Water Management

Modelling runoff and sediment rate using a neuro-fuzzy technique
Nayak and Jain

ice | proceedings

Proceedings of the Institution of Civil Engineers

doi: 10.1680/wama.900083
Paper 900083
Received 11/12/2009 Accepted 27/10/2010
Keywords: hydraulics & hydrodynamics/mathematical modelling/river engineering

ice
Institution of Civil Engineers
publishing

Modelling runoff and sediment rate using a neuro-fuzzy technique

Purna Chandra Nayak PhD
Scientist, National Institute of Hydrology Deltaic Regional Centre, Kakinada, India

Sharad Kumar Jain PhD
NEEPCO Chair Professor, Department of Water Resources Development & Management, Indian Institute of Technology, Roorkee, India

This paper demonstrates the estimation and simulation of discharge and sediment concentration for two river basins in the USA and India. The first-order Sugeno fuzzy inference system was utilised to model the stage, discharge and sediment concentration relationship. A subtractive clustering algorithm, along with a least-squares estimation, was used to generate the fuzzy rules that describe the relationship between input and output data of stage, discharge and sediment concentration, which change over time. The fuzzy rules were tuned by a back-propagation algorithm. The results are illustrated using simulation and virtual reality. A comparison was made between the estimates provided by the neuro-fuzzy model and a multi-linear regression model. Different statistical criteria were used to evaluate the performance of both models in estimating discharge and sediment concentration. Comparison of the results reveals that, in general, the neuro-fuzzy model gives better estimates than the multi-linear regression model in terms of root mean square and sum of squares errors. Furthermore, compared with the multi-linear regression model, the neuro-fuzzy model yields statistical properties of estimates that are closer to actual historical data.

1. Introduction

The catchments of many large rivers (e.g. Yellow, Mississippi, Amazon, Ganga and Brahmaputra) are facing a range of problems due to significant soil erosion, transport and deposition. Estimates of watershed sediment yield are required to solve a number of problems in hydrologic and hydraulic engineering, such as design of dams and reservoirs, transport of pollutants in rivers, soil conservation practices, design of stable channels, debris basins and bridge foundations, operation of hydropower plants and environmental impact assessment. Obviously, correct assessment of sediment transported by a river is of vital interest in hydrologic engineering.

A number of sediment yield models have been developed to address wide-ranging soil and water resources problems. Sediment rating curves are widely used to estimate sediment loads transported by rivers (Jain, 2001). The most important aspect in developing a transport model is the development of a basis for scaling, or representing, the discharge Q. Since we often wish to develop a sediment rating curve, the obvious thing to try is to develop a model based directly on Q. However, just a little thought shows that such a model could not possibly be general. It seems quite unlikely that, for example, 50 m^3/s of water could produce the same transport rate in a small creek that one could easily jump across and in a very large river that might be a kilometre wide.

Differences in channel size, shape, slope, roughness and bed material composition would produce very different sediment rates Q_s for the same Q. This implies that very different values of coefficients (fitting parameters) would be needed to estimate Q_s in different rivers assuming that a power function correctly describes sediment transport. Of course, this assumption may not always be valid.

Similarly, the measure of flow strength that has been found to provide a general description of transport rate is the bed shear stress τ. A stress is a force per area – in this case, the shear force exerted by the flowing water on an area of the bed. That the transport should depend on the fluid force applied to the bed should, hopefully, seem reasonable. The price we pay for using τ is that we have to figure out how to estimate it, which is a very difficult task.

Karim and Kennedy (1990) attempted to establish relations among the velocity, sediment discharge, bed-form geometry and friction factor of alluvial rivers. Crawford (1991) reported that a weighted non-linear least-squares method can be used to improve parameter estimates in a non-linear model for suspended sediment estimation. Lopes and Ffolliott (1993) pointed out that additional complexity is introduced into the sediment concentration and streamflow relationship due to a hysteresis effect. Sumathi and Santhana Bosu (2002) used a geographic information system and a modified universal soil loss equation to estimate sediment yield.

2. Data-driven modelling techniques for sediment modelling

Although physics-based distributed models provide reasonable accuracy in hydrologic assessments and predictions, a wider implementation of such models can typically present certain difficulties (Duan *et al.*, 1992). Due to the use of idealised sediment components, these models require sophisticated mathematical tools, a large volume of detailed spatial and temporal data pertaining to environment and water resources, and a certain degree of expertise and experience with the model. Clearly, all these are not always possible (Cigizoglu and Kisi, 2005; Kisi, 2005). Simpler approaches are therefore required in the form of a 'black-box' modelling approach, which falls in the general domain of data-driven modelling (DDM).

Various artificial intelligence (AI) techniques are extensively used for DDM. AI techniques are a consortium of methodologies that work synergistically to provide real-world computing capabilities. The techniques under this paradigm exploit the tolerance for imprecision, uncertainty, approximate reasoning and partial truth. The most commonly employed tools in AI are fuzzy sets, artificial neural networks (ANNs) and genetic algorithms, which offer the opportunity to analyse a wide variety of data types by handling real-life uncertainty with low-cost solutions.

