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

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 2005). 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 (Lafren et al.
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 & Muttill 2005; Khan & Coulibaly 2006; Cimen 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 & Dural 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, ANNL, ANNCG, 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
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:**

\[ S_t \lor D_t = f(R_t) \] (1)

**Combination 2:**

\[ S_t \lor D_t = f(R_t, T_t) \] (2)

**Combination 3:**

\[ S_t \lor D_t = f(R_t, T_t, RH_t) \] (3)

**Combination 4:**

\[ S_t \lor D_t = f(R_t, T_t, RH_t, E_t) \] (4)

**Combination 5:**

\[ S_t \lor D_t = f(R_t, T_t, RH_t, E_t, Sun_t) \] (5)

**Combination 6:**

\[ S_t \lor D_t = f(R_t, T_t, RH_t, E_t, Sun_t, Solar_t) \] (6)
Combination 7:

\[ S_t \text{ or } D_t = f(R_t, T_t, RH_t, E_t, Sun_t, Solar_t, W_t) \quad (7) \]

where \( S_t \) is sediment yield (t/ha) at time \( t \), \( D_t \) is runoff (m³/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/m²) 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:

\[
\tilde{x} = \frac{x - x_{\text{ori}}}{x_{\text{max}} - x_{\text{min}}} 
\]

where \( x_{\text{norm}} \) and \( x_{\text{ori}} \) represent the normalized and original data. \( x_{\text{max}} \) and \( x_{\text{min}} \) 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:

\[
X_{k+1} = X_k - [J_k^T J_k + \mu I]^{-1} J_k^T e 
\]

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:

\[
\omega_{k+1} = \omega_k + \alpha_k \sigma_k \rho_k \quad (10)
\]

\[
\rho_k = E'(\omega) + \alpha_k \sigma_k \rho_{k+1} \quad (11)
\]

where \( \rho_k \) and \( \rho_{k+1} \) are the conjugate directions in successive iterations. \( \alpha_k \) and \( \omega_k \) are calculated in each iteration. SCG method to calculate the Hessian matrix which is approximated by:

\[
E''(\omega_k) \rho_k = E'(\omega_k + \sigma_k \rho_k) - E'(\omega_k) + \lambda_k \rho_k \quad (12)
\]

where \( E' \) and \( E'' \) are the first and second derivative of \( E \), \( \rho_k \), \( \sigma_k \) and \( \lambda_k \) 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 \( \sigma \) should be as small as possible (\( \sigma \leq 10^{-4} \)) (Moller 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:

\[ y = w^T \varphi(x) + b \]  \hspace{1cm} (13)

where w is known as weight vector and b as bias. \( \varphi(x) \) 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 w. The parameters are estimated by minimizing the cost function \( J(w, e) \). The LS-SVM optimization problem for function estimation is formulated by minimizing the cost function:

\[ J(w, e) = \frac{1}{2}w^Tw + \gamma \sum_{i=1}^{N} e_i^2 \]  \hspace{1cm} (14)

Subject to the equality constraint

\[ y_i = w^T \varphi(x_i) + b + e_i \quad i = 1, \ldots, N \]

where \( e_i \) is the random error and \( \gamma \) 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:

\[ L(w, b, e; \alpha) = J(w, e) - \sum_{i=1}^{N} \alpha_i \left\{ w^T \varphi(x_i) + b + e_i - y_i \right\} \]  \hspace{1cm} (15)

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

\[ \frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i \varphi(x_i) \]  \hspace{1cm} (16)

\[ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0 \]  \hspace{1cm} (17)

\[ \frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i, \quad i = 1, \ldots, N \]  \hspace{1cm} (18)

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

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

\[ K(x, x_i) = \varphi(x_i)^T \varphi(x) \]  \hspace{1cm} (20)

where \( \varphi(x) \) 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:

\[ \hat{y} = \sum_{i=1}^{N} \hat{\alpha}_i K(x, x_i) + \hat{b} \]  \hspace{1cm} (21)

where \( K(x, x_i) \) is a kernel function, \( \hat{\alpha}_i \) and \( \hat{b} \) are the estimated values of \( \alpha_i \) 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:

\[ K(x, x_i) = \exp \left( -\frac{1}{2\sigma^2} \| x - x_i \| ^2 \right) \]  \hspace{1cm} (22)

where \( \sigma \) 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 \( \hat{y} \) 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:

