiation, wind speed, runoff and sediment yield from 2000 to 2008 (36 months) were considered for the training and validation of the models. Out of the total 36 months, 70% of the data, i.e., 24 months, were selected for the training and the remaining 20% of data, i.e., 12 months, were considered for the validation of the models.

Statistical evaluation indices for various models

The entire data were divided into two parts on the basis of statistical properties of the time series, such as mean and standard deviation, one for calibration (training) and another for validation. The performance of all models during calibration and validation were evaluated by various performance indices, i.e., coefficient of determination (R^2) (Willmott 1981; Legates & McCabe 1999), Nash–Sutcliffe coefficient (NSE) (Nash & Sutcliffe 1970), percent bias (PBIAS) (Gupta *et al.* 1999) and RMSE-observations standard deviation ratio (RSR). The weightage is given to statistical measures of NSE followed by coefficient of correlation (CC) and RMSE-observations standard deviation ratio (RSR) (Singh *et al.* 2004) for evaluation of the model performance during validation. Model validation is possibly the most important step in the model building sequence. The efficacy of the model is ascertained not by its performance on the training dataset but by its ability to perform well on unseen data.

RESULTS AND DISCUSSION

As stated earlier, five models, i.e., ANNLM, ANNSCG, LS-SVR, REPTree and M5 model, have been developed for runoff and sediment yield modelling of Pokhariya watershed in India.

Results of monthly runoff simulations

Artificial neural network-Levenberg–Marquardt

Performance of the ANN models, which consists of a three-layer feed forward-back propagation network, trained by LM algorithms were evaluated for runoff with the seven input combination developed, resulting in seven ANN-LM models (denoted as ANNLMD) and presented in Table 1. The number of neurons in the hidden layer were assumed to be constant and five neurons in the hidden layer were considered. From Table 1, it can be observed, that the model ANNLMD5 performed well compared to other models during calibration (NSE = 0.98, RSR = 0.07, PBIAS = 1.14, $R^2 = 0.99$) and validation (NSE = 0.98, RSR = 0.11, PBIAS = −8.77, $R^2 = 0.98$) whereas the model ANNLMD4 performed worst during calibration (NSE = 0.65, RSR = 0.49, PBIAS = 3.58, $R^2 = 0.70$) and validation (NSE = −0.50, RSR = 0.98, PBIAS = −85.2, $R^2 = 0.50$). Scatter graphs between observed and simulated runoff during calibration and validation for all the ANNLMD models are presented in Figure 2(a) and 2(b). From the graphs, it can be observed that the model converges well during calibration and tends to capture almost all the values. This is due to the well-known characteristics of the LM algorithm of quicker process learning and convergence.

Artificial neural network-scaled conjugate gradient

Performance of the ANN models, which consists of a three-layer feed forward-back propagation network, trained by SCG algorithms were evaluated for runoff with the seven input combination developed, resulting in seven ANN-SCG models (denoted as ANNSCGD) and presented in Table 2. The number of neurons in the hidden layer were assumed to be constant and five neurons in the hidden layer were considered. From Table 2, it can be observed, that the model ANNSCGD7 performed well compared to other models during calibration (NSE = 0.43, RSR = 0.72, PBIAS = 12.72, $R^2 = 0.45$) and validation (NSE = 0.27, RSR = 0.89, PBIAS = −83.29, $R^2 = 0.70$) whereas the model ANNSCGD4 performed worst during calibration (NSE = 0.30, RSR = 0.70, PBIAS = 12.32, $R2 = 0.32$) and validation (NSE = −0.63, RSR = 1.21, PBIAS = −125.47, $R^2 = 0.51$). Scatter graphs between observed and simulated runoff during calibration and validation for all the ANNSCGD models are presented in Figure 3(a) and 3(b).

Least square-support vector regression

The LS-SVR model has two parameters (γ, σ) to be determined. In this study, the regularization parameter (γ), which determines the trade-off between the training error minimization and smoothness of the estimated function, was calibrated to see the effect of regularization parameter on runoff, keeping the RBF kernel function parameter (σ^2) constant with numerical value 0.9. The regularization parameter (γ) was calibrated during model development and varied from 1 to 10. The model developed employing the LS-SVR for runoff prediction was denoted as 'LSSVRD'. The model-generated runoff was compared against the available observed runoff at the outlet of Pokhariya watershed.

Table 1 | ANN model result by LM algorithm during calibration and validation

		Calibration				Validation			
Model number	ANN structure	NSE	RSR	PBIAS	R^2	NSE	RSR	PBIAS	R^2
ANNLMD1	1-5-1	0.69	0.48	5.23	0.71	0.76	0.53	−34.39	0.85
ANNLMD2	2-5-1	0.72	0.48	16.11	0.76	0.8	0.43	−36.61	0.88
ANNLMD3	3-5-1	0.7	0.47	−1.96	0.71	0.37	0.69	−70.38	0.72
ANNLMD4	4-5-1	0.65	0.49	3.58	0.68	−0.5	0.98	−85.2	0.45
ANNLMD5	5-5-1	0.98	0.07	1.14	0.98	0.98	0.11	−8.77	0.98
ANNLMD6	6-5-1	0.75	0.34	1.76	0.75	0.65	0.59	−39.22	0.81
ANNLMD7	7-5-1	0.87	0.23	5.89	0.87	0.53	0.49	−49.39	0.75

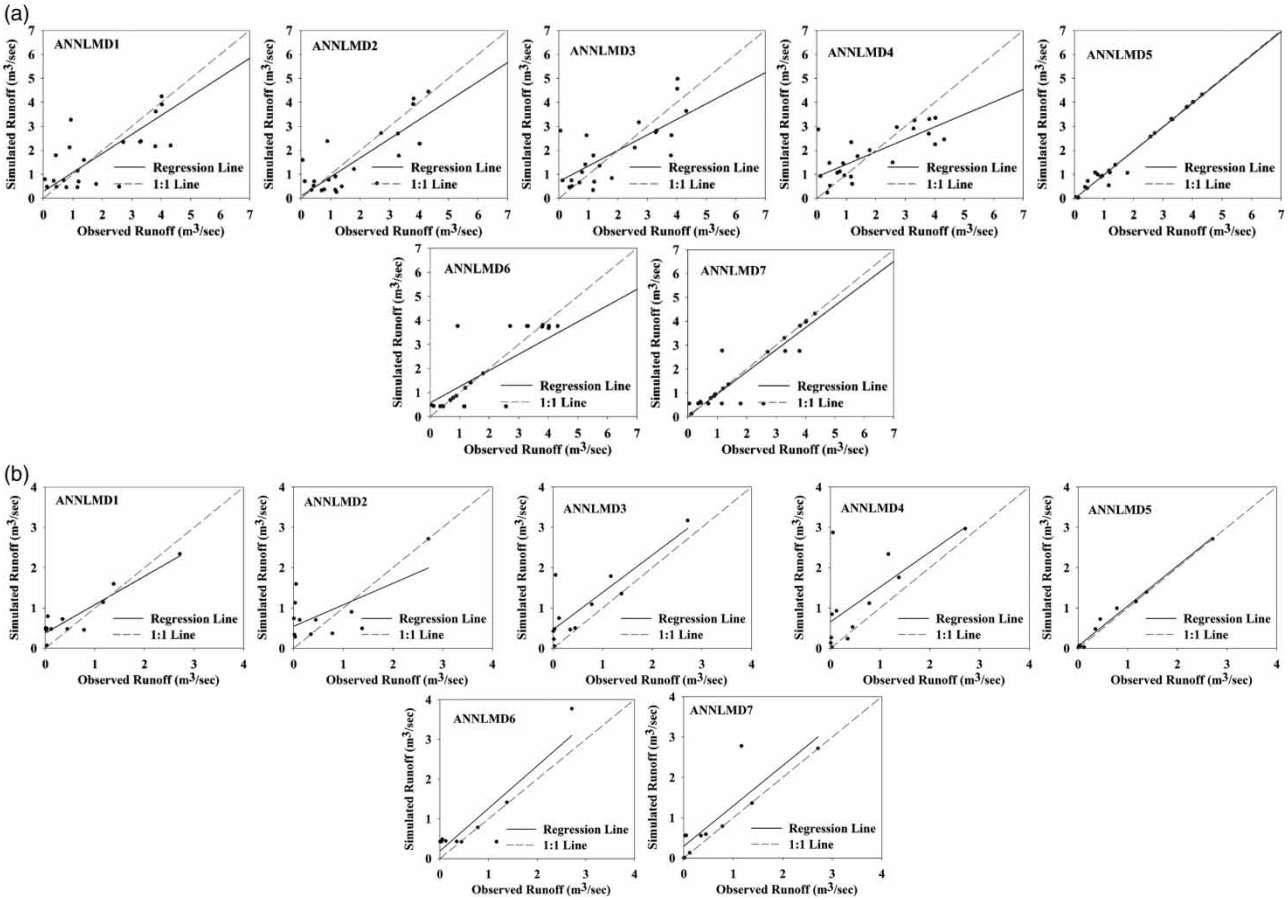

Figure 2 | Monthly observed and simulated runoff using the ANNLM model during (a) calibration and (b) validation.