The AI applications in suspended sediment modelling are relatively new compared with other water resources domains. Jain (2001) applied an ANN and a conventional curve-fitting approach to estimate sediment rate and found that the ANN results were much closer to the observed values than the curve-fitting results. Cigizoglu (2004) showed that feed-forward back-propagation models could provide negative suspended sediment estimations for some observed low sediment values. Kisi (2005) first reported the application of a neuro-fuzzy (NF) model for estimation of suspended sediment, and found that NF and ANN models performed better than the regression and rating curve techniques in this application. He also found that the NF model was more flexible than the considered ANN model, with more options of incorporating the fuzzy nature of a real-world system. Cigizoglu and Kisi (2005) reported the use of a range-dependent neural network (RDNN) and found it to be superior to conventional ANN applications. Azamathullah *et al.* (2009) reported on the use of adaptive neuro-fuzzy inference systems (Anfis) for bed load prediction.

While developing the NF model, Kisi (2005) used grid partitioning to generate fuzzy rules. Grid partitioning refers to dividing the input space into rectangular grids, and the fuzzy rules are confined to the corners of the grid (Brown and Harris, 1994). There is, however, a serious disadvantage of the grid partitioning approach: the very regular partition of the input space may be unable to produce a rule set of acceptable size that is able to optimally handle a given dataset. If, for example, the dataset contains regions with several small clusters of different classes, then small rule patches have to be created to correctly classify the data in this region. Due to the grid partition, however, the fine resolution needed in this particular area is also propagated to areas that are much easier to handle, perhaps because they contain data belonging to a single class only. The grid partitioning approach enforces a large number of small identical rule patches, although one large patch would theoretically be able to correctly classify the data in this region. Effective partition of the input space can decrease the number of rules and thus increase speed in both the learning and application phases.

To eliminate the problems associated with grid partitioning, scatter partitioning of the input space into rule patches has been proposed (Chiu, 1994; Setnes, 2000). In this type of partitioning, the antecedent parts of the fuzzy rules are positioned at arbitrary locations in the input space. This means that rules are no longer confined to the corners of a rectangular grid, but can be freely chosen (e.g. by a clustering algorithm working on the training data). An appealing feature of the clustering approach for partitioning is simultaneous identification of the antecedent membership functions, which are attributed to a minimum number of fuzzy rules from the dataset.

In this investigation the subtractive clustering technique is used for generation of fuzzy rules from the input/output data and the back-propagation algorithm is used for optimisation of membership function parameters in the NF model. The developed model is used to establish an integrated stage–discharge–sediment concentration relation for river sites. Based on comparison of the results for two gauging sites, the NF model results are much closer to the observed values than those produced by the conventional technique.

3. NF model development

The hybrid intelligent system proposed in this paper is a combination of the Takagi–Sugeno (TS) fuzzy model with an ANN. Various types of fuzzy rule-based models have been proposed (e.g. Mamdani and Assilan, 1975; Takagi and Sugeno, 1985; Tsukamoto, 1979) and each is characterised by their consequent function only. The TS fuzzy model resulted from an effort to develop a systematic approach to generate fuzzy rules from a given input–output dataset (Sugeno and Kang, 1988; Takagi and Sugeno, 1985) in which the rule consequents are typically taken to be either crisp numbers or linear functions of the inputs. The first-order TS model is described as follows.

Consider a function $y = f(\boldsymbol{x})$ being mapped by the TS model, in which y is the dependent variable and $\boldsymbol{x}$ is the (k-dimensional) column vector of independent variables that have a causal relationship with y. Assume that n example (pattern) pairs $[\boldsymbol{x}, y]$ are available for parameter estimation. Considering m rules, the mathematical functioning of the TS model is

$$\begin{aligned} &\boldsymbol{R}_i\text{: If } x_1 \text{ is } A_{i,1} \text{ AND } \ldots \text{ AND } x_k \text{ is } A_{i,k} \\ &\qquad \text{THEN } y_i = \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} + b_i \end{aligned} \tag{1}$$

where $\boldsymbol{x} \in \Re^k$ are the input variables (antecedents), $y_i \in \Re$ is the output (consequent) of the ith rule R_i, and $\boldsymbol{a}_i$ and b_i are the parameters of the consequent model. $\boldsymbol{a}_i^{\mathrm{T}}$ is the transpose the matrix $\boldsymbol{a}_i$ and A_i is the membership function (such as a Gaussian), which has the form

$$A(x_k) = \exp \frac{1}{2}\left(\frac{x_k - c_i}{\sigma_i}\right)^2 \tag{2}$$

where $\{c_i, \sigma_i\}$ is the parameter set, called the centre and spread function. These parameters (called premise parameters or antecedent parameters) – with maximum equal to 1 and minimum equal to 0 – determine the shape of the membership function. A_i is the antecedent fuzzy set (membership function) of the ith rule such that

$$A_i(x_k)\colon \Re^k \rightarrow [0, 1] \qquad i = 1\colon m \tag{3}$$

In the case of univariate membership functions $\mu_{ij}(x_j)$, the fuzzy antecedent in the TS model is typically defined as an AND conjunction by means of the product operator

