Application of artificial neural network, fuzzy logic and decision tree algorithms for modelling of streamflow at Kasol in India

A. R. Senthil kumar, Manish Kumar Goyal, C. S. P. Ojha, R. D. Singh and P. K. Swamee

ABSTRACT

The prediction of streamflow is required in many activities associated with the planning and operation of the components of a water resources system. Soft computing techniques have proven to be an efficient alternative to traditional methods for modelling qualitative and quantitative water resource variables such as streamflow, etc. The focus of this paper is to present the development of models using multiple linear regression (MLR), artificial neural network (ANN), fuzzy logic and decision tree algorithms such as M5 and REPTree for predicting the streamflow at Kasol located at the upstream of Bhakra reservoir in Sutlej basin in northern India. The input vector to the various models using different algorithms was derived considering statistical properties such as auto-correlation function, partial auto-correlation and cross-correlation function of the time series. It was found that REPtree model performed well compared to other soft computing techniques such as MLR, ANN, fuzzy logic, and M5P investigated in this study and the results of the REPTree model indicate that the entire range of streamflow values were simulated fairly well. The performance of the naïve persistence model was compared with other models and the requirement of the development of the naïve persistence model was also analysed by persistence index.

Key words | fuzzy logic, M5, neural networks, REPTree, streamflow

A. R. Senthil kumar
R. D. Singh
National Institute of Hydrology,
Roorkee,
India

Manish Kumar Goyal (corresponding author)
Indian Institute of Technology,
Guwahati 781039,
India
E-mail: *vipmkgoyal@gmail.com*

C. S. P. Ojha
P. K. Swamee
Indian Institute of Technology,
Roorkee,
India

INTRODUCTION

The relationship between rainfall and runoff is a highly non-linear and complex process and its determination is very important for hydrologic engineering design and management purposes. It is dependent on numerous factors such as initial soil moisture, land use, watershed geomorphology, evaporation, infiltration, distribution and rainfall duration, amongst others. Many rainfall-runoff models such as empirical, lumped and distributed models have been developed and used for simulating the streamflow at the catchment outlet. Empirical models estimate the peak runoff from the whole catchment for the purpose of design of storage structures. Lumped models like unit hydrograph (Chow *et al.* 1988) have been developed to estimate the runoff hydrograph from a storm event. Complexity and less accuracy of these conventional models some times force the modeller to think of alternative modelling techniques such as artificial neural network (ANN), fuzzy logic which provide better results. There are many reported applications of ANN in rainfall-runoff modelling (Senthil kumar *et al.* 2005).

Hsu *et al.* (1995) used the linear least square simplex (LLSSIM) procedure for identifying the structure and parameters of three layered feed forward ANN models and demonstrated the potential of ANN models for simulating the non-linear hydrologic behaviour of watersheds. Raman & Sunilkumar (1995) used ANNs for the synthesis of inflows to two reservoirs, Mangalam and Pothundy, located in the Bharathapuzha, Kerala. Danh *et al.* (1999) developed two back propagation neural network models to forecast the daily river flows in two basins in Vietnam and compared the results with the tank model. Tokar & Markus (2000) applied the ANN technique to model watershed runoff in three basins with different climatic and physiographic characteristics and concluded that ANNs could be powerful tools in modelling the rainfall-runoff process for various

doi: 10.2166/wst.2013.491

time scale, topography, and climatic patterns. Zhang & Govindaraju (2000) used a modular neural network structure to handle complex sets of rainfall-runoff data. Different modules within the network were trained to learn subsets of the input space in an expert mode. It was concluded that modular neural networks predicted extreme events of runoff better than the singular neural network models. Sudheer *et al.* (2003) developed a technique to improve peak flow estimation in river flow models. They concluded that the model built on the transformed data outperformed the model built on raw data. The peak flow estimates were improved by data transformation.

