HYDROLOGICAL PROCESSES
Hydrol. Process. **19**, 955–968 (2005)
Published online 16 September 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/hyp.5553

Fuzzy computing based rainfall–runoff model for real time flood forecasting

P. C. Nayak,[1] K. P. Sudheer[2]* and K. S. Ramasastri[3]

[1] *National Institute of Hydrology, Deltaic Regional Centre, Siddartha Nagar, Kakinada, India–533 003*
[2] *Department of Civil Engineering, Indian Institute of Technology Madras, Chennai, India –600036*
[3] *National Institute of Hydrology, Roorkee, India –247 667*

Abstract:

This paper analyses the skills of fuzzy computing based rainfall–runoff model in real time flood forecasting. The potential of fuzzy computing has been demonstrated by developing a model for forecasting the river flow of Narmada basin in India. This work has demonstrated that fuzzy models can take advantage of their capability to simulate the unknown relationships between a set of relevant hydrological data such as rainfall and river flow. Many combinations of input variables were presented to the model with varying structures as a sensitivity study to verify the conclusions about the coherence between precipitation, upstream runoff and total watershed runoff. The most appropriate set of input variables was determined, and the study suggests that the river flow of Narmada behaves more like an autoregressive process. As the precipitation is weighted only a little by the model, the last time-steps of measured runoff are dominating the forecast. Thus a forecast based on expected rainfall becomes very inaccurate. Although good results for one-step-ahead forecasts are received, the accuracy deteriorates as the lead time increases. Using the one-step-ahead forecast model recursively to predict flows at higher lead time, however, produces better results as opposed to different independent fuzzy models to forecast flows at various lead times. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS fuzzy modelling; flood forecasting; clustering algorithm

INTRODUCTION

River flow forecasting has always been one of the most important issues in hydrology. To date, a wide variety of rainfall–runoff models has been developed and applied for flood forecasting. The approaches used for river flow forecasting cover a wide range of methods from completely black-box models to very detailed conceptual models (Porporato and Ridolfi, 2001). These rainfall–runoff models encompass a broad spectrum of more or less plausible descriptions of rainfall–runoff relationships and processes. The regular emergence of new models, across the whole spectrum of models, is a testament to the fact that a single superior model does not yet exist that works on all watersheds, and indeed never will be produced, despite continuing advances and enhancing of our modelling techniques (Beven, 1996a, b).

The rainfall–runoff relationship is the most complex hydrological process to explain, owing to tremendous spatial and temporal variability of basin characteristics and rainfall patterns, as well as a number of other variables associated with modelling the physical processes (Tokar and Markus, 2000). The transformation from rainfall to basin runoff involves many hydrological components that are believed to be highly non-linear, time varying, spatially distributed and not easily described by simple models. Owing to the difficulties associated with non-linear model structure identification and parameter estimation, very few truly non-linear system theoretic hydrological models have been reported (e.g. Amorocho and Brandstetter, 1971; Ikeda *et al*., 1976).

* Correspondence to: K. P. Sudheer, Indian Institute of Technology Madras, Dept of Civil Engineering, Chennai, 600036, India.
E-mail: sudheer@iitm.ac.in

Received 25 March 2003
Accepted 4 February 2004

In most cases, linearity or piecewise linearity has been assumed (Hsu *et al*., 1995). Recently a growing interest in the modelling of non-linear relationships has developed, and a variety of test procedures for detecting the non-linearities have evolved. If the aim of analysis is prediction, however, it is not sufficient to uncover the non-linearities. One needs to describe them through an adequate non-linear model. Unfortunately, for many applications the theory does not guide the model building process by suggesting the relevant input variables or the correct functional form. This particular difficulty makes it attractive to consider 'atheoretical' but flexible classes of statistical models (Anders and Korn, 1999). Fuzzy logic based approach is one such powerful tool to relate sets of predictor variables to forecast variables in non-linear ways.

Since Zadeh (1965) published the fuzzy set theory as an extension of classic set theory, the fuzzy set theory has been widely used in many fields of application, such as pattern recognition, data analysis, system control, etc. (Kruse *et al*., 1994; Theodoridis and Koutroumbas, 1999). The main advantages of the fuzzy applications are that the fuzzy theory is more logical and scientific in describing the properties of an object. The most unique characteristic of this theory, in contrast to classic mathematics, is its operation on various membership functions (MF) instead of the crisp real values of the variables. This heuristic permits fuzzy theory to be a powerful tool whenever it handles imprecise data or ambiguous non-linear relationships between the variables. Fuzzy theory appears to be extremely effective at handling dynamic, non-linear and noisy data, especially when the underlying physical relationships are not fully understood. As hydrologists are still uncertain about many of the aspects of the physical processes in the watershed, fuzzy theory has proved to be a very attractive tool enabling them to investigate such problems. The past decade has witnessed a few applications of fuzzy logic approach in water resources (Fujita *et al*., 1992; Zhu and Fujita, 1994; Zhu *et al*., 1994; See and Openshaw, 1999; Stuber *et al*., 2000; Hundecha *et al*., 2001).