$$A_i(x_k) = \prod_{j=1}^{k} \mu_{ij}(x_k) \tag{4}$$

For the nth input patterns $\boldsymbol{x}_n$, the total output $\hat{y}(n)$ of the model is computed by aggregating the contribution of individual rules $y_i(n)$

$$\hat{y}(n) = \sum_{i=1}^{m} u_{ni} y_i(n) \tag{5}$$

where $\hat{y}(n)$ is the estimated output for the pattern x_n and u_{ni} is the normalised degree of fulfilment of the antecedent clause of rule $\boldsymbol{R}_i$, defined as

$$u_{ni} = \frac{A_i(x_n)}{\sum_{i=1}^{m} A_i(x_n)} \tag{6}$$

In this way, a weighted average of the individual rule outputs is computed and a non-linear function can be approximated.

The basic structure of a fuzzy system consists of three conceptual components

(*a*) a rule base, which contains a selection of fuzzy rules

(*b*) a database that defines the membership function used in the fuzzy rules

(*c*) a reasoning mechanism, which performs the inference procedure upon the rules and a given condition to derive a reasonable output conclusion (Nayak *et al.*, 2005).

A fuzzy inference system (FIS) implements a non-linear mapping from its input space to an output space. A FIS can utilise human expertise by storing its essential components in a rule base and database, and performs fuzzy reasoning to infer the overall output value. The derivation of if–then rules and corresponding membership functions depends heavily on the a priori knowledge about the system under consideration.

There is no systematic way to transform the knowledge and experience of human experts into the knowledge base of a FIS. Recall that the ANN learning mechanisms do not rely on human expertise. Due to the highly parallel structure of an ANN, it is hard to extract structured knowledge from either the weights or the architecture of the ANN, although many efforts have been made in this direction (e.g. Jain *et al.*, 2008). The weights of the ANN represent the coefficients of the hyper-plane that partitions the input space into two regions with different output values. If one can visualise the hyper-plane structure from the training data, then the subsequent learning procedures in an ANN can be reduced. Conversely, a priori knowledge is usually obtained from human experts and it is most appropriate to express the knowledge as a set of fuzzy if–then rules. The limitations that arise when these techniques (ANN and FIS) are individually used can be addressed by creating hybrid systems that combine both techniques. A common way to integrate an FIS with an ANN is to embed the FIS in a general ANN architecture to create an adaptive neural network and to use the learning algorithms of the ANN to estimate membership function parameters (Jang, 1993). However, conventional ANN learning algorithms (e.g. gradient descent) cannot be directly applied to such a system because the transfer function of the FIS need not usually be 'non-differentiable'. This problem can be tackled by using differentiable functions in the inference system or by not using the standard neural learning algorithm.

4. Cluster estimation

The subtractive clustering method (Chiu, 1994) is an extension of the mountain clustering method (Yager and Filev, 1994) where the potential is calculated for the data rather than grid points defined on the data space. Each historical data point is considered a potential cluster centre. A measure of the potential of a data point x_i as a cluster centre is defined as

$$p_i = \sum_{j=1}^{n} \mathrm{e}^{-\alpha \| x_i - x_j \| 2} \tag{7}$$

where

8. $$\alpha = 4/\gamma_a^2$$

and γ_a is a positive constant. Thus, the measure of potential for a data point is a function of its distance from all other data points. The data point with the highest potential is selected as the first cluster centre. Let x_1^* be the location of the first cluster centre and p_1^* its potential value. Before proceeding to find the second cluster centre, the amount of potential from the next data point is subtracted based on its distance from the first cluster centre. Then the potential of each data point x_i may be revised as

9. $$p_i = p_i - p_1^* \mathrm{e}^{-\beta \|x_i - x_1^*\|^2}$$

where

10. $$\beta = 4/\gamma_b^2$$

and γ_b is a positive constant, which is effectively the radius that defines the neighbourhood that will have measurable reductions in the potential of the other data points. To avoid obtaining closely spaced cluster centres, γ_b may be set to be somewhat greater than γ_a; a good choice is $\gamma_b = 1{\cdot}5\gamma_a$.

When the potentials of all the data points have been revised according to Equation 9, the algorithm then selects the data point with the highest remaining potential as the second cluster centre. The process is repeated until a given threshold for the potential is obtained such that $p_k^*/p_1^* < \bar{\varepsilon}$. The choice of $\bar{\varepsilon}$ is an important factor affecting the results; if $\bar{\varepsilon}$ is too large too few data points will be accepted as cluster centres and if $\bar{\varepsilon}$ is too small, too many cluster centres will be generated. Figure 1 shows the criteria for accepting or rejecting cluster centres. In the present investigation, $\bar{\varepsilon} = 0{\cdot}5$, as suggested by Chiu (1994), is considered for analysis.