Table 2 | ANN model result by SCG algorithm during calibration and validation

		Calibration				Validation			
Model number	ANN Structure	NSE	RSR	PBIAS	R^2	NSE	RSR	PBIAS	R^2
ANNSCGD1	1-5-1	0.36	0.79	17.02	0.41	−0.25	1.1	−74.91	0.31
ANNSCGD2	2-5-1	0.26	0.72	19.25	0.33	−0.52	1.19	−111.36	0.43
ANNSCGD3	3-5-1	0.34	0.71	15.98	0.38	−0.52	0.89	−88.77	0.51
ANNSCGD4	4-5-1	0.3	0.70	12.32	0.32	−0.63	1.21	−125.47	0.51
ANNSCGD5	5-5-1	0.32	0.8	12.55	0.36	−0.34	1.17	−120.24	0.52
ANNSCGD6	6-5-1	0.26	0.79	17.75	0.33	−0.51	1.28	−105.07	0.38
ANNSCGD7	7-5-1	0.43	0.72	12.72	0.45	0.27	0.89	−83.29	0.70

The model LS-SVRD8 was found to be the best among all the LS-SVR models developed (Table 3). During calibration, the values of NSE, RSR, PBIAS and R^2 were found to be 0.98, 0.12, 3.41 and 0.99, respectively, and 0.96, 0.99, 0.22, −18.81 and 0.98, respectively, during validation. Moreover, slight variation with respect to the model evaluation criteria was observed during calibration and validation for all the models developed using the LS-SVR model. These statistical

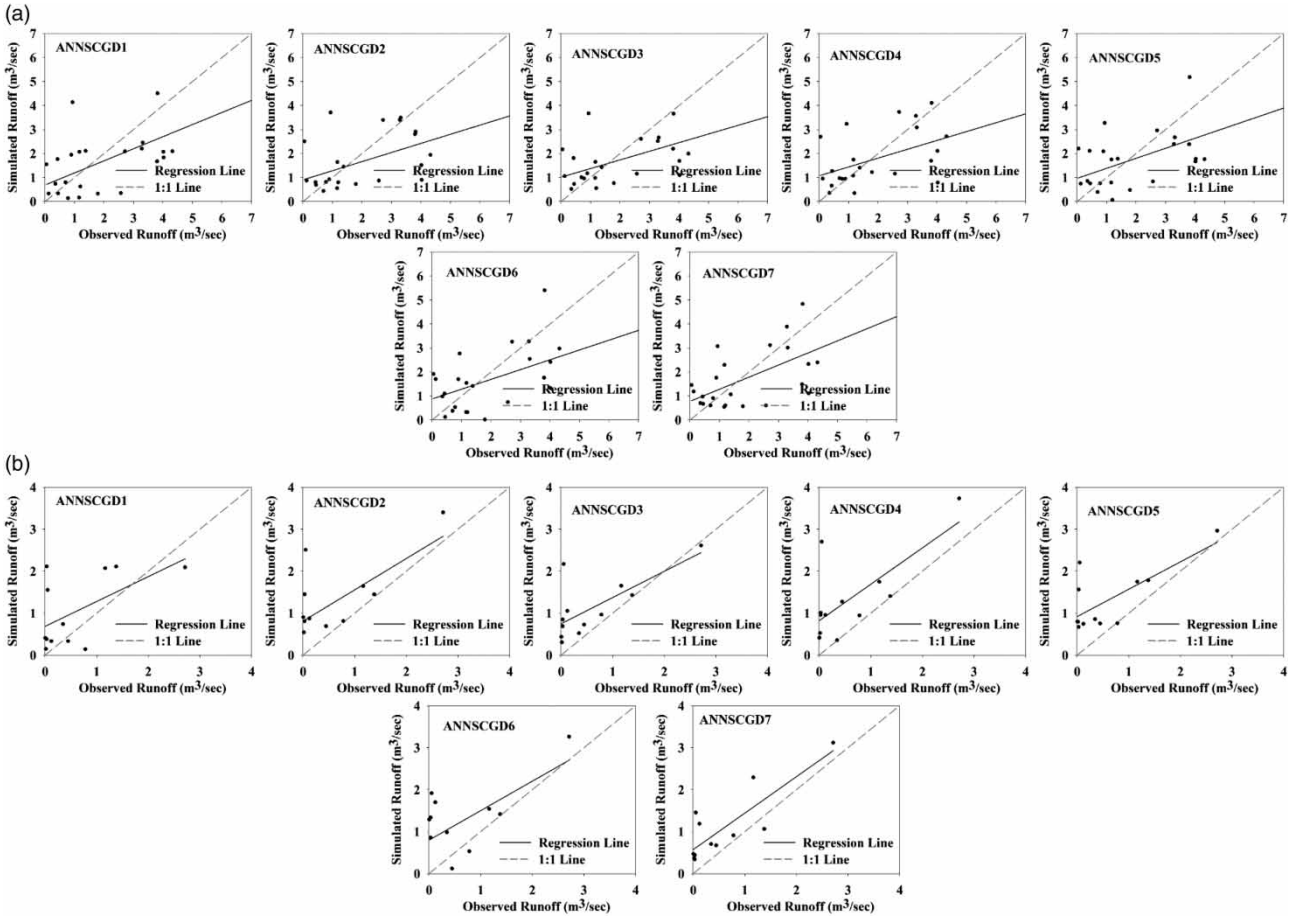

Figure 3 | Monthly observed and simulated runoff using the ANNSCG model during (a) calibration and (b) validation.

Table 3 | Results of SVR during calibration and validation

			Calibration				Validation			
Model number	**γ (Gamma)**	**σ^2 (Sigma square)**	**NSE**	**RSR**	**PBIAS**	**R^2**	**NSE**	**RSR**	**PBIAS**	**R^2**
LS-SVRD1	1	0.9	0.71	0.49	14.65	0.95	0.62	0.71	−45.48	0.90
LS-SVRD2	2	0.9	0.86	0.34	9.84	0.98	0.62	0.71	−45.48	0.90
LS-SVRD3	3	0.9	0.92	0.26	7.46	0.99	0.79	0.52	0.00	0.92
LS-SVRD4	4	0.9	0.95	0.21	6.02	0.99	0.90	0.36	−26.90	0.97
LS-SVRD5	5	0.9	0.96	0.18	5.05	0.99	0.84	0.39	−33.19	0.93
LS-SVRD6	6	0.9	0.97	0.16	4.35	0.99	0.95	0.26	−21.78	0.98
LS-SVRD7	7	0.9	0.98	0.14	3.82	0.99	0.94	0.26	−22.25	0.97
LS-SVRD8	8	0.9	0.98	0.12	3.41	0.99	0.96	0.22	−18.81	0.98
LS-SVRD9	9	0.9	0.99	0.11	3.08	0.99	0.94	0.23	−6.78	0.97
LS-SVRD10	10	0.9	0.99	0.10	2.81	0.99	0.95	0.20	−4.49	0.97

measures confirm good agreement between observed and predicted runoff values during calibration and validation. However, the peak values are captured well but not with reasonable accuracy. This may be due to the uncalibrated regularization parameter (γ), which is the driving parameter for the LS-SVR model of error minimization and smoothness of the calculated function. Scatter plots between observed and simulated runoff are presented in Figure 4(a) and 4(b).