Several soft computing learning algorithms have real world applications in image processing, speech processing, control engineering, medicine, classification owing to the advantage of their robustness in noisy environment and flexibility in solving problems (Chiu 1994; Quinlan 1996). Recently, various approaches such as neural network and fuzzy logic have been applied in many areas of water resources due to their capability in representing any nonlinear processes by selecting the significant input vector of the trained networks (Govindaraju & Rao 2000; Kisi *et al.* 2006; Ajmera & Goyal 2012). In this paper, ANN and fuzzy logic have been applied to streamflow modelling at Kasol. A new methodology of the art rule induction and tree algorithms namely, M5 Model Tree and REPTree have also been explored for modelling the streamflow as a part of this paper.

M5 MODEL TREE

Model trees technique provide a structural representation of the data and a piecewise linear fit of the class (Quinlan 1996). They have a conventional decision tree structure but use linear functions at the leaves instead of discrete class labels as shown in Figure 1. Like conventional decision tree learners, M5 builds a tree by splitting the data based on the values of predictive attributes. Instead of selecting attributes by an information theoretic metric, M5 chooses attributes that minimize intra-subset variation in the class values of instances that go down each branch (Goyal & Ojha 2012).

REPTree

REPTree algorithm is a fast decision tree learner. It builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). The algorithm only sorts values for numeric attributes once (Daud & Corne 2007; Senthil kumar *et al.* 2013). REPTree uses standard techniques from C4.5 and classification and regression trees methods (Quinlan 1996).

For further details about M5 model tree and REPTree, one can refer to Witten & Frank (2005). The implementations of various algorithms were carried out in Weka (Witten *et al.* 1999). Weka is written in Java and is freely available from www.cs.waikato.ac.nz/~ml.

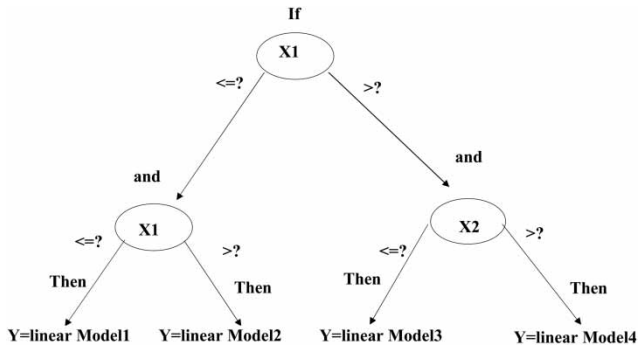

Figure 1 | Model tree.

PERFORMANCE EVALUATION OF VARIOUS MODELS

The whole data length is divided into two based on statistical properties of the time series such as mean and standard deviation, one for calibration (training) and another for validation of the ANN model. The performance during calibration and validation is evaluated by performance indices such as root mean square error (RMSE), model efficiency (EFF) (Nash Sutcliffe efficiency index) and coefficient of correlation (CORR).

STUDY AREA AND DATA AVAILABILITY

The catchment area of the Sutlej river up to Kasol was considered for this study. The study area is located on the upstream of Bhakra reservoir. The catchment area up to Kasol is 56,980 km^2. The location of the reservoir is presented in Figure 2. For the current application, the daily rainfall values for Kalpa, Rampur, Rackchham, Berthin, Bhakra, Kahu, Kasol, Kaza, Namagia, and Suni were available for the years 1987 to 2000. The rainfall values of all the stations have been used. The maximum rainfall values

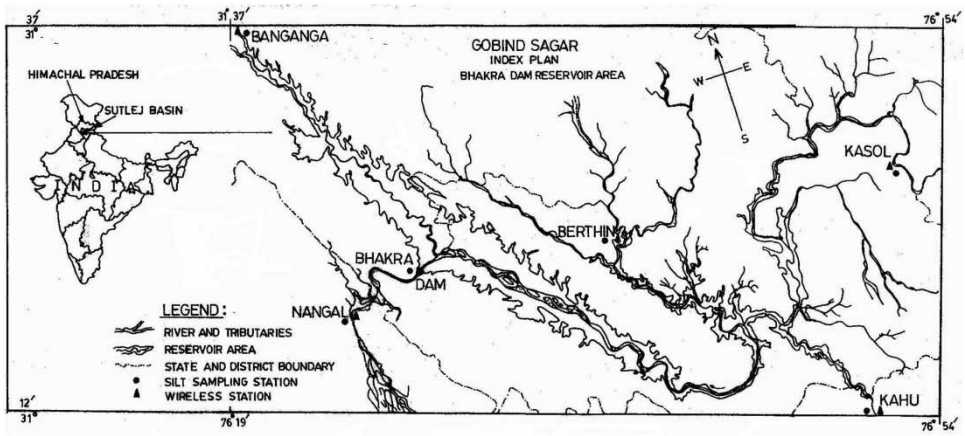

Figure 2 | Index map of Bhakra reservoir.