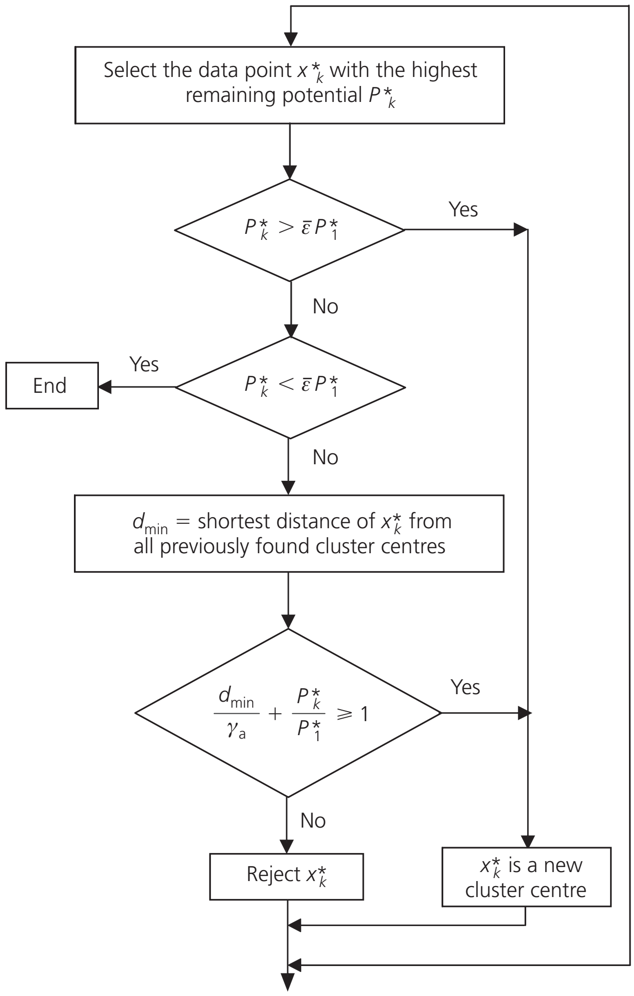

Figure 1. Flow chart of the criteria for accepting cluster centres

5. Case studies

The proposed approach requires uninterrupted time series of measured data pertaining to river stage, discharge and sediment concentration at a location. The data should be of sufficient length to obtain stable estimates of the parameters. The current investigation considered data pertaining to Thebes gauging station (United States Geological Survey (USGS) station No. 07022000), which is located in the state of Illinois and is operated by the USGS. The drainage area at the Thebes site is 1 847 190 km^2. Daily time series of river stage, discharge and sediment concentration for this station were used by Jain (2001). After examining the data and noting the periods in which there were gaps in one or more of the three variables, periods for calibration and validation were chosen: data for 1 January to 30 September 1990 were used for calibration and data for 15 January to 10 August 1991 were used for validation. It may be noted that the periods from which calibration and validation data were chosen for the Thebes site span approximately the same temporal seasons (January–September and January–August).

A second basin – Indravati basin, India – was chosen for comparison. Daily discharge and sediment concentration data from 1994 to 2006 for Pathagudem and Nawarangpur gauging sites (Figure 2) were employed for the current analysis. The number of input variables was determined using standard statistical method reported by Sudheer *et al.* (2002). Data from 1 March 1994 to 31 December 2000 were used for model calibration and data from 1 January 2001 to 31 May 2006 were used for validation of the model parameters.

6. Model application

The variables were scaled to limit between 0 and 1, as the activation function warrants. All available data were divided into a calibration set and a validation set as described earlier. Model parameters were identified using the calibration dataset, and the model was tested for its performance on the validation dataset. The combinations of input data of stage, discharge and sediment

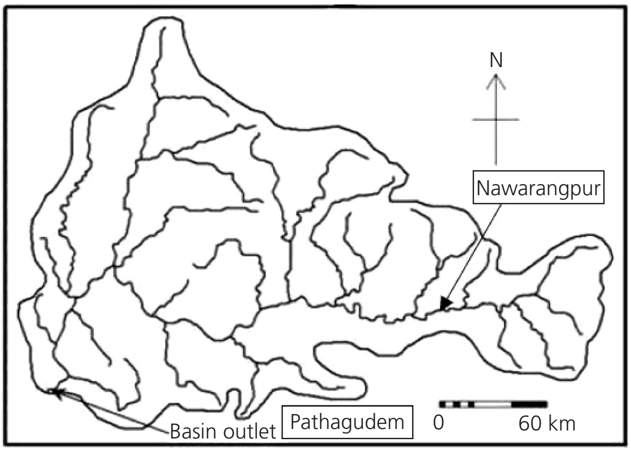

Figure 2. Indravati Basin, India

concentration data evaluated in the present study are listed in Table 1.

The main parameter that needs to be identified in the NF model is the clustering radius. (Note that the radius specifies the range of influence of the cluster centre of the each input and output dimension.) Assuming that the cluster radius falls within a hyperbox of unit dimensions, a smaller cluster radius will yield more cluster in a data and hence a greater number of rules. Simultaneously, it increases model complexity and decreases parsimony. The clustering radius was identified through a trial-and-error procedure by varying the clustering radius from 0·1 to 1·0 in increments of 0·05 for the developed models with different inputs as shown in Table 1. For each radius, the variance explained by the NF model was computed. The model that explained the maximum variance was selected for further analysis. The best models obtained for different combinations of input vectors for the Thebes station with changing clustering radius and their optimal fuzzy if–then rules are presented in Table 2.