This paper investigates the potential of fuzzy theory in real time flood forecasting by developing a rainfall–runoff model. The paper presents a cluster estimation method integrated with a least-squares estimation algorithm to provide a fast and robust method for identifying fuzzy models from input and output data. The underlying principles of the clustering algorithm are also discussed. A fuzzy rule based model is developed to forecast hourly flood discharge at a given streamflow gauge station at different lead times. The applicability of the method is demonstrated by modelling the river flow for Narmada basin in India.

FUZZY MODELLING

The basic structure of fuzzy modelling, commonly known as fuzzy inference system (FIS), is a rule-based or knowledge-based system consisting of three conceptual components: a rule base that consists of a collection of fuzzy IF–THEN rules; a database that defines the membership function (MF) used in fuzzy rules; and a reasoning mechanism that combines these rules into a mapping routine from the inputs to the outputs of the system, to derive a reasonable output conclusion (Figure 1). The FIS can take either fuzzy sets or crisp values as input, but the overall outputs are always fuzzy sets. Therefore, a defuzzification strategy is required to convert a fuzzy set to a crisp value. A FIS implements a non-linear mapping routine from its input space to output space. This mapping routine is accomplished by a number of fuzzy IF–THEN rules from the rule base; each of these rules describes the local behaviour of the mapping routine. The parameters of the IF–THEN rules (known as *antecedents* or *premise* in fuzzy modelling) define a fuzzy region of the input space, and the output parameters (known as *consequent* in fuzzy modelling) specify the corresponding output. Hence, the efficiency of the FIS depends on the number of fuzzy IF–THEN rules used for computation.

The application of fuzzy theory normally includes three procedures, i.e. fuzzification, logic decision and defuzzification. Fuzzification involves the identification of the input variables and the control variable (i.e. the output), the division of both input and the control variable into different domains, and choosing a membership function. Logic decision involves the design of the IF–THEN rule, and the determination of output fuzzy set. Defuzzification involves the determination of crisp output from the fuzzy outputs of the IF–THEN inference system. Division of the input and control variable is generally done using an unsupervised clustering algorithm,

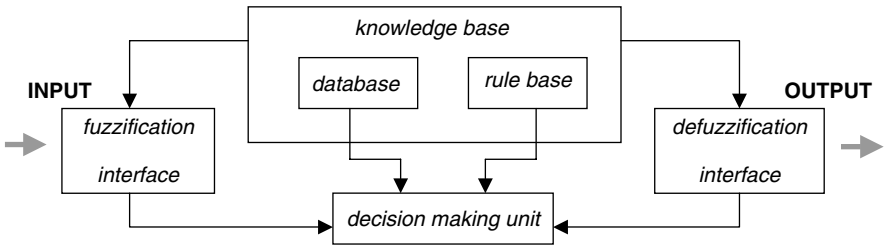

Figure 1. Fuzzy inference system

and is one of the most important aspects in fuzzy modelling. Ideally, for n domains and p input variables there could be n^p different IF–THEN rules. Hence the domain partitioning has to be done carefully; otherwise it may lead to overparameterization. Chiu (1994) presented an efficient method, *subtractive clustering,* for estimating the cluster centres of numerical data. The essential advantage of their method is the computational efficiency and robustness.

In the applications of the fuzzy system in control and forecasting, there are mainly two approaches, the first one being the Mamdani approach and the other the Takagi–Sugeno approach (Kruse *et al*., 1994). For the Mamdani approach (Mamdani and Assilian, 1975), which has been used in some hydrological applications (Schulz and Huwe, 1997; Schulz *et al*., 1999), there are three clear procedures, i.e. fuzzification, logic decision and defuzzification, as described earlier. The Takagi–Sugeno approach (Takagi and Sugeno, 1985), however, does not have an explicit defuzzification procedure. Rather, it amalgamates the logic decision and defuzzification procedures into one composite procedure. To the knowledge of the authors, the Takagi–Sugeno fuzzy system has not yet been applied in the hydrological field for flood forecasting. A definite advantage of the Takagi–Sugeno fuzzy system of first order is that the output function is a first-order polynomial of the input variables, and hence a least-square error procedure can be used for the consequent parameter estimation.