REPTree and M5 model

The performance of the two decision trees, namely, REPTree and M5 model, were also evaluated and presented in Table 4. For the REPTree model, the statistical indicators for evaluation of models, i.e., NSE, RSR, PBIAS and R^2 were found to be 0.71, 0.49, 8.99 and 0.73, respectively, during calibration and 0.54, 0.76, −58.22 and 0.74 during validation. For the M5

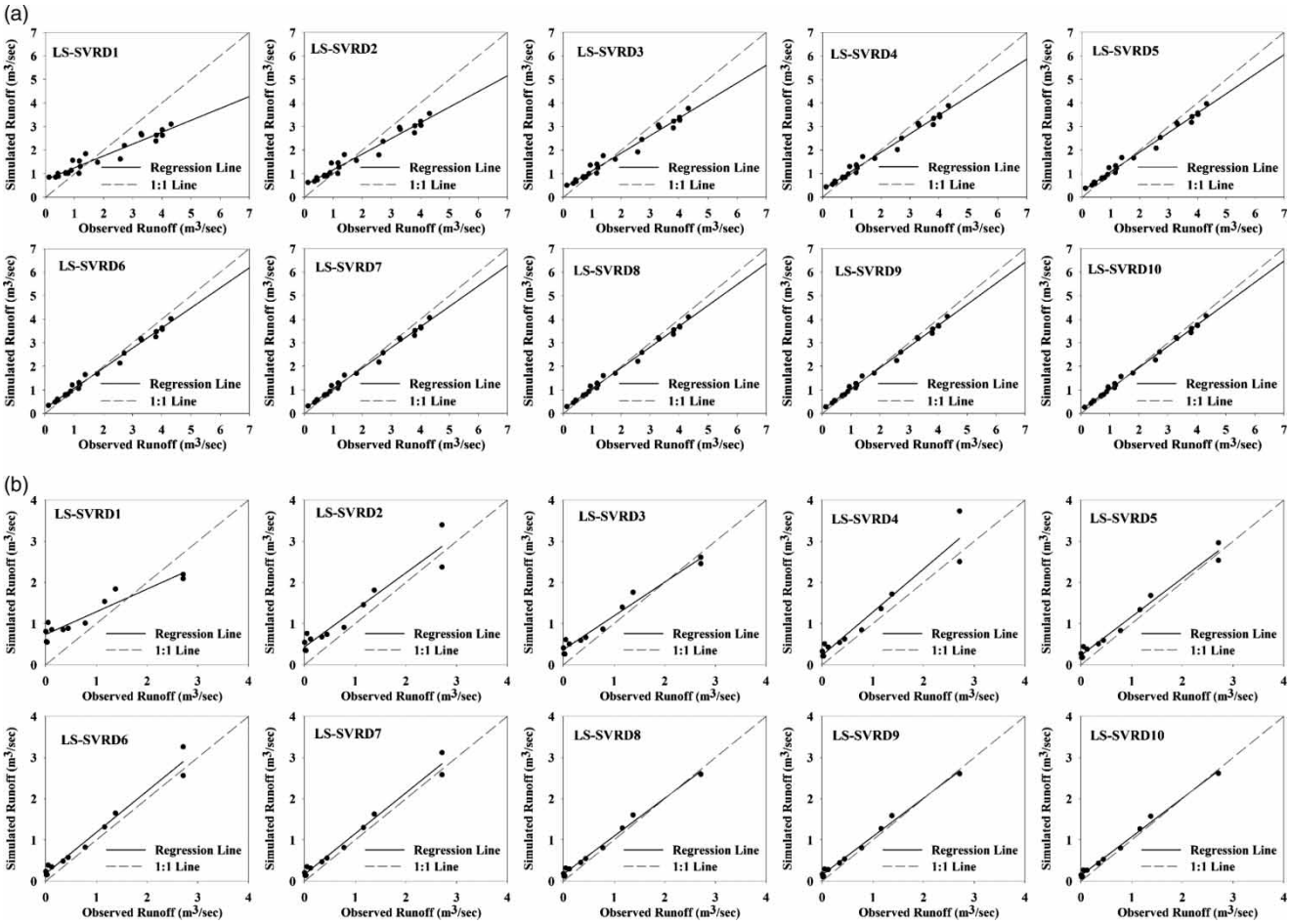

Figure 4 | Monthly observed and simulated runoff using the LS-SVR model during (a) calibration and (b) validation.

Table 4 | Results of REPTree and M5 models during calibration and validation

		Calibration				Validation			
Model	**Inputs**	**NSE**	**RSR**	**PBIAS**	R^2	**NSE**	**RSR**	**PBIAS**	R^2
M5	S_t or $D_t = f(R_t, T_t, RH_t, \ldots$ $\ldots E_t, \text{Sun}_t, \text{Solar}_t, W_t)$	0.92	0.24	3.69	0.92	0.92	0.23	−13.54	0.95
REPTree		0.71	0.49	8.99	0.73	0.54	0.76	−58.22	0.74

model, the evaluation criteria of NSE, RSR, PBIAS and R^2 were found to be 0.92, 0.24, 3.69 and 0.92, respectively, during calibration and 0.92, 0.23, −13.54 and 0.95, respectively, during validation. It is clear from the time series and scatter plots in Figure 5(a) and 5(b), as well as the evaluation criteria, that the REPTree model successfully mimics the runoff of the Pokhariya watershed during calibration. Moreover, the model also performed well in predicting the low and high peak runoff conditions. The M5 regression tree model conducted using WEKA, outperformed the REPTree model during both calibration and validation. A comparison between the statistical measures, in Table 4, clearly shows the superiority of M5 models and REPTree models with the same input combination. The result also shows that the M5 model is more accurate during validation when compared to REPTree. Figure 6(a) and 6(b) show the time series and scatter plots between observed and simulated runoff, using the M5 model.

The overall performance of the five machine learning models developed for runoff modelling in the Pokhariya watershed was considered as highly satisfactory in terms of the model evaluation criteria selected. The ANNLM5 model was rated as the best model for predicting runoff with NSE of 0.98 during validation, followed by LS-SVR8 with NSE of 0.95 during the same period. The M5 model tree appears to be the third best model for runoff forecasting with NSE value of 0.92. The performances of the other models are also satisfactory and good for field application.

Results of monthly sediment yield simulations

Artificial neural network-Levenberg–Marquardt

Performance of the ANN models, which consists of a three-layer feed forward-back propagation network, trained by LM algorithms, were evaluated for sediment yield with the

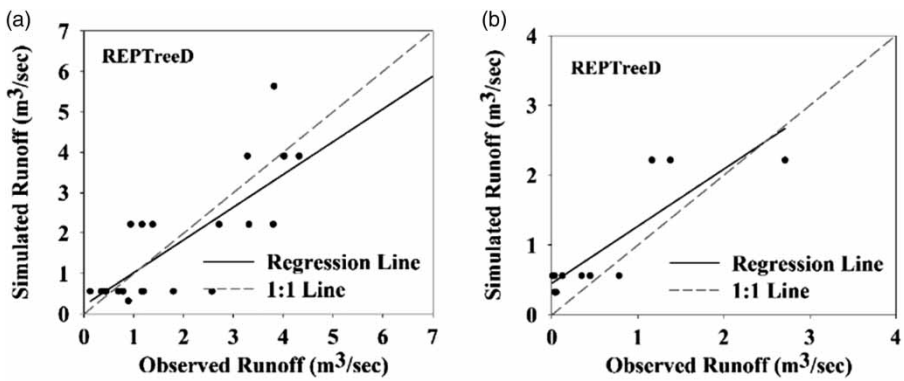

Figure 5 | Observed and simulated runoff using the REPTree model during (a) calibration and (b) validation.