recorded at Kalpa, Rampur, Rackchham, Berthin, Bhakra, Kahu, Kasol, Kaza, Namagia, and Suni during 1987 to 2000 were 104, 86, 66, 213.2, 281.5, 224, 141, 80, 42.8 and 170 mm respectively. The rainfall stations Berthin, Bhakra and Kahu received heavier rainfall in a single day than other rainfall stations and are located nearer to the reservoir. The rainfall stations Rampur, Suni and Kasol are located along the course of the Sutlej river. The discharge values at Kasol for the same period were also available and the discharge measured on 1 August 2000 was 2689.28 cumecs. From correlation analysis it was learnt that the rainfall values at Kasol, Rampur and Suni and discharge at Kasol had significant influence/impact on the streamflow at Kasol.

MODEL DEVELOPMENT

The input vector to ANN model for the streamflow prediction mainly consists of the antecedent rainfall and discharge of upstream stations (Minns & Hall 1996; Sudheer *et al*. 2003; Srinivasuslu & Jain 2009). The input vector is generally selected by the trial and error. The simple correlation between the dependent and independent variables helps in selecting the significant input vector to the model. Determining the number of antecedent rainfall and discharge values involves finding the lags of rainfall and discharge values that have significant influence on the predicted discharge. The input vector is selected based on the statistical procedure presented by Sudheer *et al*. (2002). They reported that the statistical parameters such as auto-correlation function (ACF), partial auto-correlation function (PACF) and cross-correlation function (CCF) could be used to find out the significant lag values of input variables.

The ACF and PACF of discharge at Kasol as well as the CCF between discharge at Kasol and rainfall values at Kasol, Rampur and Suni is computed (Salas *et al*. 1980). The partial auto-correlation coefficient of discharge at Kasol for lag 1 is 0.98. The cross-correlation coefficients of discharge at Kasol with rainfall at Kasol, Rampur and Suni for lag 0 are 0.38, 0.27 and 0.31 respectively and are higher than all other lagged cross-correlation coefficient values.

Based on the values of PACF and CCF of the data series, the following input vector is selected for neural network training.

$$\text{Inflkas}(t) = f(\text{inflkas}(t-1), \text{rainkas}(t), \text{rainram}(t), \text{rainsun}(t)) \quad (1)$$

in which inflkas is discharge at Kasol, and, rainkas, rainram and rainsun are rainfall at Kasol, Rampur and Suni, respectively.

RESULTS AND DISCUSSION

The ANN models are trained using back propagation algorithms. The whole data set is divided into two sets in order to train and validate the ANN model (Rumelhart *et al*. 1986). The data from 1991 to 2000 are considered for the training of the model since it contained the extreme discharge values. The data from 1987 to 1990 are considered for the validation of the model. The software used for the training of the ANN model for streamflow is MATLAB (The MathWorks, Inc. 2001). The performance of the ANN model during