METHODS

In the current study, a Takagi–Sugeno fuzzy inference system has been developed using the *subtractive clustering* algorithm integrated with a linear least-squares estimate algorithm for developing a rainfall–runoff model. The basic concepts about the fuzzy theory and its applications, such as fuzzy set, membership functions, the domain partitions and fuzzy IF–THEN inference rules, which have been introduced in numerous hydrological papers (Fujita *et al*., 1992; Zhu and Fujita, 1994; Zhu *et al*., 1994; Stuber *et al*., 2000; Hundecha *et al*., 2001), are not reproduced in the body of this paper. As *subtractive clustering* is a relatively new concept, the details of the method are presented in the following sections.

Subtractive clustering

Subtractive clustering (Chiu, 1994) is an extension of the mountain clustering method (Yager and Filev, 1994), where the potential is calculated for the data rather than the grid points defined on the data space. As a result, clusters are elected from the system training data according to their potential. Subtractive clustering compared with mountain clustering has an advantage in that there is no need to estimate a resolution for the grid.

Consider a collection of n data points $[x_1, x_2, \ldots x_n]$ in an M dimensional space. Without the loss of generality, it is assumed that the data points have been normalized in each dimension so that their

coordinate ranges in each dimensions are equal; i.e. the data points are bounded by a hypercube. Each data point is considered as a potential cluster centre and a measure of the potential of the data point x_i is defined as

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

where

$$\alpha = \frac{4}{r_a^2} \tag{2}$$

and r_a is a positive constant. Thus, the measure of potential for a data point is a function of its distance to all other data points. A data point with many neighbouring data points will have a high potential value. The constant r_a is effectively the radius defining a neighbourhood; a data point outside this radius has little influence on the potential. After the potential of every data point has been computed, 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^* be its potential value. Then the potential of each data point x_i may be revised by the formula

$$p_i = p_i - p_1^* e^{-\beta\|x_i - x_1^*\|^2} \tag{3}$$

where

$$\beta = \frac{4}{r_b^2} \tag{4}$$

and r_b is a positive constant. That is, subtract an amount of potential from the first cluster centre. The data point near the first cluster centre will have a greatly reduced potential, and therefore is unlikely to be selected as the next cluster centre. The constant r_b is effectively the radius defining a neighbourhood; which will have measurable reductions in potential. To avoid obtaining closely spaced cluster centres, r_b may be set to be somewhat greater than r_a; a good choice is $r_b = 1{\cdot}5r_a$.

When the potential of all the data points has been revised according to Equation (3), the data point with the highest remaining potential is selected as the second cluster centre. Similarly, the potential of each data point is reduced according to their distance to the second cluster centre. In general, after the kth cluster has been obtained, the potential of each data point is revised by the formula

$$p_i = p_i - p_k^* e^{-\beta\|x_i - x_k^*\|^2} \tag{5}$$

where x_k^* is the location of the kth cluster centre and p_k^* is the potential value. The process is repeated until a given threshold for the potential is obtained such that $\frac{P_k^*}{P_1^*} < \varepsilon$. The choice of ε is an important factor affecting the results; if ε is too large too few data points will be accepted as cluster centres and if ε is too small, too many cluster centres will be generated.

Fuzzy model identification

When cluster estimation is applied to a collection of input and output data, each cluster centre is in essence a prototypical data point that represents a characteristic behaviour of the system. Hence, each cluster centre can be used as the basis of a rule that illustrates the system behaviour.

Consider a set of c cluster centres $\{x_1^*, x_2^*, \ldots, x_c^*\}$ in an M dimensional space. Let the first N dimensions correspond to input variables and the last $M - N$ dimensions correspond to output variables. Each vector $\boldsymbol{x}_i^*$ may be decomposed into two component vectors $\boldsymbol{y}_i^*$ and $\boldsymbol{z}_i^*$, where $\boldsymbol{y}_i^*$ contains the first N elements of $\boldsymbol{x}_i^*$ (i.e. the coordinates of the cluster centre in input space) and $\boldsymbol{z}_i^*$ contains the last $M - N$ elements (i.e. the coordinates of the cluster centre in output space).