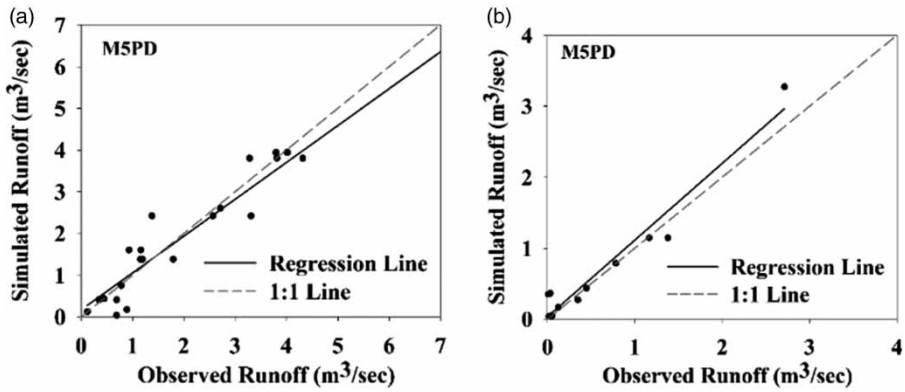

Figure 6 | Observed and simulated runoff using the M5 model during (a) calibration and (b) validation.

seven input combination developed, resulting in seven ANN-LM models (denoted as ANNLMS) and presented in Table 5. The number of neurons in the hidden layer were assumed to be constant and five neurons in the hidden layer were considered. From Table 5, it can be observed that, the model ANNLMS7 performed well compared to other models during calibration (NSE = 0.58, RSR = 0.62, PBIAS = 15.03, $R^2 = 0.65$) and validation (NSE = −0.93, RSR = 1.29, PBIAS = 37.55, $R^2 = 0.57$) whereas the model ANNLMS1 performed worst during calibration (NSE = −0.05, RSR = 1.00, PBIAS = −14.63, $R^2 = 0.01$) and validation (NSE = −2.61, RSR = 1.86, PBIAS = −18.17, $R^2 = 0.05$). Scatter graphs between observed and simulated sediment yield during calibration and validation for all the ANNLM models are presented in Figure 7(a) and 7(b).

Artificial neural network-scaled conjugate gradient

Performance of the ANN models, which consists of a three-layer feed forward-back propagation network, trained by SCG algorithms, were evaluated for sediment yield with the seven input combination developed, resulting in seven ANN-SCG models (denoted as ANNSCGS) and presented in Table 6. The number of neurons in the hidden layer were assumed to be constant and five neurons in the hidden layer were considered. From Table 6, it can be observed, that the model ANNSCGS5 performed better than other models during calibration (NSE = 0.59, RSR = 0.59, PBIAS = 1.22, $R^2 = 0.59$) and validation (NSE = −0.64, RSR = 1.14, PBIAS = −22.22, $R^2 = 0.41$) whereas the model ANNSCGS2 performed worst during calibration (NSE = 0.38, RSR = 0.76, PBIAS = 3.88, $R^2 = 0.38$) and validation (NSE = −2.55, RSR = 1.54, PBIAS = −51.09, $R^2 = 0.00$). Scatter graphs between observed and simulated sediment yield during calibration and validation for all the ANNSCGS models are presented in Figure 8(a) and 8(b).

Least square-support vector regression

The model developed employing the LS-SVR for sediment yield prediction was denoted as 'LSSVRS'. The regularization parameter (γ) was calibrated, keeping the RBF kernel function parameter (σ^2) constant with numerical value 0.9, similar to the model developed during runoff simulation. The model generated sediment yield was compared against the observed sediment yield. The model LS-SVRS10 was found to be the best among all the LS-SVR models developed (Table 7). During calibration, the values of NSE, RSR, PBIAS and R^2 were found to be 0.98, 0.00, 1.51 and 0.99, respectively, and 0.98, 0.16, −4.94 and 0.99, respectively, during validation. Moreover, slight variation with respect to the model evaluation criteria was observed during calibration and validation for all the models developed using the LS-SVR model. These statistical measures confirm good agreement between observed and predicted sediment values during calibration and validation. Scatter plots between observed and simulated sediment yield are presented in Figure 9(a) and 9(b).

REPTree and M5 model

Performance of the two decision trees, namely, REPTree and M5 model, were also evaluated and presented in Table 8. For

Table 5 | ANN model result by LM algorithm during calibration and validation

Model number	ANN structure	Calibration				Validation			
		NSE	RSR	PBIAS	R^2	NSE	RSR	PBIAS	R^2
ANNLMS1	1-5-1	−0.05	1.00	−14.63	0.01	−2.61	1.86	−18.17	0.05
ANNLMS2	2-5-1	0.42	0.74	−5.42	0.44	−2.57	1.56	29.20	0.06
ANNLMS3	3-5-1	−0.24	1.04	−29.18	0.02	−2.57	1.66	0.00	0.16
ANNLMS4	4-5-1	0.47	0.71	−13.88	0.55	−0.95	1.26	−1.84	0.00
ANNLMS5	5-5-1	0.25	0.82	1.36	0.26	−8.06	3.12	107.36	0.04
ANNLMS6	6-5-1	0.48	0.69	5.67	0.57	−0.22	1.03	27.61	0.61
ANNLMS7	7-5-1	0.58	0.62	15.03	0.65	−0.93	1.29	37.55	0.57

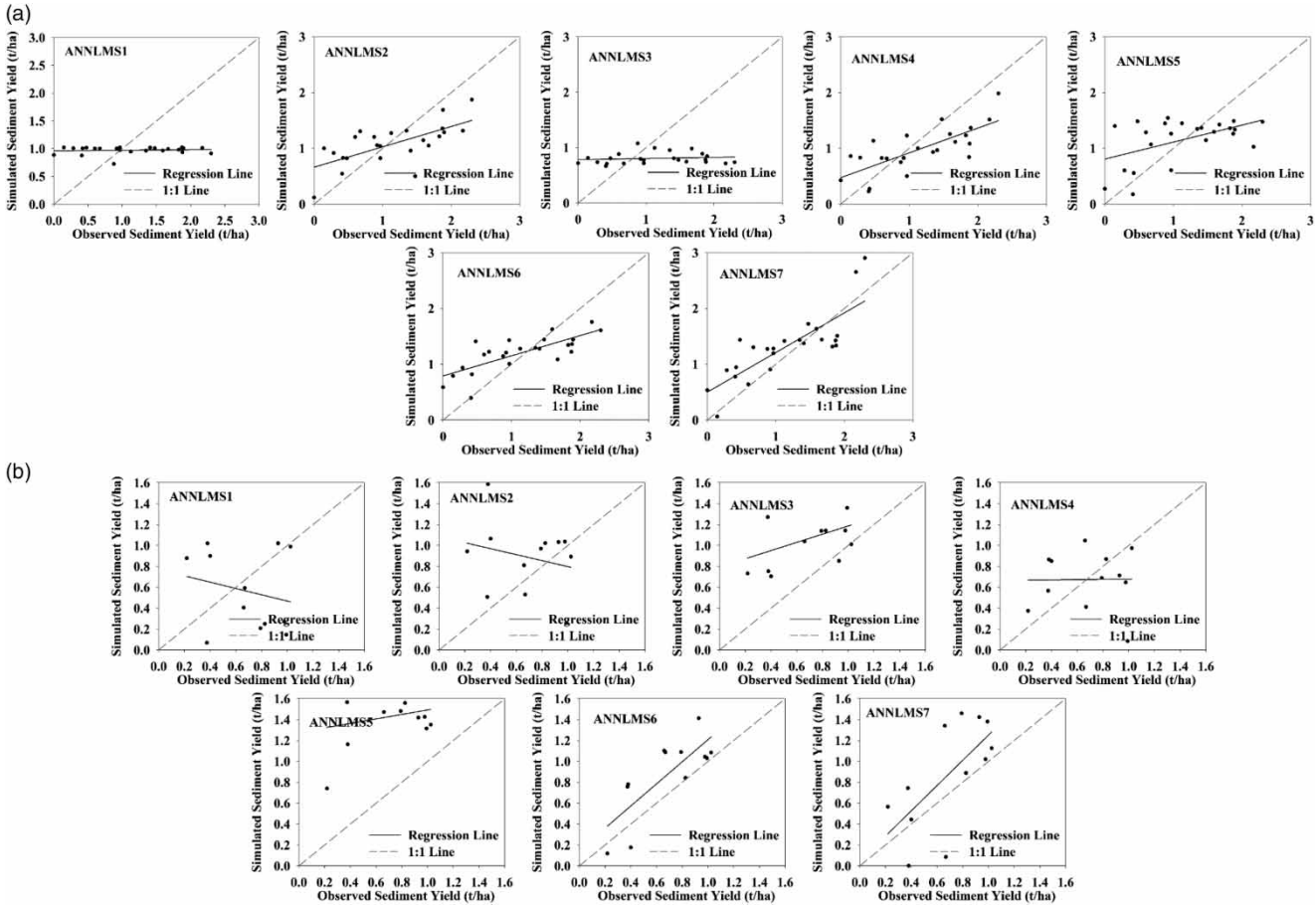

Figure 7 | Monthly observed and simulated sediment yield using the ANNLM model during (a) calibration and (b) validation.