To determine the optimal model for estimating discharge and sediment concentration, different statistical indices were considered – the coefficient of correlation (Corr), the coefficient of efficiency (Eff) (Nash and Sutcliffe, 1970), root mean square error (RMSE) between the observed and forecasted values and the sum of squares error (SSE).

	Inputs
Model 1	H_t, H_{t-1}, S_{t-1} and Q_{t-1}
Model 2	H_t, H_{t-1}, H_{t-2}, S_{t-1} and Q_{t-1}
Model 3	H_t, H_{t-1}, H_{t-2}, S_{t-1}, Q_{t-2} and Q_{t-1}
Model 4	H_t, H_{t-1}, H_{t-2}, S_{t-1}, S_{t-2}, Q_{t-1} and Q_{t-2}
Model 5	H_t, H_{t-1}, S_{t-1}, S_{t-1}, Q_{t-1} and Q_{t-2}

Table 1. Different combinations of input data (H is stage, S is sediment concentration, Q is discharge and t is time)

	Clustering radius	No. of fuzzy if–then rules
Discharge		
Model 1	0·9	3
Model 2	0·7	3
Model 3	0·4	6
Model 4	0·4	5
Model 5	0·4	5
Sediment		
Model 1	0·5	4
Model 2	0·3	6
Model 3	0·4	5
Model 4	0·6	4
Model 5	0·4	5

Table 2. Number of if–then rules for different models

11.
$$\text{Corr} = \left\{ \frac{\sum_{t=1}^{n} (y_t^{\text{o}} - \bar{y}^{\text{o}})(y_t^{\text{c}} - \bar{y}^{\text{c}})}{\left[\sum_{t=1}^{n} (y_t^{\text{o}} - \bar{y}^{\text{o}})^2\right]^{1/2} \left[\sum_{t=1}^{n} (y_t^{\text{c}} - \bar{y}^{\text{c}})^2\right]^{1/2}} \right\}$$

12.
$$\text{Eff} = 1 - \frac{\sum_{t=1}^{n} (y_t^{\text{o}} - y_t^{\text{c}})^2}{\sum_{t=1}^{n} (y_t^{\text{o}} - \bar{y}^{\text{o}})^2}$$

13.
$$\text{RMSE} = \left[\frac{\sum_{t=1}^{n} (y_t^{\text{o}} - y_t^{\text{c}})^2}{n} \right]^{1/2}$$

where y_t^{o} and y_t^{c} are the observed and computed values at time t and $\bar{y}^{\text{o}}$ and $\bar{y}^{\text{c}}$ are the mean of the observed and computed values corresponding to n patterns, respectively.

7. Result and discussions

The values of the performance indices in estimating discharge and sediment concentration are presented in Table 3 for both the calibration and validation periods. The correlation statistics, which evaluate the linear correlation between the observed and computed values, are consistent (>0·923) for all models during calibration and validation in estimating sediment concentration and discharge. The model efficiency, which measures the capability of the model to explain variance in the data, is acceptable (99% for discharge computation and >80% for sediment concen-

	Corr		Eff: %		RMSE: mg/l		SSE	
	CL	VL	CL	VL	CL	VL	CL	VL
Discharge								
Model 1	0·998	0·999	99·590	99·680	249·620	151·169	$1{\cdot}689 \times 10^7$	$4{\cdot}708 \times 10^6$
Model 2	0·998	0·999	99·640	99·690	232·539	148·050	$1{\cdot}494 \times 10^7$	$4{\cdot}515 \times 10^6$
Model 3	0·998	0·998	99·660	99·640	225·626	158·443	$1{\cdot}380 \times 10^7$	$5{\cdot}172 \times 10^6$
Model 4	0·998	0·998	99·660	99·630	226·634	162·287	$1{\cdot}392 \times 10^7$	$5{\cdot}425 \times 10^6$
Model 5	0·998	0·998	99·660	99·490	226·524	189·571	$1{\cdot}391 \times 10^7$	$7{\cdot}403 \times 10^6$
Sediment								
Model 1	0·924	0·923	85·330	81·770	144·128	145·501	$5{\cdot}629 \times 10^6$	$4{\cdot}361 \times 10^6$
Model 2	0·933	0·931	87·000	83·650	135·681	127·806	$4{\cdot}989 \times 10^6$	$2{\cdot}912 \times 10^6$
Model 3	0·927	0·939	85·840	87·860	141·614	118·716	$5{\cdot}435 \times 10^6$	$2{\cdot}903 \times 10^6$
Model 4	0·936	0·928	87·680	81·980	132·050	144·641	$4{\cdot}726 \times 10^6$	$4{\cdot}310 \times 10^6$
Model 5	0·929	0·945	86·370	89·200	138·915	111·966	$5{\cdot}130 \times 10^6$	$2{\cdot}583 \times 10^6$

Table 3. Statistical performance indices for calibration (CL) and validation (VL) for different models

tration) for all models; according to Shamseldin (1997) these values are very satisfactory. The RMSE, which is a quantitative measure of the model error in units of the variable, was good for all the models. Overall, all the models show similar performance in estimating sediment concentration and discharge (Table 3).