Each cluster centre x_i^* may be considered as a fuzzy rule that describes the system behaviour. Given an input vector $\boldsymbol{y}$, the degree to which rule i is fulfilled is defined as

$$\mu_i = \mathrm{e}^{-\alpha\|y-y_i^*\|^2} \tag{6}$$

where α is the constant defined by Equation (2). The output vector $\boldsymbol{z}$ may be computed as

$$\boldsymbol{z} = \frac{\sum_{i=1}^{c} \mu_i z_i^*}{\sum_{i-1}^{c} \mu_i} \tag{7}$$

Equations (6) and (7) provide a simple and direct way to translate a set of cluster centres into a FIS. This computational scheme may be viewed in terms of a FIS using traditional fuzzy IF–THEN rules. Each rule has the form

$$\text{if } y_1 \text{ is } A_1 \text{ and } y_2 \text{ is } A_2 \text{ and} \ldots \text{ then } z_1 \text{ is } B_1 \text{ and } z_2 \text{ is } B_2 \ldots \tag{8}$$

where y_i is the ith input variable and z_j is the jth output variable; A_i is a fuzzy set defined by an exponential membership function and B_j is a singleton. This computational scheme is equivalent to an inference method that uses multiplication as the AND operator, weights the output of each rule by the rule's firing strength and the overall output is determined as the weighted average of each rule output. Given the same number of cluster numbers, the modelling accuracy of the system can be improved significantly if z_i^* in Equation (7) is considered as an optimized linear function of the input variables instead of a simple constant. It follows

$$z_i^* = G_i y + h_i \tag{9}$$

where G_i is $(M - N) \times N$ constant matrix, and $\boldsymbol{h}_i$ is a constant column vector with $(M - N)$ elements. Thus the computational scheme becomes equivalent to a first-order Takagi–Sugeno type FIS. Expressing z_i^* as a linear function of the input allows a significant degree of rule optimization to be performed without adding much computational complexity. Optimizing the parameters in the consequent equation with respect to training data therefore reduces to a linear least-squares estimation (LSE) problem (Takagi and Sugeno, 1985). Such problems can be solved easily and the solution is always global. For details about alteration from the equation parameter optimization problem into the LSE problem readers are referred to Takagi and Sugeno(1985). The LSE computes the optimal G_i and h_i and minimizes the error between the output of the FIS and the output of the training data.

Fuzzy rainfall–Runoff model development

There are no fixed rules for developing a fuzzy model, even though a general framework can be followed based on previous successful applications in engineering. The goal of an FIS is to generalize a relationship of the form of

$$\boldsymbol{Y}^m = f(\boldsymbol{X}^n) \tag{10}$$

where X^n is an n dimensional input vector consisting of variables $x_1, \ldots, x_i, \ldots, x_n$: Y^m is an m-dimensional output vector of the resulting variables of interest consisting of $y_1, \ldots, y_i, \ldots, y_m$. In rainfall–runoff modelling, values of x_i may be rainfall/runoff values at different lag times and the value of y_i is generally the flow for the subsequent period. How many antecedent rainfall/runoff values, however, should be included in the vector X^n is not known a priori. A firm understanding of the hydrological system under consideration would play an important role in successful implementation of FIS. This would help in avoiding loss of information that may result if key input variables are omitted, and also prevent inclusion of spurious input variables

that tend to confuse the training process. Sudheer *et al*. (2002) present a statistical approach to identify the appropriate input vector that can best represent the process. Their method is based on the heuristic that the potential influencing variables corresponding to different time lags can be identified through statistical analysis of the data series. The approach is based on cross-, auto- and partial autocorrelations between the variables in question, and the current study used this method in identifying the input vector.

Sudheer *et al*. (2003) suggest that by following the guidelines used in traditional statistical modelling, the model performance can be improved in the case of soft computing based models. In most traditional statistical models, the data have to be normally distributed before the model coefficients can be estimated efficiently. If the data are not normally distributed, suitable transformations to normality have to be applied. Data transformations often are used to simplify the structure of the data so that they follow a convenient statistical model (Sudheer *et al*., 2003). In the current study, log-normal transformation is used and the deterministic component in the runoff and rainfall series was removed prior to the modelling. The variables are scaled to limit between 0 and 1 as the activations function warrants. The total available data have been divided into two sets, a calibration set and a validation set. The parameters of the model are identified using the calibration data set, and the model is tested for its performance on the validation data set. The resulting hydrographs from the model are analysed statistically using various indices utilized for performance analysis of models. The goodness of fit statistics considered are the root mean square error (RMSE) between the computed and observed runoff, coefficient of correlation (CORR), the model efficiency (Nash and Sutcliffe, 1970) and percentage error in peak flow estimation (%EMF). The general nomenclature of the model used in this paper is M-*I*-*J*, where *I* is the time horizon of forecast in hours and *J* is the model number.