Table 1 | Results of different models during calibration and validation

		Calibration			Validation		
Models	**Input combinations**	**CORR**	**RMSE**	**EFF%**	**CORR**	**RMSE**	**EFF%**
ANNDIS1(4-1-1)	rainram(t), rainsun(t), rainkas(t), inflkas(t–1)	0.98	61.368	97.66	0.98	69.949	97.26
ANNDIS2(4-2-1)	rainram(t), rainsun(t), rainkas(t), inflkas(t–1)	0.98	59.488	97.80	0.98	68.409	97.38
ANNDIS3(4-3-1)	rainram(t), rainsun(t), rainkas(t), inflkas(t–1)	0.98	56.929	97.99	0.98	78.692	96.53
ANNDIS4(4-4-1)	rainram(t), rainsun(t), rainkas(t), inflkas(t–1)	0.98	56.785	98.00	0.98	74.339	96.90
ANNDIS5(4-5-1)	rainram(t), rainsun(t) rainkas(t), inflkas(t–1)	0.99	56.142	98.04	0.98	74.604	96.88
ANNDIS6(4-6-1)	rainram(t), rainsun(t), rainkas(t), inflkas(t–1)	0.99	55.705	98.07	0.98	78.261	96.57
ANNDIS7(4-7-1)	rainram(t), rainsun(t), rainkas(t), inflkas(t–1)	0.99	55.332	98.10	0.98	75.603	96.80
MLRM1	rainram(t), rainsun(t), rainkas(t), inflkas(t–1)	0.98	63.86	97.47	0.98	73.53	96.97
FLM1	rainram(t), rainsun(t), rainkas(t), inflkas(t–1)	0.99	61.75	97.63	0.99	70.03	97.25
M5M1	rainram(t), rainsun(t), rainkas(t), inflkas(t–1)	0.99	59.74	97.78	0.99	70.55	97.21
REPTreeM1	rainram(t), rainsun(t), rainkas(t), inflkas(t–1)	0.99	56.67	98.00	0.98	76.89	96.69

calibration and validation with the input combination derived from statistical procedure given by Sudheer *et al.* (2002) is given in Table 1. The number of neurons in the hidden layer is found by trial and error as in the previous case. The model ANNDIS2 performed better than other ANN models during calibration (CORR = 0.98, RMSE = 59.488, EFF = 97.80%) and validation (CORR = 0.98, RMSE = 68.409, EFF = 97.38%) and the optimum structure of the ANN model is found to be two neurons in the hidden layer. A multiple linear regression (MLR) model is developed for the prediction of discharge using the same data set and input vector considered in the development of ANN model. The MLR model yields the following equation

$$\mathrm{INFL}_{\mathrm{Kasol},t} = 1.41\mathrm{R}_{\mathrm{Rampur},t} + 1.18\mathrm{R}_{\mathrm{Suni},t} + 0.63\mathrm{R}_{\mathrm{Kasol},t} + 0.97\mathrm{INFL}_{\mathrm{Kasol},t-1} + 4.40 \quad (2)$$

The values of the performance criteria from various models for both calibration and validation are presented in Table 1. The calibration and validation results are compared with the performance indices of best models.

CALIBRATION/TRAINING RESULTS

It can be inferred from Table 1 that the model REPTreeM1 using the REPTree algorithm outperformed all other models explored in this study. In terms of RMSE, model MLRM1using MLR performed worst (63.86) and REPTreeM1 model performed the best (56.67). The CORR for the REPTreeM1 model was 0.99, whereas the value of the CORR for the ANN with back propagation model ANNDS12 was 0.98 and similar value was observed for MLRM1 also. In terms of percentage error in peak flow estimation, the REPTree model performed best (−11.90%) while all other model performed in the range of 37–42%.

VALIDATION/TESTING RESULTS

Model ANNDIS2 using ANN outperformed all other models investigated in this study in terms of RMSE. RMSE was 68.40 for ANNDIS2 while the model REPTreeM1 performed worst (76.89). The best value of correlation coefficient was 0.99 from the FLM1 and M5M1 model

while the other model REPTreeM2 performed equally well (0.98). In term of efficiency, all the models performed well (~97%). ANN model ANNDIS2 performed best for percentage error in peak discharge (−13.46%) while M5 model performed worst (−25.73%).

The observed and predicted streamflow from the ANN, MLRs, fuzzy logic, M5 and REPTree models for calibration and validation are shown in Figure 3. The graphical results also indicate that the REPTree models perform a lot better than the other models investigated in this study. The scatter plot of the results of REPTree during validation is shown in Figure 4. The persistence model for the simulation of discharge is also developed and is given as

$$Q_t = 0.99Q_{t-1} \tag{3}$$

The persistence index is computed for both the calibration and validation of the model using the following equation

$$\text{PI} = 1 - \frac{\sum_{t=1}^{n}(Yo_t - Yc_t)^2}{\sum(Yo_{t-1} - Yo_t)^2} \tag{4}$$

where Yo, Yc are observed and computed discharges respectively. The naïve persistence model is very good if PI is equal to 1. The development of naïve persistence model is not at all required if PI is equal to −1. The performance of the naïve persistence model during calibration and validation is given in Table 2. The PI during calibration and validation are 0.00381 and 0.0033. This clearly indicates that the performance of naïve persistence model is not comparable with the other models such as ANN, MLR, fuzzy logic, M5P and REPTree.