To determine the combination of input variables that best preserves the statistical properties of the historical data, various statistical parameters of the observed and simulated responses were computed. The results, shown in Table 4, indicate that all the models are similar in performance but model 2 best preserves the statistical properties of the historic sediment and discharge data. Table 4 also shows that the skewness of the output data produced by model 2 is very close to that of the observed series and the model is able to maintain the statistical properties closely while estimating sediment transport rate and discharge. This observation is significant as the skewness of the data plays an important role in overall model performance (Sudheer *et al.*, 2003).

The variations of observed and computed discharge and sediment concentration over time are presented in Figures 3 and 4, respectively. Figure 3 shows that the discharge hydrograph produced by the NF model matches the observed hydrograph very closely. The match between observed and computed graphs of sediment concentration data is also very close, as shown in Figure 4.

Jain (2001) used an ANN to develop a rating relation for the Thebes site using the same data. For the training data, the best values of the correlation coefficient and sum of squares error (SSE) for the discharge obtained by Jain were 0·993 and $1{\cdot}194 \times 10^7$; for the sediment concentration, the best values were 0·961 and $3{\cdot}019 \times 10^6$. Likewise, for the training data, the best values of correlation and SSE for the discharge data were 0·995

	Model 1	Model 2	Model 3	Model 4	Model 5	Observed
Discharge						
Skewness	0·22	0·23	0·22	0·21	0·19	0·25
Average	7568·59	7582·03	7561·72	7573·45	7542·91	7586·21
Std dev.	2622·15	2636·28	2630·08	2627·14	2634·85	2658·17
Sediment						
Skewness	1·39	1·24	1·02	1·35	1·07	1·20
Average	495·78	486·98	481·69	496·75	473·95	465·05
Std dev.	370·42	373·19	325·31	378·17	320·84	341·59

Table 4. Summary statistics of the different models

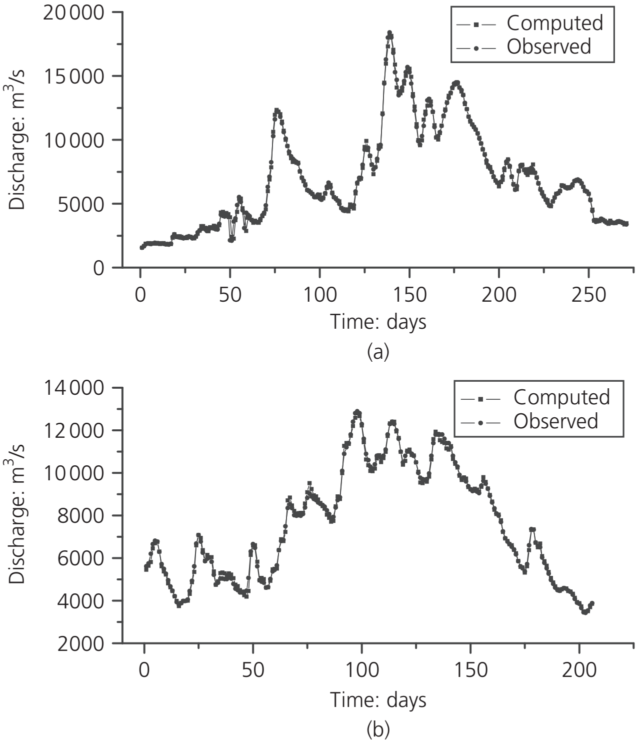

Figure 3. Plots for observed and computed discharges (using NF model) during (a) calibration and (b) validation

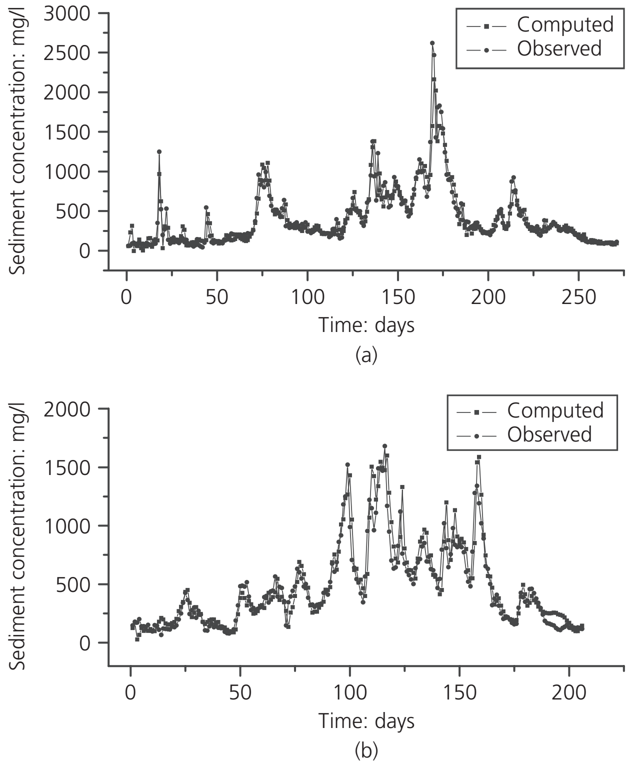

Figure 4. Plots for observed and computed sediment concentration during (a) calibration and (b) validation

and $1{\cdot}747 \times 10^7$, and 0·926 and $4{\cdot}245 \times 10^6$ for sediment concentration, respectively. The results obtained in this study show much improvement over these results.