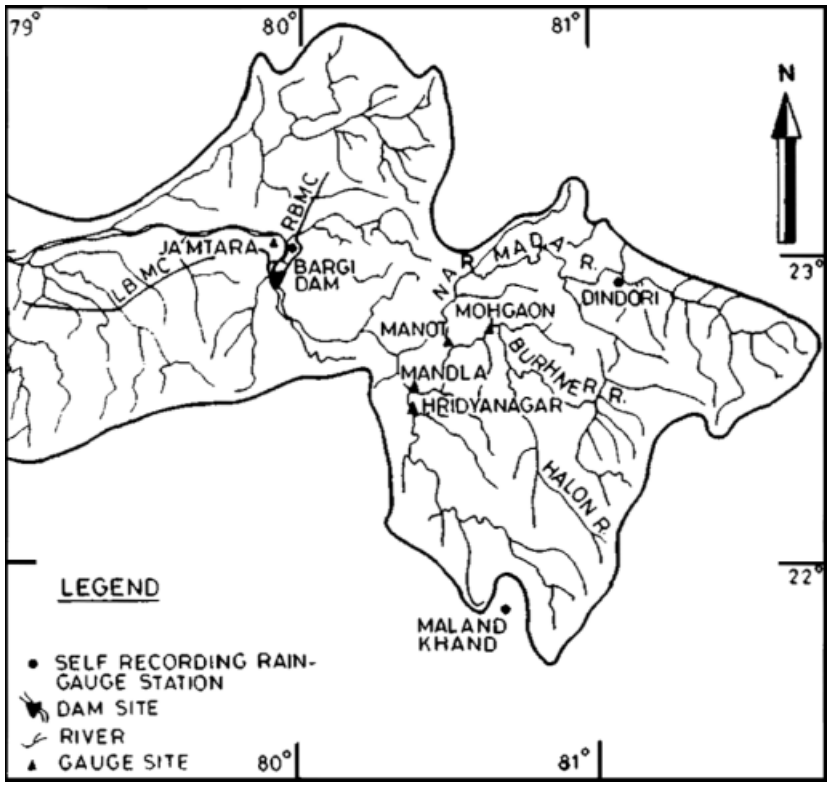

Figure 2. Location of Mandala gauging station on Narmada River

Table I. Variables in the input vector to the fuzzy model: $Q(t)$, $R(t)$ are river flow and rainfall respectively at any time t

Gauging station and rainfall	Variables in the input vector
Manot	$Q(t-2)$, $Q(t-3)$, $(t-4)$
Mohegaon	$Q(t-4)$, $Q(t-5)$, $Q(t-6)$
Hridyanagar	$Q(t-1)$
Dindori	$Q(t-11)$, $Q(t-12)$, $Q(t-13)$
Mandala	$Q(t-1)$, $Q(t-2)$, $Q(t-3)$, $Q(t-4)$, $Q(t-5)$, $Q(t-6)$
Areal rainfall	$R(t-16)$, $R(t-17)$, $R(t-18)$

Case study

The potential of fuzzy theory in real time flood forecasting is illustrated by developing a rainfall–runoff model for Narmada River basin in India (Figure 2). The study aimed at forecasting the flow at Mandala gauging station, which has a total drainage area of 13 120 km^2. The rainfall and runoff data available for the monsoon season during years 1989 to 1993 on hourly interval have been used in the study. The rainfall data are available in the form of areal averages for the entire basin. The hourly runoff data for upstream gauging stations, namely Manot, Mohegaon, Hridyanagar and Dindori are also used in the study. The identified input vector according to Sudheer *et al*. (2002) included a total number of 19 variables, as detailed in Table I, where Q represents the runoff, R is the rainfall and t indicates the prediction time. The output of the model is the flow at Mandala.

The model is trained using data for 3 years (1989–1991) and validated on the rest of the data (1992–1993). The cluster radius is fixed as 0·5 after several trials. It may be noted that the radius specifies the range of influence of the cluster centre of each input and output dimension. Assuming that the cluster radius falls within the hyper box of unit dimension, a smaller cluster radius will usually yield more clusters in the data, and hence a greater number of rules. Simultaneously it increases the model complexity and hence decreases the parsimony. Exponential membership function (Equation 6) is used in the model.

RESULTS AND DISCUSSIONS

Forecasting at shorter times (1 h)

The values of performance indices for the 1-h lead forecast fuzzy model (M-1-1) for Narmada basin are presented in Table II. The correlation statistics evaluate the linear correlation between the observed and the computed runoff, which is consistent during calibration as well as the validation period. The fuzzy model performance is very good in terms of the efficiency statistic, as the calibration and validation efficiency is greater than 99% (Table II). The RMSE statistic is a measure of residual variance and is indicative of the model's ability to predict high flows (Hsu *et al*., 1995). Considering the magnitude of the peak flow during the period of study (11 794 m^3 s^{-1}), the fuzzy model is able to forecast the flows with reasonable accuracy, as can be evidenced from the low RMSE values. The %EMF statistic is a measure of the percentage error in estimating peak flow in a data series, and the model prediction of peak flow is good as the estimation error is less than 2%.