Table 6 | ANNSCG model result for sediment yield during calibration and validation

		Calibration				Validation			
Model number	ANN structure	NSE	RSR	PBIAS	R^2	NSE	RSR	PBIAS	R^2
ANNSCGS1	1-5-1	0.07	0.93	1.00	0.07	−2.38	1.72	−53.73	0.03
ANNSCGS2	2-5-1	0.38	0.76	3.88	0.38	−2.55	1.54	−51.09	0.00
ANNSCGS3	3-5-1	0.31	0.76	2.25	0.31	−2.55	1.86	0.00	0.26
ANNSCGS4	4-5-1	0.48	0.67	2.68	0.49	−2.92	1.71	−55.05	0.01
ANNSCGS5	5-5-1	0.59	0.59	1.22	0.59	−0.64	1.14	−22.22	0.41
ANNSCGS6	6-5-1	0.21	0.86	6.39	0.24	−1.99	1.82	−62.69	0.56
ANNSCGS7	7-5-1	0.23	0.83	3.89	0.24	−1.22	1.45	−40.30	0.13

the REPTree model, the statistical indicators for evaluation of models, i.e., NSE, RSR, PBIAS and R^2 were found to 0.57, 0.54, 23.02 and 0.73, respectively, during calibration and 0.77, 0.46, −28.7 and 0.83 during validation. For the M5 model, the evaluation criteria of NSE, RSR, PBIAS and R^2 were found to be 0.82, 0.32, 15.72 and 0.89, respectively, during calibration and 0.89, 0.27, 26.72 and 0.93, respectively, during validation. It is clear from the time series and scatter

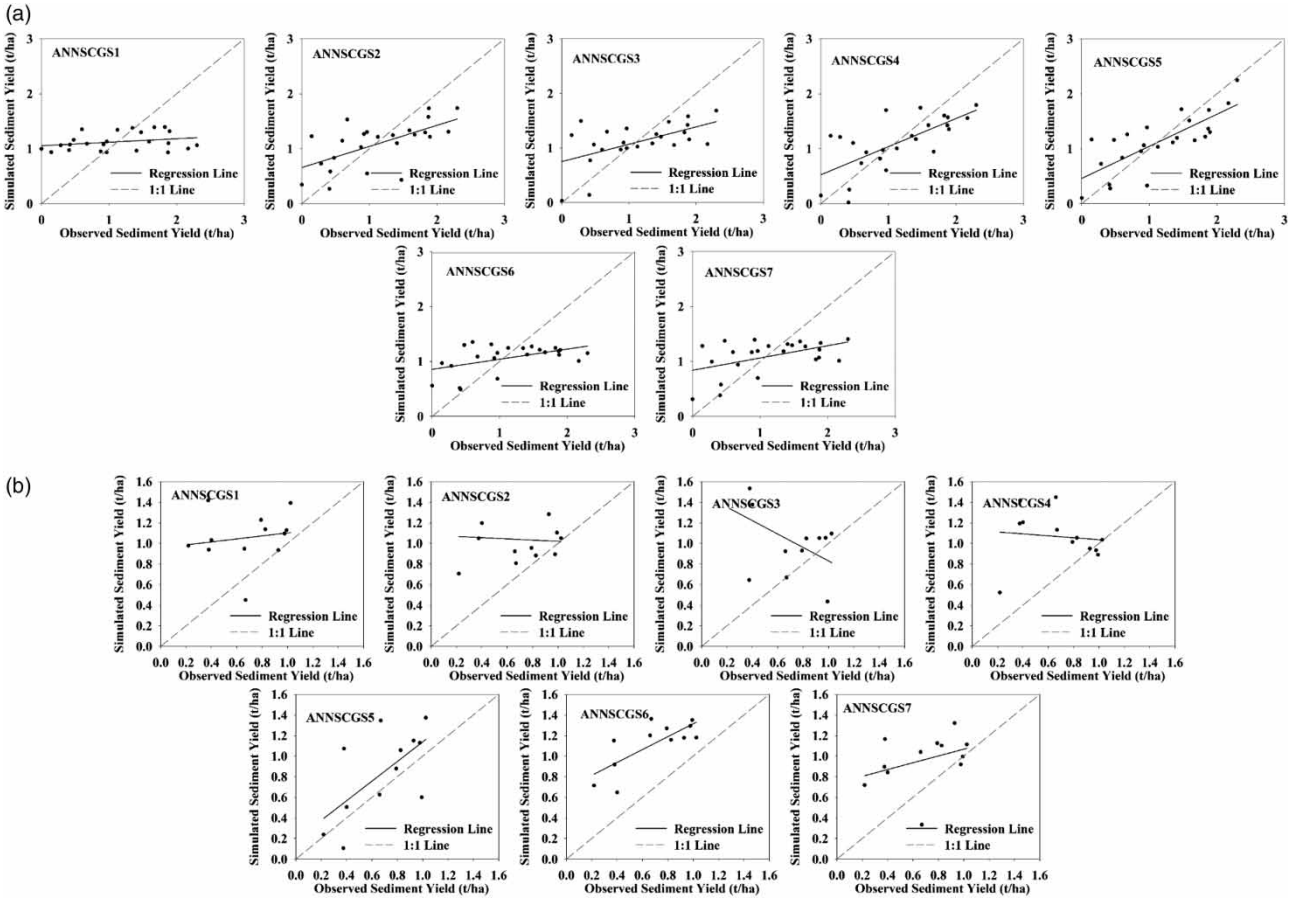

Figure 8 | Monthly observed and simulated sediment yield using the ANNSCG model during (a) calibration and (b) validation.

Table 7 | Results of SVR during calibration and validation

			Calibration				Validation			
Model number	**γ (Gamma)**	**σ^2 (Sigma square)**	**NSE**	**RSR**	**PBIAS**	**R^2**	**NSE**	**RSR**	**PBIAS**	**R^2**
LS-SVRS1	1	0.9	0.71	0.54	6.96	0.97	0.47	0.71	−23.00	0.89
LS-SVRS2	2	0.9	0.86	0.38	4.86	0.98	0.74	0.50	−16.05	0.94
LS-SVRS3	3	0.9	0.92	0.29	3.77	0.99	0.74	0.39	0.00	0.96
LS-SVRS4	4	0.9	0.94	0.24	3.09	0.99	0.90	0.32	−10.18	0.97
LS-SVRS5	5	0.9	0.96	0.20	2.62	0.99	0.93	0.27	−8.64	0.98
LS-SVRS6	6	0.9	0.97	0.17	2.28	0.99	0.94	0.24	−7.51	0.98
LS-SVRS7	7	0.9	0.98	0.15	2.02	0.99	0.96	0.21	−6.64	0.99
LS-SVRS8	8	0.9	0.98	0.00	1.82	0.99	0.96	0.19	−5.96	0.99
LS-SVRS9	9	0.9	0.98	0.00	1.65	0.99	0.97	0.17	−5.40	0.99
LS-SVRS10	10	0.9	0.98	0.00	1.51	0.99	0.98	0.16	−4.94	0.99