To compare the performance of the developed NF model over multiple linear regression (MLR), an MLR model was developed for the same data using the same input variables as model 2. The MLR model was

14. $$X_t = aH_t + bH_{t-1} + cH_{t-2} + dS_{t-1} + eQ_{t-1}$$

where X_t is the variable to be computed (either discharge or sediment transport). The performance indices for the MLR model in estimating discharge and sediment concentration values are given in Table 5. It is evident from the table that the performance of the MLR model in estimating discharge is close to the NF model in terms of correlation and efficiency statistics during training as well as validation. However, the RMSE and SSE for the MLR model are inferior compared with the NF model for calibration and validation. Similarly, the performance of the NF model in estimating sediment concentration is also much better than the MLR model, as indicated by statistical indices.

The behaviour of residuals or errors (the difference between observed and predicted values of a variable) provides useful information about the adequacy and strengths of a mathematical model. The distribution of estimated errors over the entire range of data by the NF and MLR models for the validation period is presented in Figure 5. In general, the errors of the MLR model are more scattered around the horizontal axis (representing zero error) than the errors of the NF model: the NF model clearly performs better than the MLR model in estimating both discharge and sediment concentration.

8. Application to Indravati basin, India

In order to reinforce the conclusion reached in Section 7, the developed NF model was applied to the Indravati basin, India. The two case studies (Thebes and Indravati) were selected because they are hydrologically different in terms of climate, river flow, sedimentation rate and timescale of data (hourly data for Thebes and daily data for Indravati). These case studies will also help to further evaluate the statistical significance of the proposed model. The procedure discussed earlier was adopted to determine the optimal structure for the NF model. The MLR model was also developed to compare the model performance in estimating sediment concentration.

The superior performance of the NF model is confirmed by the results shown in Table 6: the efficiency index is higher for the NF model than the MLR model for both Pathagudem and Nawarangpur basins. The NF model estimations are also much better than MLR in terms of the RMSE. Both case studies thus indicate that the NF model may be preferable for estimating discharge and sediment concentrations.

	Corr		Eff: %		RMSE: mg/l		SSE	
	CL	VL	CL	VL	CL	VL	CL	VL
Sediment	0·905	0·943	0·820	0·884	159·837	116·254	$6{\cdot}924 \times 10^6$	$2{\cdot}784 \times 10^6$
Discharge	0·997	0·998	0·994	0·997	306·834	156·204	$2{\cdot}550 \times 10^7$	$5{\cdot}030 \times 10^6$

Table 5. Statistical performance indices for calibration (CL) and validation (VL) of the MLR model

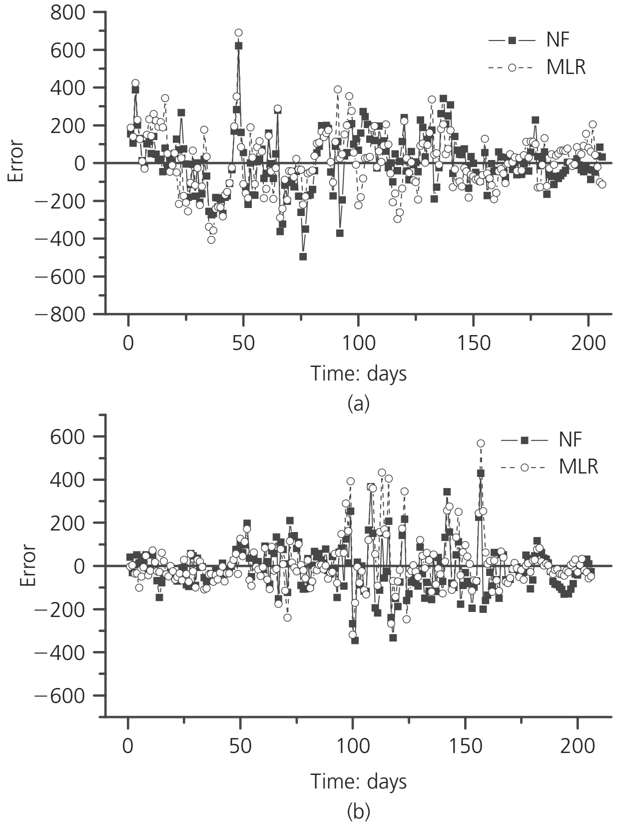

Figure 5. Distribution of errors with time: (a) discharge and (b) sediment concentration during validation

Model	Statistical index	Pathagudem		Nawarangpur	
		CL	VL	CL	VL
NF	Corr	0·89	0·91	0·85	0·76
	Eff: %	79·12	82·42	73·04	0·58
	RMSE: mg/l	117·80	104·30	249·60	147·60
MLR	Corr	0·88	0·90	0·84	0·76
	Eff: %	76·92	81·47	69·53	54·75
	RMSE: mg/l	183·40	136·50	334·70	276·30