\[ SDR = \text{sd}(T) - \sum_{i} \frac{|T_i|}{|T|} \cdot \text{sd}(T_i) \] (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, ANNCG, 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 ANNSCG 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, $R^2 = 0.52$) and validation (NSE = 0.65, 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 ($\gamma$, $\sigma$) to be determined. In this study, the regularization parameter ($\gamma$), 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 ($\sigma^2$) constant with numerical value 0.9. The regularization parameter ($\gamma$) 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>
<thead>
<tr>
<th>Model number</th>
<th>ANN structure</th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NSE</td>
<td>RSR</td>
<td>PBIAS</td>
</tr>
<tr>
<td>ANNLMD1</td>
<td>1-5-1</td>
<td>0.69</td>
<td>0.48</td>
</tr>
<tr>
<td>ANNLMD2</td>
<td>2-5-1</td>
<td>0.72</td>
<td>0.48</td>
</tr>
<tr>
<td>ANNLMD3</td>
<td>3-5-1</td>
<td>0.7</td>
<td>0.47</td>
</tr>
<tr>
<td>ANNLMD4</td>
<td>4-5-1</td>
<td>0.65</td>
<td>0.49</td>
</tr>
<tr>
<td>ANNLMD5</td>
<td>5-5-1</td>
<td>0.98</td>
<td>0.07</td>
</tr>
<tr>
<td>ANNLMD6</td>
<td>6-5-1</td>
<td>0.75</td>
<td>0.34</td>
</tr>
<tr>
<td>ANNLMD7</td>
<td>7-5-1</td>
<td>0.87</td>
<td>0.23</td>
</tr>
</tbody>
</table>
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 ANN models during (a) calibration and (b) validation.

Table 3 | Results of SVR during calibration and validation

<table>
<thead>
<tr>
<th>Model number</th>
<th>( \gamma ) (Gamma)</th>
<th>( \sigma^2 ) (Sigma square)</th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>NSE</td>
<td>RSR</td>
</tr>
<tr>
<td>LS-SVRD1</td>
<td>1</td>
<td>0.9</td>
<td>0.71</td>
<td>0.49</td>
</tr>
<tr>
<td>LS-SVRD2</td>
<td>2</td>
<td>0.9</td>
<td>0.86</td>
<td>0.34</td>
</tr>
<tr>
<td>LS-SVRD3</td>
<td>3</td>
<td>0.9</td>
<td>0.92</td>
<td>0.26</td>
</tr>
<tr>
<td>LS-SVRD4</td>
<td>4</td>
<td>0.9</td>
<td>0.95</td>
<td>0.21</td>
</tr>
<tr>
<td>LS-SVRD5</td>
<td>5</td>
<td>0.9</td>
<td>0.96</td>
<td>0.18</td>
</tr>
<tr>
<td>LS-SVRD6</td>
<td>6</td>
<td>0.9</td>
<td>0.97</td>
<td>0.16</td>
</tr>
<tr>
<td>LS-SVRD7</td>
<td>7</td>
<td>0.9</td>
<td>0.98</td>
<td>0.14</td>
</tr>
<tr>
<td>LS-SVRD8</td>
<td>8</td>
<td>0.9</td>
<td>0.98</td>
<td>0.12</td>
</tr>
<tr>
<td>LS-SVRD9</td>
<td>9</td>
<td>0.9</td>
<td>0.99</td>
<td>0.11</td>
</tr>
<tr>
<td>LS-SVRD10</td>
<td>10</td>
<td>0.9</td>
<td>0.99</td>
<td>0.10</td>
</tr>
</tbody>
</table>
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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Inputs</th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NSE</td>
<td>RSR</td>
</tr>
<tr>
<td>M5</td>
<td>$S_r$ or $D_t = f(R_o, T_r, R_H, \ldots$</td>
<td>0.92</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>$\ldots E_t, S_{un}, S_{olar}, W_t$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REPTree</td>
<td></td>
<td>0.71</td>
<td>0.49</td>
</tr>
</tbody>
</table>
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
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 = −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 ANNLMS 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-SCGS 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 ($\gamma$) was calibrated, keeping the RBF kernel function parameter ($\sigma^2$) 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.