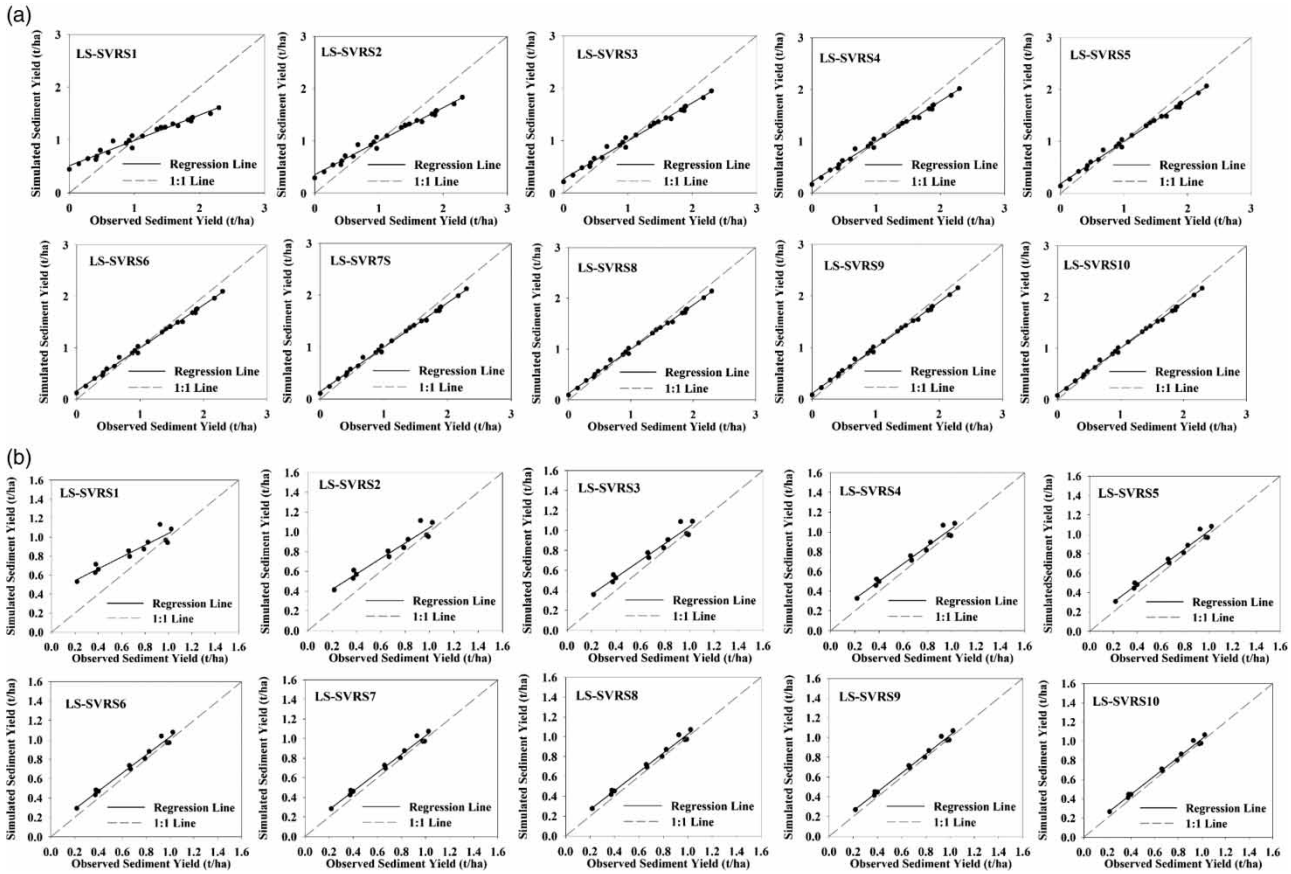

Figure 9 | Monthly observed and simulated sediment yield using the LS-SVR model during (a) calibration and (b) validation.

Table 8 | Results of REPTree and M5 models during calibration and validation

		Calibration				Validation			
Model	**Inputs**	**NSE**	**RSR**	**PBIAS**	R^2	**NSE**	**RSR**	**PBIAS**	R^2
M5	S_t or $D_t = f(R_t, T_t, RH_t, \ldots$ $\ldots E_t, \mathrm{Sun}_t, \mathrm{Solar}_t, W_t)$	0.82	0.32	15.72	0.89	0.89	0.27	26.72	0.93
REPTree		0.57	0.54	23.02	0.73	0.77	0.46	−28.74	0.83

plot (Figures 10 and 11), as well as the evaluation criteria, that the REPTree model successfully mimics the sediment yield of the Pokhariya watershed during calibration. Moreover, the model also performed well in predicting the low and high peak sediment yield conditions. The M5 regression tree model conducted using WEKA, outperformed the REPTree model during both calibration and validation. A comparison between the statistical measures in Table 8 clearly shows the superiority of the M5 model and REPTree model with the same input combination. The result also shows that the M5 model is more accurate during validation when compared to REPTree. Figures 10 and 11 show the time series and scatter plots between observed and simulated sediment yield using the M5 model.

The overall performance of the five machine learning models developed for sediment modelling in the Pokhariya watershed was considered as highly satisfactory in terms of the model evaluation criteria selected. The LSSVRS10

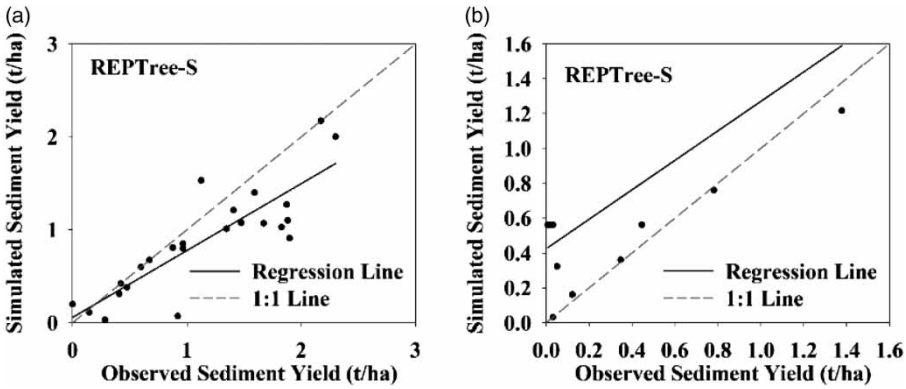

Figure 10 | Observed and simulated sediment yield using the REPTree model during (a) calibration and (b) validation.

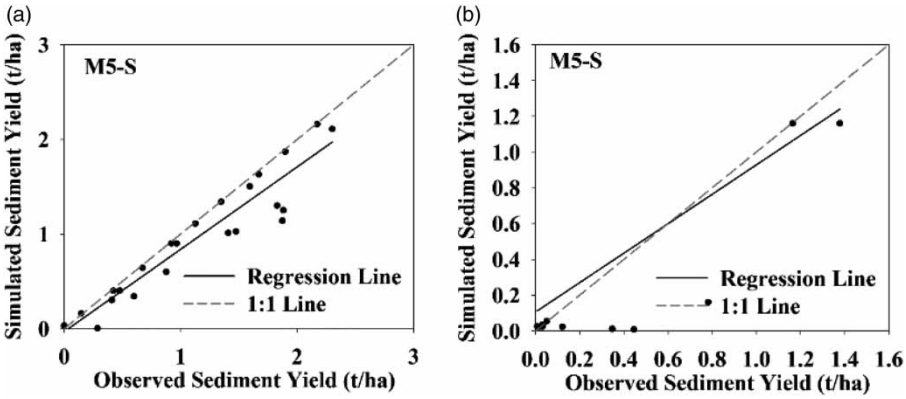

Figure 11 | Observed and simulated sediment yield using the M5 model during (a) calibration and (b) validation.

model was rated as the best model for predicting sediment yield with NSE of 0.98 during validation followed by LSSVRS9 with NSE of 0.97 during the same period. The M5 model tree appears to be the third best model for sediment yield simulation with NSE value of 0.89. The performances of the other models are also satisfactory and good for field application except for the ANN models. However, the values of RSR from all the other models are highly acceptable.

CONCLUSIONS

In the present study, five data-driven models, namely, ANN-LM, ANN-SCG, LS-SVR, REPTree and M5 model have been investigated for modelling runoff and sediment yield of the Pokhariya watershed. The highly nonlinear nature of the rainfall–runoff–sediment process is appropriate for the evaluation of these models. The monthly runoff and sediment yield data from the years 2000 to 2008 were used to calibrate and validate all models. The seven input variables to all the models comprised a combination of rainfall, average temperature, relative humidity, pan evaporation, sunshine duration, solar radiation and wind speed. The performance evaluation of each model was carried out employing common statistical evaluation indices, i.e., NSE, RSR, PBIAS and R^2. The following conclusions are drawn from the present study:

(1) A good agreement between observed and predicted values for runoff as well as sediment yield were obtained for all the models investigated.
(2) The ANNLM model outperformed LS-SVR and all the other models investigated during calibration and validation for runoff modelling whereas the LS-SVR model surpassed the ANN model and other models for sediment yield modelling. It can be concluded that the

LS-SVR model can be used as a tool for predicting the sediment yield at a single point of interest in the Pokhariya watershed, Jharkhand.

(3) Within decision tree models, M5 model tree is better in simulating sediment yield and runoff than its near counterpart, the REPTree model and marginally inferior when compared to LS-SVR and ANN models.

The results from the study conclude that the machine learning techniques are promising in the simulation of runoff and sediment yield in the study watershed and can be applied in real-life conditions. However, one should be prudent when applying machine learning techniques in real-life problems since the output from these models are only as good as the quality of input datasets employed. Moreover, the issue pertaining to the analysis of hydrological time series in developing countries is the insufficiency of data, which means the length of time series is short (Qian & Leung 2007; Wang *et al.* 2014, 2015). Therefore, data uncertainty should be assessed when employing these models to meet the need of scientific and sustainable management of water resources.