Table 6. Performance indices for NF and MLR model for Indravati basin, India

9. Conclusions

The first-order Sugeno fuzzy inference system was applied to estimate discharge and sedimentation concentration using observed data of stage, discharge and sediment concentration. The identified model was applied to two different basins in the USA and India. Subtractive clustering, along with a least-squares estimation, was used to develop fuzzy rules. A number of fuzzy if–then rules were generated by changing the clustering radius, and optimal rules were found after several trials. A linear (MLR) model was also developed for estimation of discharge and sediment concentration in order that the results could be compared with the NF model results. For the sake of true comparison, model inputs for both models were kept exactly the same. Application of the NF and MLR models for estimation of river discharge and sediment concentration using historic stage and discharge data showed that the RMSE and SSE for the NF model were much smaller than those of the MLR model. Furthermore, the NF model was able to preserve the statistical properties of the estimates. The results show that the estimates of discharge and sediment concentration provided by the NF model are much better than the MLR model for both basins studied.

REFERENCES

Azamathullah HMd, Chang CK, Ab Ghani A *et al.* (2009) An ANFIS-based approach for predicting the bed load for moderately-sized rivers. *IAHR Journal of Hydro-Environment Research* **3(1)**: 35–44.

Brown M and Harris C (1994) *Neurofuzzy Adaptive Modeling and Control*. Prentice Hall. Upper Saddle River, NJ.

Chiu SL (1994) Fuzzy model identification based on cluster estimation. *Journal of Intelligent and Fuzzy Systems* **2**: 267–278.

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

Cigizoglu HK and Kisi O (2005) Flow prediction by three back propagation techniques using *k*-fold partitioning of neural network training data. *Nordic Hydrology* **36(1)**: 49–64.

Crawford CG (1991) Estimation of suspended sediment rating curves and mean suspended sediment loads. *Journal of Hydrology* **129(1–4)**: 331–348.

Duan Q, Sorooshian S and Gupta VK (1992) Effective and efficient global optimization for conceptual rainfall runoff models. *Water Resources Research* **28(4)**: 1015–1031.

Jain SK (2001) Development of integrated sediment rating curves using ANNs. *ASCE Journal of Hydraulic Engineering* **127(1)**: 30–37.

Jain SK, Nayak PC and Sudheer KP (2008) Models for estimating evapotranspiration using artificial neural networks and their physical interpretation. *Hydrologic Processes Journal* **22(13)**: 2225–2234.

Jang JSR (1993) ANFIS: adaptive network based fuzzy inference system. *IEEE Transactions on Systems, Man and Cybernetics* **23(3)**: 665–683.

Karim MF and Kennedy JF (1990) Menu of coupled velocity and sediment–discharge relations for rivers. *ASCE Journal of Hydraulic Engineering* **116(8)**: 978–996.

Kisi O (2005) Suspended sediment estimation using neuro-fuzzy and neural network approaches. *Hydrological Sciences Journal* **50(4)**: 683–696.

Lopes VL and Ffolliott PF (1993) Modeling sediment processes on small watersheds: a conceptual framework. II. Concentrated flow processes. *International Journal of Sediment Research* **81**: 1–23.

Mamdani EH and Assilan S (1975) An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies* **7(1)**: 1–13.

Nash JE and Sutcliffe JV (1970) River flow forecasting through conceptual models: 1. A discussion of principles. *Journal of Hydrology* **10(3)**: 282–290.

Nayak PC, Sudheer KP, Rangan DM and Ramasastri KS (2005) Short-term flood forecasting with a neurofuzzy model. *Water Resources Research* **41(4)**: doi:10.1029/2004WR003562.

Setnes M (2000) Supervised fuzzy clustering for rule extraction. *IEEE Transactions on Fuzzy Systems* **8(4)**: 416–424.

Shamseldin AY (1997) Application of a neural network technique to rainfall runoff modelling. *Journal of Hydrology* **199(3–4)**: 272–294.

Sudheer KP, Gosain AK and Ramasastri KS (2002) A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrological Processes* **16(6)**: 1325–1330.

Sudheer KP, Nayak PC and Ramasastri KS (2003) Improving peak flow estimates in artificial neural network river flow models. *Hydrological Processes* **17(3)**: 677–686.

Sugeno M and Kang G (1988) Structure identification of fuzzy model. *Fuzzy Sets and Systems* **26(1)**: 15–33.

Sumathi I and Santhana Bosu S (2002) Using GIS for facilitating sediment yield estimation. *Proceedings of the 3rd Asian Conference for Information Technology in Agriculture, Afita, Beijing*. 454–456.

Takagi T and Sugeno M (1985) Fuzzy identification of systems and its application to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics* **15(1)**: 116–132.

Tsukamoto Y (1979) An approach to fuzzy reasoning method. In *Advances in Fuzzy Set Theory and Application* (Gupta MM, Ragade RK and Yager RR (eds)). North-Holland, Amsterdam, pp. 137–149.

Yager R and Filev D (1994) Generation of fuzzy rules by mountain clustering. *Journal of Intelligent & Fuzzy Systems* **2(3)**: 209–219.