A second model is developed by removing the rainfall values from the input vector, and the model parameters are re-estimated (M-1-2). This analysis is performed with the heuristic that as the upstream runoff values are incorporated in the input vector the prediction may not be sensitive to rainfall information. This modification in input vector produces results comparable to that of M-1-1, and it seems to indicate that the downstream prediction is not sensitive to rainfall if upstream discharge values are considered in the input (see Table II). It further adds that it may be necessary to test the sensitivity

Table II. Performance indices for 1 h lead forecast models

	Model: M-1-1		Model: M-1-2		Model: M-1-3	
	Calibration	Validation	Calibration	Validation	Calibration	Validation
Correlation	0·9986	0·9972	0·9986	0·9975	0·9982	0·9963
Efficiency (%)	99·72	99·43	99·71	99·48	99·94	99·25
RMSE	39·7927	32·5844	40·3914	31·0911	44·9987	34·5489
%EMF	−1·6308	−1·9113	−1·7416	−1·884	−0·4609	−0·9723

of tributary flow also, and consequently another model is developed by using only discharge information at the Mandala gauging station (M-1-3). Interestingly, the results reveal that the model predictions are in good agreement with the observed flows for a M-1-3 model (see Table II). This is well in line with the suggestion of Campolo *et al*. (1999) that the capacity of a basin to respond to a perturbation is more accurate when recent discharge values are used. Further, it significantly reduces the complexity of the model. Figure 3 depicts the predicted discharge hydrograph on a scatter plot for all three analyses.

An important characteristic of the hydrograph is the peak flow. The error in peak prediction defined as the difference between the observed and the computed values at the peak flow expressed as percentage of the observed peak flow can be used to evaluate the reliability of the model in forecasting the floods. The error on predicted peaks is less than 2% for all the models. It may be noted that M-1-3 improves the peak flow

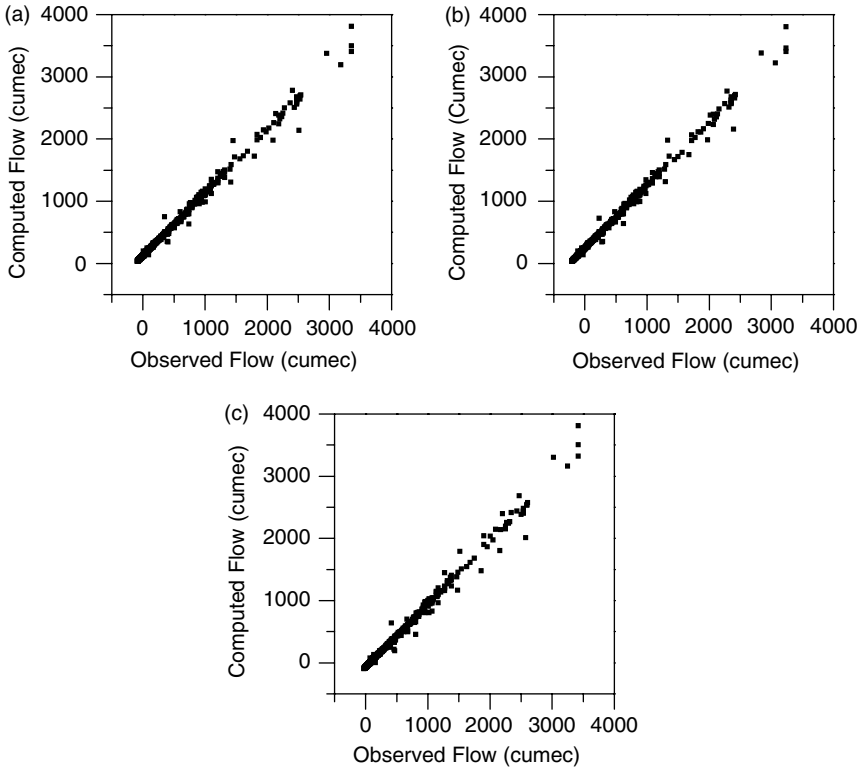

Figure 3. Scatter plot of the observed and the computed river flows during validation year 1993: (a) M-1-1 model, (b) M-1-2 model and (c) M-1-3 model

predictions over the other two models as the errors are less than 1%. In all the cases, five cluster centres are found optimal using the clustering algorithm resulting in five fuzzy IF–THEN rules.