---

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

<table>
<thead>
<tr>
<th>Model number</th>
<th>ANN structure</th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NSE</td>
<td>RSR</td>
<td>PBIAS</td>
</tr>
<tr>
<td>ANNLMS1</td>
<td>1-5-1</td>
<td>−0.05</td>
<td>1.00</td>
</tr>
<tr>
<td>ANNLMS2</td>
<td>2-5-1</td>
<td>0.42</td>
<td>0.74</td>
</tr>
<tr>
<td>ANNLMS3</td>
<td>3-5-1</td>
<td>−0.24</td>
<td>1.04</td>
</tr>
<tr>
<td>ANNLMS4</td>
<td>4-5-1</td>
<td>0.47</td>
<td>0.71</td>
</tr>
<tr>
<td>ANNLMS5</td>
<td>5-5-1</td>
<td>0.25</td>
<td>0.82</td>
</tr>
<tr>
<td>ANNLMS6</td>
<td>6-5-1</td>
<td>0.48</td>
<td>0.69</td>
</tr>
<tr>
<td>ANNLMS7</td>
<td>7-5-1</td>
<td>0.58</td>
<td>0.62</td>
</tr>
</tbody>
</table>
the REPTree model, the statistical indicators for evaluation of models, i.e., NSE, RSR, PBIAS and $R^2$ were found to be 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

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

<table>
<thead>
<tr>
<th>Model number</th>
<th>ANN structure</th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NSE</td>
<td>RSR</td>
<td>PBIAS</td>
</tr>
<tr>
<td>ANNSCGS1</td>
<td>1-5-1</td>
<td>0.07</td>
<td>0.93</td>
</tr>
<tr>
<td>ANNSCGS2</td>
<td>2-5-1</td>
<td>0.38</td>
<td>0.76</td>
</tr>
<tr>
<td>ANNSCGS3</td>
<td>3-5-1</td>
<td>0.31</td>
<td>0.76</td>
</tr>
<tr>
<td>ANNSCGS4</td>
<td>4-5-1</td>
<td>0.48</td>
<td>0.67</td>
</tr>
<tr>
<td>ANNSCGS5</td>
<td>5-5-1</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>ANNSCGS6</td>
<td>6-5-1</td>
<td>0.21</td>
<td>0.86</td>
</tr>
<tr>
<td>ANNSCGS7</td>
<td>7-5-1</td>
<td>0.23</td>
<td>0.83</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Model number</th>
<th>$\gamma$ (Gamma)</th>
<th>$\sigma^2$ (Sigma square)</th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>NSE</td>
<td>RSR</td>
</tr>
<tr>
<td>LS-SVRS1</td>
<td>1</td>
<td>0.9</td>
<td>0.71</td>
<td>0.54</td>
</tr>
<tr>
<td>LS-SVRS2</td>
<td>2</td>
<td>0.9</td>
<td>0.86</td>
<td>0.38</td>
</tr>
<tr>
<td>LS-SVRS3</td>
<td>3</td>
<td>0.9</td>
<td>0.92</td>
<td>0.29</td>
</tr>
<tr>
<td>LS-SVRS4</td>
<td>4</td>
<td>0.9</td>
<td>0.94</td>
<td>0.24</td>
</tr>
<tr>
<td>LS-SVRS5</td>
<td>5</td>
<td>0.9</td>
<td>0.96</td>
<td>0.20</td>
</tr>
<tr>
<td>LS-SVRS6</td>
<td>6</td>
<td>0.9</td>
<td>0.97</td>
<td>0.17</td>
</tr>
<tr>
<td>LS-SVRS7</td>
<td>7</td>
<td>0.9</td>
<td>0.98</td>
<td>0.15</td>
</tr>
<tr>
<td>LS-SVRS8</td>
<td>8</td>
<td>0.9</td>
<td>0.98</td>
<td>0.00</td>
</tr>
<tr>
<td>LS-SVRS9</td>
<td>9</td>
<td>0.9</td>
<td>0.98</td>
<td>0.00</td>
</tr>
<tr>
<td>LS-SVRS10</td>
<td>10</td>
<td>0.9</td>
<td>0.98</td>
<td>0.00</td>
</tr>
</tbody>
</table>
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 LSSVR S10

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Inputs</th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>M5</td>
<td>( S_t ) or ( D_t = f(R_t, T_t, RH_t, ... ) ( ... E_t, Sun_t, Solar_t, W_t ) )</td>
<td>NSE 0.82, RSR 0.32, PBIAS 15.72, ( R^2 ) 0.89</td>
<td>NSE 0.89, RSR 0.27, PBIAS 26.72, ( R^2 ) 0.95</td>
</tr>
<tr>
<td>REPTree</td>
<td></td>
<td>NSE 0.57, RSR 0.54, PBIAS 23.02, ( R^2 ) 0.73</td>
<td>NSE 0.77, RSR 0.46, PBIAS -28.74, ( R^2 ) 0.83</td>
</tr>
</tbody>
</table>

Figure 9 | Monthly observed and simulated sediment yield using the LS-SVR 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


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


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