REFERENCES

Agarwal, A. & Singh, R. D. 2004 Runoff modelling through back propagation artificial neural network with variable rainfall-runoff data. *Water Resources Management* **18** (3), 285–300.

Agarwal, A., Singh, R. D., Mishra, S. K. & Bhunya, P. K. 2005 ANN-based sediment yield models for Vamsadhara river basin India. *Water SA* **31** (1), 85–100.

Anmala, J., Zhang, B. & Govindaraju, R. S. 2000 Comparison of ANNs and empirical approaches for predicting watershed runoff. *Journal of Water Resources Planning and Management* **126** (3), 156–166.

Arnold, J. G. & Allen, P. M. 1999 Automated methods for estimating base flow and groundwater recharge from streamflow records. *Journal of the American Water Resources Association* **35** (2), 411–424.

Babel, M. S., Sirisena, T. A. J. G. & Singhrattna, N. 2017 Incorporating large-scale atmospheric variables in long-term seasonal rainfall forecasting using artificial neural networks: an application to the Ping Basin in Thailand. *Hydrology Research* **48** (3), 867–882. doi: 10.2166/nh.2016.212.

Bachmair, S. & Weiler, M. 2012 Hillslope characteristics as controls of subsurface flow variability. *Hydrology and Earth System Sciences* **16** (10), 3699–3715.

Bhattacharya, B. & Solomatine, D. P. 2005 Neural networks and M5 model trees in modelling water level–discharge relationship. *Neurocomputing* **63**, 381–396.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. 1984 *Classification and Regression Trees*. CRC Press, Boca Raton, FL, USA.

Chiang, Y. M., Chang, L. C. & Chang, F. J. 2004 Comparison of static-feed forward and dynamic-feedback neural networks for rainfall–runoff modeling. *Journal of Hydrology* **290** (3–4), 297–311.

Choy, K. Y. & Chan, C. W. 2003 Modelling of river discharges and rainfall using radial basis function networks based on support vector regression. *International Journal of Systems Science* **34** (14–15), 763–773.

Cigizoglu, H. K. 2004 Estimation and forecasting of daily suspended sediment data by multi-layer perceptrons. *Advances in Water Resources* **27** (2), 185–195.

Çimen, M. 2008 Estimation of daily suspended sediments using support vector machines. *Hydrological Sciences Journal* **53** (3), 656–666.

Clark, L. A. & Pregibon, D. 1992 Tree-based models. In: *Statistical Models in S* (J. M. Chambers & T. J. Hastie, eds). Wadsworth, Pacific Grove, CA, USA, pp. 377–419.

Cobaner, M., Unal, B. & Kisi, O. 2009 Suspended sediment concentration estimation by an adaptive neuro-fuzzy and neural network approaches using hydro-meteorological data. *Journal of Hydrology* **367** (1), 52–61.

Cohen Liechti, T., Matos, J. P., Boillat, J. L., Portela, M. M. & Schleiss, A. J. 2014a Hydraulic–hydrologic model for water resources management of the Zambezi basin. *Journal of Applied Water Engineering and Research* **2** (2), 105–117.

Cohen Liechti, T., Matos, J. P., Ferràs Segura, D., Boillat, J. L. & Schleiss, A. J. 2014b Hydrological modelling of the Zambezi River Basin taking into account floodplain behaviour by a modified reservoir approach. *International Journal of River Basin Management* **12** (1), 29–41.

De Vos, N. J. & Rientjes, T. H. M. 2005 Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation. *Hydrology and Earth System Sciences Discussions* **2** (1), 365–415.

Dibike, Y. B., Velickov, S., Solomatine, D. & Abbott, M. B. 2001 Model induction with support vector machines: introduction and applications. *Journal of Computing in Civil Engineering* **15** (3), 208–216.

Feidas, H. 2010 Validation of satellite rainfall products over Greece. *Theoretical and Applied Climatology* **99** (1–2), 193–216.

Fernando, D. A. K. & Jayawardena, A. W. 1998 Runoff forecasting using RBF networks with OLS algorithm. *Journal of Hydrologic Engineering* **3** (3), 203–209.

French, M. N., Krajewski, W. F. & Cuykendall, R. R. 1992 Rainfall forecasting in space and time using a neural network. *Journal of Hydrology* **137** (1), 1–31.

Galelli, S. & Castelletti, A. 2013 Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. *Hydrology and Earth System Sciences* **17** (7), 2669–2684.

Gorgij, A. D., Kisi, O. & Moghaddam, A. A. 2017 Groundwater budget forecasting, using hybrid wavelet-ANN-GP modelling: a case study of Azarshahr Plain, East Azerbaijan, Iran. *Hydrology Research* **48** (2), 455–467. doi: 10.2166/nh.2016.202.

Goyal, M. K. & Ojha, C. S. P. 2012 Downscaling of precipitation on a lake basin: evaluation of rule and decision tree induction algorithms. *Hydrology Research* **43** (3), 215–230.

Goyal, M. K., Burn, D. H. & Ojha, C. S. P. 2012 Evaluation of machine learning tools as a statistical downscaling tool: temperatures projections for multi-stations for Thames River Basin, Canada. *Theoretical and Applied Climatology* **108** (3–4), 519–534.

Goyal, M. K., Bharti, B., Quilty, J., Adamowski, J. & Pandey, A. 2014 Modeling of daily pan evaporation in subtropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS. *Expert Systems with Applications* **41** (11), 5267–5276.

Gupta, H. V., Sorooshian, S. & Yapo, P. O. 1999 Status of automatic calibration for hydrologic models: comparison with multilevel expert calibration. *Journal of Hydrologic Engineering* **42**, 135–143.

Hagan, M. T., Demuth, H. B., Beale, M. H. & De Jesús, O. 1996 *Neural Network Design*, Vol. 20. PWS Publishing Company, Boston, MA, USA.

Han, D. A. W. E. I. & Cluckie, I. 2004 Support vector machines identification for runoff modeling. In: *Proceedings of the Sixth International Conference on Hydro Informatics*, June, Singapore, pp. 21–24.

Hornik, K., Stinchcombe, M. & White, H. 1989 Multilayer feedforward networks are universal approximators. *Neural Networks* **25**, 359–366.

Hyafil, L. & Rivest, R. L. 1976 Constructing optimal binary decision trees is NP-complete. *Information Processing Letters* **5** (1), 15–17.

Iorgulescu, I. & Beven, K. J. 2004 Nonparametric direct mapping of rainfall-runoff relationships: An alternative approach to data analysis and modeling. *Water Resources Research* **40** (8), 1–11.

Jain, S. K. 2001 Development of integrated sediment rating curves using ANNs. *Journal of Hydraulic Engineering* **127** (1), 30–37.

Jain, S. K. 2012 Modeling river stage–discharge–sediment rating relation using support vector regression. *Hydrology Research* **43** (6), 851–861.

Jain, M. K. & Kothyari, U. C. 2000 Estimation of soil erosion and sediment yield using GIS. *Hydrological Sciences Journal* **45** (5), 771–786.

Khan, M. S. & Coulibaly, P. 2006 Application of support vector machine in lake water level prediction. *Journal of Hydrologic Engineering* **113**, 199–205.

Kim, S., Kisi, O., Seo, Y., Singh, V. P. & Lee, C. J. 2017 Assessment of rainfall aggregation and disaggregation using data-driven models and wavelet decomposition. *Hydrology Research* **48** (1), 99–116. doi: 10.2166/nh.2016.314.

Kisi, O. & Cimen, M. 2011 A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *Journal of Hydrology* **399** (1), 132–140.

Kisi, O. & Cimen, M. 2012 Precipitation forecasting by using wavelet-support vector machine conjunction model. *Engineering Applications of Artificial Intelligence* **25** (4), 783–792.

Krishna, B., Satyaji Rao, Y. R. & Vijaya, T. 2008 Modelling groundwater levels in an urban coastal aquifer using artificial neural networks. *Hydrological Processes* **22** (8), 1180–1188.

Kumar, D., Pandey, A., Sharma, N. & Flügel, W. A. 2014 Modeling suspended sediment using artificial neural networks and TRMM-3B42 Version 7 rainfall dataset. *Journal of Hydrologic Engineering*. doi:10.1061/ASCEHE.1943-5584.0001082.