Forecasting at larger times (>1 *h*)

The results discussed suggest that the information given as input to the model is sufficient to capture the transformation. However, if the target is to predict the flow at time t, supplying the observed value at time $t-1$ (see Table I) is too stringent a requirement for the model and makes it impractical to use for flood risk warning. One way of addressing this issue is by using the computed flow at time $t+i$ recursively to forecast the flow at $t+i+1$, for any value of step i. Alternatively, one can also develop different fuzzy models by using the flow at the lead time of interest as the output of the model. In the current study, both these options have been explored. However, the former option was explored only for model M-1-3, as the other models require information at $t-1$ for other variables also (see Table I).

Model M-1-3 was used to develop forecasts up to 24 h in advance, in a recursive way. It is to be noted that the performance of any model may be good for smaller lead times, but may become worse as the lead time increases. This can be observed in Figure 4 in which the RMSE is small for forecasts 1 and 2 h ahead but increases thereafter. It is interesting to note that rate of rise of RMSE is linear with lead time, and the reasons for these phenomena need to be explored further. In the procedure, the error on the predicted flow $Q_{(t+1)}$ for hour $(t+1)$ will definitely affect the forecasted flow $Q_{(t+2)}$ on the hour $(t+2)$, and this is applicable to subsequent flows also. This way, the error is accumulated as the lead time increases and this error accumulation is the obvious reason for an increasing trend in error with increase in lead time (Figure 4). Sudheer and Jain (2003) suggest that in discharge estimation using the computed discharge at previous time-steps recursively, an artificial neural network model fails to preserve the accuracy in discharges computed at subsequent time periods, and the results of this study indicate that their observation holds valid in fuzzy modelling also.

Figure 5 shows the forecasted and observed hydrograph of a typical flood event at three different lead times (1, 3 and 6 h) for model M-1-3. It can be observed that the forecasted flows at 1 h in advance are satisfactory except for a small portion at the tail end of the recession limb. As the forecast lead time increases, the accuracy decreases, however the lower flow forecasts at the tail end of the recession limb are not significantly affected. It is observed that the forecasts at a 3 h lead time have a phase lag, and the forecasted values are overestimated as a result of error accumulation at previous steps, as explained earlier. However, the reasons for this phase lag are to be explored further.

It appears that when assessing the performance of a stream-flow forecasting model for its applicability in forecasting stream flows at larger lead times, it is not only important to evaluate the average prediction error but also the distribution of prediction errors. It is important to know whether the model is predicting higher magnitude flows badly or the lower magnitude flows badly, which may help in further refining the model.

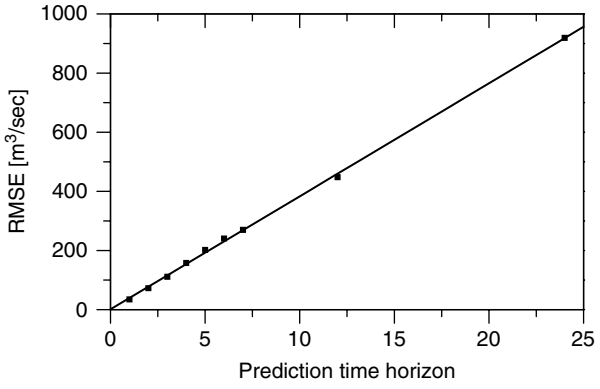

Figure 4. Testing RMSE of model M-1-3 at different lead time forecasts

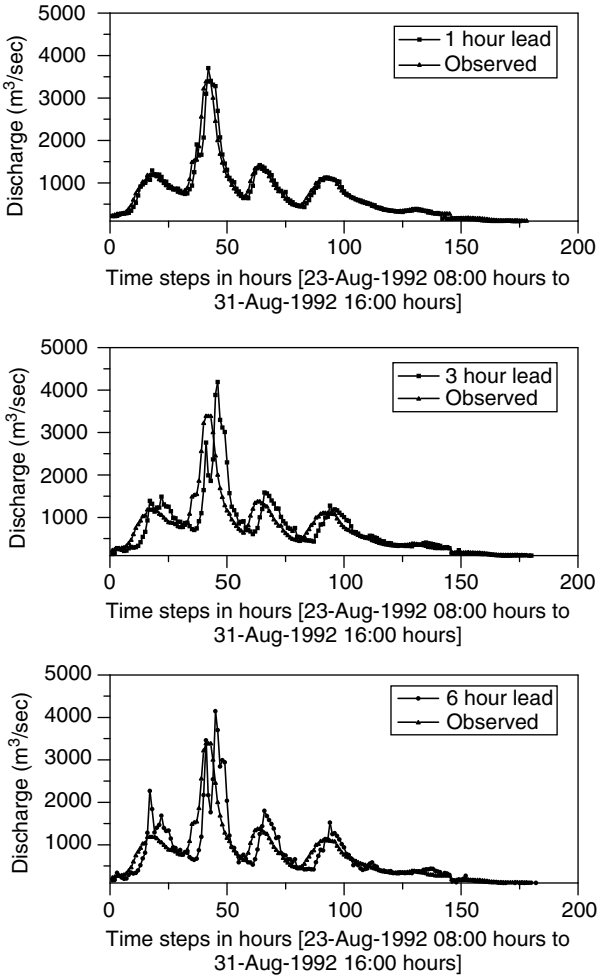

Figure 5. The forecast hydrograph at different lead times by the M-1-3 model for a typical flood event