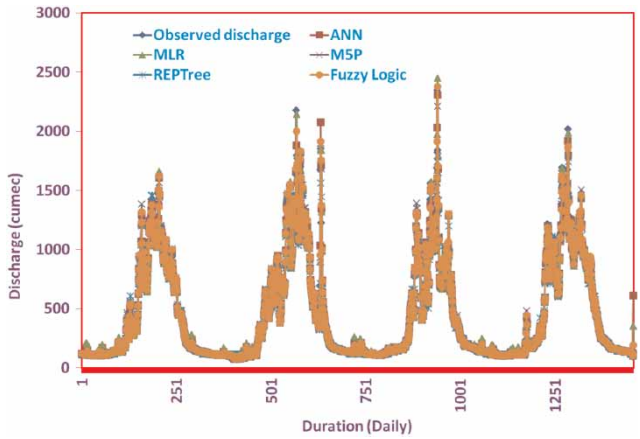

Figure 3 | The results of ANN, MLR, fuzzy logic, M5P, REPTree during validation.

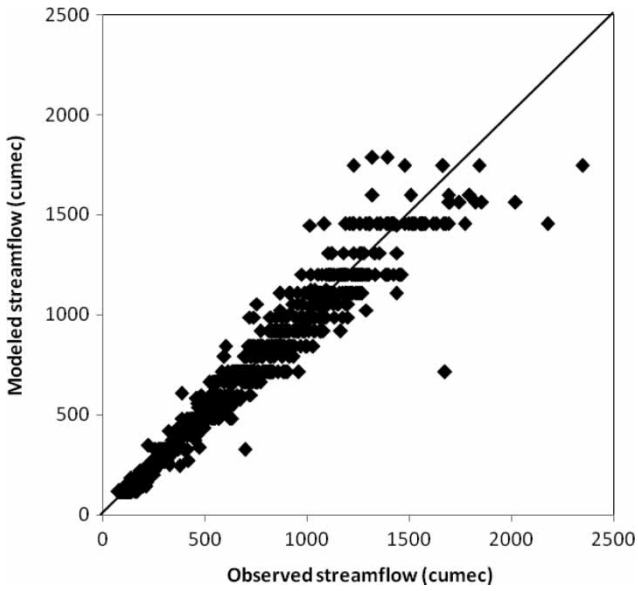

Figure 4 | Scatter plot for the result of REPTree model during validation.

SUMMARY AND CONCLUSIONS

This study suggests a better model on comparison of performance of the various modelling techniques for the purpose of modelling of the streamflow at Kasol gauging site of Sutlej river basin in Northern India. The techniques investigated include MLR, ANN with back propagation algorithm, fuzzy logic and decision tree algorithms such as M5 and REPTree. The statistical parameters ACF, PACF and CCF had been used for selection of input vector to the various models. The data corresponding to 1987 to 2000 were employed to calibrate and validate all model structures. The performance of each model structure was evaluated using common performance criteria. The naïve persistence model was developed and compared with the results of other models.

Table 2 | Results of persistence model during calibration and validation

	Calibration				Validation			
Models	**CORR**	**RMSE**	**EFF%**	**Persistence index**	**CORR**	**RMSE**	**EFF%**	**Persistence index**
PI	0.98	76.08	96.75	0.00381	0.98	67.58	97.16	0.0033

Persistence indices during calibration and validation indicates that the development of naïve persistence model is not required.

The results obtained in this study have been able to demonstrate that: (1) REPTree model was able to consistently outperform in terms of performance criteria barring few exceptions; (2) REPTree modelling approach provide better insight into the developed models; and (3) a less computational time by REPTree model compared to the other techniques reveals that training of REPTree model is much faster.

REFERENCES

Ajmera Tapesh, K. & Goyal Manish Kumar 2012 Development of stage discharge rating curve using model tree and neural networks: an application to Peachtree Creek in Atlanta. *Expert Systems With Applications* **39** (5), 5702–5710.