Kumar, D., Pandey, A., Sharma, N. & Flügel, W. A. 2016 Daily suspended sediment simulation using machine learning approach. *Catena* **138**, 77–90.

Laflen, J. M., Lane, L. J. & Foster, G. R. 1991 WEPP: a new generation of erosion prediction technology. *Journal of Soil and Water Conservation* **461**, 34–38.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of goodness of fit measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **351**, 233–241.

Licznar, P. & Nearing, M. A. 2003 Artificial neural networks of soil erosion and runoff prediction at the plot scale. *Catena* **51** (2), 89–114.

Lin, G. F. & Chen, L. H. 2004 A non-linear rainfall-runoff model using radial basis function network. *Journal of Hydrology* **289** (1), 1–8.

Liong, S. Y. & Sivapragasam, C. 2002 Flood stage forecasting with support vector machines. *JAWRA: Journal of the American Water Resources Association* **38** (1), 173–186.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling and Software* **15** (1), 101–124.

Marquez, A. M. & Guevara-Pérez, E. 2010 Comparative analysis of erosion modeling techniques in a basin of Venezuela. *Journal of Urban and Environmental Engineering* **4** (2), 81–104.

Mattera, D. & Haykin, S. 1999 Support vector machines for dynamic reconstruction of a chaotic system. In: *Advances in Kernel Methods: Support Vector Learning* (B. Scholkopf, C. J. C. Burges & A. J. Smola, eds). MIT Press, Cambridge, MA, USA, pp. 211–241.

Misra, D., Oommen, T., Agarwal, A., Mishra, S. K. & Thompson, A. M. 2009 Application and analysis of support vector machine based simulation for runoff and sediment yield. *Biosystems Engineering* **103** (4), 527–535.

Møller, M. F. 1993 A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks* **6** (4), 525–533.

Mount, N. J., Abrahart, R. J., Dawson, C. W. & Ab Ghani, N. 2012 The need for operational reasoning in data-driven rating curve prediction of suspended sediment. *Hydrological Processes* **262** (6), 3982–4000.

Naik, M. G., Rao, E. P. & Eldho, T. I. 2009 Finite element method and GIS based distributed model for soil erosion and sediment yield in a watershed. *Water Resources Management* **23** (3), 553–579.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models. Part I – a discussion of principles. *Journal of Hydrology* **10** (3), 282–290.

Obradovic, D. & Deco, G. 1996 An information theory based learning paradigm for linear feature extraction. *Neurocomputing* **12** (2), 203–221.

Pandey, A., Chowdary, V. M., Mal, B. C. & Billib, M. 2008 Runoff and sediment yield modeling from a small agricultural watershed in India using the WEPP model. *Journal of Hydrology* **348** (3–4), 305–319.

Partal, T. & Cigizoglu, H. K. 2008 Estimation and forecasting of daily suspended sediment data using wavelet–neural networks. *Journal of Hydrology* **358** (3), 317–331.

Prabhanjan, A., Rao, E. P. & Eldho, T. I. 2014 Application of SWAT model and geospatial techniques for sediment-yield modeling in ungauged watersheds. *Journal of Hydrologic Engineering*. doi: 10.1061/ASCEHE.1943-5584.0001123.

Qian, Y. & Leung, L. R. 2007 A long-term regional simulation and observations of the hydroclimate in China. *Journal of Geophysical Research: Atmospheres* **112** (D14), 1–11.

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

Raghuwanshi, N. S., Singh, R. & Reddy, L. S. 2006 Runoff and sediment yield modeling using artificial neural networks: Upper Siwane River, India. *Journal of Hydrologic Engineering* **11** (1), 71–79.

Sarma, B., Sarma, A. K., Mahanta, C. & Singh, V. P. 2015 Optimal ecological management practices for controlling sediment yield and peak discharge from hilly urban areas. *Journal of Hydrologic Engineering*. doi: 10.1061/ASCEHE.1943-5584.0001154.

Sauquet, E. & Catalogne, C. 2011 Comparison of catchment grouping methods for flow duration curve estimation at ungauged sites in France. *Hydrology and Earth System Sciences Discussions* **15**, 2421–2435.

Shamseldin, A. Y. 1997 Application of a neural network technique to rainfall-runoff modelling. *Journal of Hydrology* **199** (3), 272–294.

Singh, J., Knapp, H. V. & Demissie, M. 2004 *Hydrologic Modelling of the Iroquois River Watershed Using HSPF and SWAT*. ISWS CR 2004-08. Illinois State Water Survey, Champaign, IL, USA. Available at: www.sws.uiuc.edu/pubdoc/CR/ISWSCR2004-08.pdf (accessed 8 September 2005).

Singh, K. K., Pal, M. & Singh, V. P. 2010 Estimation of mean annual flood in Indian catchments using backpropagation neural network and M5 model tree. *Water Resources Management* **24** (10), 2007–2019.

Sivapragasam, C. & Muttil, N. 2005 Discharge rating curve extension–a new approach. *Water Resources Management* **19** (5), 505–520.

Smola, A. J. 1996 Regression estimation with support vector learning machines. Master's Thesis, Technische Universitat Munchen, Germany.

Solomatine, D. P. & Dulal, K. N. 2003 Model trees as an alternative to neural networks in rainfall–runoff modelling. *Hydrological Sciences Journal* **48** (3), 399–411.

Solomatine, D. P. & Xue, Y. 2004 M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. *Journal of Hydrologic Engineering* **9** (6), 491–501.

Suykens, J. A. & Vandewalle, J. 1999 Least squares support vector machine classifiers. *Neural Processing Letters* **9** (3), 293–300.

Torgo, L. 1997 Functional models for regression tree leaves. In: *Machine Learning, Proceedings of the 14th International Conference* (D. Fisher, ed.). Morgan Kaufmann, pp. 385–393.

Tripathi, M. P., Panda, R. K. & Raghuwanshi, N. S. 2003 Identification and prioritisation of critical sub-watersheds for soil conservation management using the SWAT model. *Biosystems Engineering* **85** (3), 365–379.

Vapnik, V. 1995 *The Nature of Statistical Learning Theory*. Springer Verlag, New York, USA.

Vapnik, V. & Chervonenkis, A. 1964 A note on one class of perceptrons. *Automation and Remote Control* **25** (1).

Vapnik, V. & Lerner, A. 1963 Pattern recognition using generalized portrait method. *Automation and Remote Control* **24**, 774–780.

Verbyla, D. L. 1987 Classification trees: a new discrimination tool. *Canadian Journal of Forest Research* **17** (9), 1150–1152.

Wang, D., Singh, V. P., Shang, X., Ding, H., Wu, J., Wang, L. & Wang, Z. 2014 Sample entropy-based adaptive wavelet de-noising approach for meteorologic and hydrologic time series. *Journal of Geophysical Research: Atmospheres* **119** (14), 8726–8740.

Wang, D., Ding, H., Singh, V. P., Shang, X., Liu, D., Wang, Y. & Zou, X. 2015 A hybrid wavelet analysis–cloud model data-extending approach for meteorologic and hydrologic time series. *Journal of Geophysical Research: Atmospheres* **120** (9), 4057–4071.

Wei, W. & Watkins Jr, D. W. 2011 Data mining methods for hydroclimatic forecasting. *Advances in Water Resources* **34** (11), 1390–1400.

Willmott, C. J. 1981 On the validation of models. *Physical Geography* **2**, 184–194.

Wu, C. L., Chau, K. W. & Li, Y. S. 2008 River stage prediction based on a distributed support vector regression. *Journal of Hydrology* **358** (1), 96–111.

Yang, C. C., Prasher, S. O., Lacroix, R., Sreekanth, S., Patni, N. K. & Masse, L. 1997 Artificial neural network model for subsurface-drained farmlands. *Journal of Irrigation and Drainage Engineering* **123** (4), 285–292.

Yu, X. & Liong, S. Y. 2007 Forecasting of hydrologic time series with ridge regression in feature space. *Journal of Hydrology* **332** (3), 290–302.

Zhang, B. & Govindaraju, R. S. 2003 Geomorphology-based artificial neural networks GANNs for estimation of direct runoff over watersheds. *Journal of Hydrology* **273** (1), 18–34.

First received 31 May 2016; accepted in revised form 5 November 2016. Available online 6 January 2017