Chiu, S. 1994 Fuzzy model identification based on cluster estimation. *J. Intell. Fuzzy Sys.* **2** (3), 267–278.

Chow, V. T., Maidment, D. R. & Mays, L. W. 1988 *Applied Hydrology*. McGraw-Hill Book Company, New York, USA.

Danh, N. T., Phien, H. N. & Gupta, A. D. 1999 Neural network models for river flow forecasting. *Water SA* **25** (1), 33–39.

Daud, M. N. R. & Corne, D. W. 2007 Human readable rule induction in medical data mining: a survey of existing algorithms. In: *Proceedings of WSEAS European Computing Conference*, Athens, Greece.

Govindaraju, R. S. & Rao, A. R. 2000 *Artificial Neural Networks in Hydrology*. Kluwer Academic, Dordrecht, The Netherlands.

Goyal Manish Kumar & 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.

Hsu, K.-L., Gupta, H. V. & Sorooshian, S. 1995 Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res.* **31** (10), 2517–2530.

Kisi, O., Karahan, M. E. & Sen, Z. 2006 River suspended sediment modelling using a fuzzy logic approach. *Hydrol. Process*. **20**, 4351–4362.

Minns, A. W. & Hall, M. J. 1996 Artificial neural networks as rainfall runoff models. *Hydrol. Sci. J.* **41** (3), 399–418.

Quinlan, J. R. 1996 Improved use of continuous attributes in C4.5. *J. Artif., Intell. Res* **4**, 77–90.

Raman, H. & Sunilkumar, N. 1995 Multivariate modelling of water resources time series using artificial neural networks. *Hydrol. Sci. J.* **40** (2), 145–163.

Rumelhart, D., Hinton, E. & Williams, J. 1986 *Learning Internal Representation by Error Propagation, Parallel Distributed Processing, Vol. 1*, MIT Press, Cambridge, Mass, pp. 318–362.

Salas, J. D., Delleur, J. W., Yevjevich, V. & Lane, W. L. 1980 *Applied Modelling of Hydrologic Time Series*. Water Resources Publications, P.O. Box 2841, Littleton, Colorado 80161, USA.

Senthil kumar, A. R., Sudheer, K. P., Jain, S. K. & Agarwal, P. K. 2005 Rainfall-runoff modelling using artificial neural networks: comparison of network types. *Hydrol. Process.* **19**, 1277–1291.

Senthil kumar, A. R., Goyal Manish Kumar, Ojha, C. S. P., Singh, R. D., Swamee, P. K. & Nema, R. K. 2013 Application of ANN, fuzzy logic and decision tree algorithms for the development of reservoir operating rules. *Water Resources Management*, **27** (3), 911–925.

Srinivasulu, S. & Jain, A. 2009 River flow prediction using an integrated approach. *J. Hydrol. Eng., ASCE* **14** (1), 75–83.

Sudheer, K. P., Gosain, A. K. & Ramasastri, K. S. 2002 A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrol. Process* **16**, 1325–1330.

Sudheer, K. P., Nayak, P. C. & Ramasastri, K. S. 2003 Improving peak flow estimates in artificial neural network river flow models. *Hydrolog. Process* **17** (3), 677–686.

The MathWorks, Inc. 2001 *ANN Toolbox User's Guide*. 3 Apple Hill Drive, Natick, MA 01760–2098.

Tokar, A. S. & Markus, M. 2000 Precipitation-runoff modelling using artificial neural networks and conceptual models. *J. Hydrol. Eng.* **5** (2), 156–161.

Witten, I. H. & Frank, E. 2005 *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn, Morgan Kaufmann, San Francisco.

Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S. J. 1999 Weka: practical machine learning tools and techniques with java implementations. In: *Emerging Knowledge Engineering and Connectionist-Based Information Systems*, pp. 192–196.

Zhang, B. & Govindaraju, R.S. 2000 Prediction of watershed runoff using Bayesian concepts and modular neural networks. *Water Resour. Res.* **36** (3), 753–762.

First received 20 April 2013; accepted in revised form 31 July 2013. Available online 24 October 2013