The statistical performance evaluation criteria used so far in this study are global statistics and do not provide any information on the distribution of errors. It is to be noted that the coefficient of efficiency can be high (80 or 90%) even for poor models, and the best models do not produce values that, on first examination, are impressively higher (Garrick *et al*., 1978; Legates and McCabe, 1999). The RMSE statistic indicates only the model's ability to predict a value away from mean (Hsu *et al*., 1995). Moreover, the fuzzy model is trained by minimizing the sum squared error at the output layer that is similar to RMSE. The correlation statistics provide information only on the strength of the relationship between the observed and the computed values. Therefore, in order to test the robustness of the model developed, it is important to test the model using some other performance evaluation criteria such as average absolute relative error (AARE) and threshold statistics (Jain *et al*., 2001; Jain and Ormsbee, 2002). The AARE and threshold statistics (TS) not only give the performance index in terms of predicting flows but also the distribution of the prediction errors.

These criteria can be computed as

$$\text{AARE} = \frac{1}{n}\sum_{i=1}^{n}|\text{RE}_t| \quad \text{in which, } \text{RE}_t = \frac{Q_t^{\text{o}} - Q_t^{\text{c}}}{Q_t^{o}}100 \tag{11}$$

where RE_t is the relative error in forecast at time t expressed as a percentage, Q_t^o is the observed stream flow at time t, Q_t^c is the computed stream flow at time t, and n is the total number of testing patterns. Clearly the smaller the value of AARE the better is the performance.

The threshold statistic for a level of $x\%$ is a measure of the consistency in forecasting errors from a particular model. The threshold statistic is represented as TS_x and expressed as a percentage. This criterion can be expressed for different levels of absolute relative error from the model. It is computed for $x\%$ level (TL) as

$$TS_x = \frac{Y_x}{n} \times 100 \tag{12}$$

where Y_x is the number of computed stream flows (out of n total computed) for which absolute relative error is less than $x\%$ from the model.

The computed values of AARE and TS_x from model M-1-3 are presented in Table III. It is clear from Table III that the relative error increases as the lead time increases. Further it is evident that even though the model M-1-3 forecasts the flows 1 h in advance within 1% error on average (see Table II), only about 48·9% of the testing data have been forecast with this accuracy. This confirms previous considerations that using computed values at previous time-steps to obtain a forecast for subsequent steps may not produce satisfactory results. It must be pointed out that 85% of the predicted flows up to 6 h are with absolute relative errors less than 15%. It is observed that the residual errors at different lead times of forecast are uncorrelated up to 6-h lead time, thus increasing the confidence in the model developed. For larger lead times, however, there is a correlation between TL and TS, suggesting less confidence in the model's use for predictions at longer lead times.

The earlier stated heuristic that the use of separate fuzzy models to compute output representing the stream flow at different lead times may give better or similar performance is also evaluated. Consequently, eight different fuzzy models are developed to forecast flows at 2, 3, 4, 5, 6, 7, 12 and 24 h in advance (these models are represented as M-J-4, where J represents the lead time). It may be noted that the inputs to all these models are the same as that of M-1-1. The performance of these models in terms of RMSE and efficiency statistics against prediction time horizon is presented in Figure 6. It appears that the error is non-linearly varying along the prediction time horizon in this analysis. However, a slight difference in the average rise rate of RMSE for calibration and validation is found. From the comparison of the results obtained during calibration and validation, it can be observed that the rate of variation of efficiency increases slightly at 6 h. This observation is in line with Campolo *et al*. (1999) who noted that the displacement between the two efficiency curves increases as the prediction time horizon becomes greater than the minimum time-lag information in the input vector. In other words, this confirms the previous considerations about the relationship between the time horizon and information available for prediction.

Table IV presents the AARE and threshold statistics of the M-J-4 models. It may be noted that for 2 h in advance, the M-2-4 model predicted only 23·66% of the total testing data with less than 1% absolute

Table III. The AARE and threshold statistics from M-1-3 during testing for flood forecasting

	Prediction time horizon					
	1 h	2 h	3 h	4 h	5 h	6 h
AARE (%)	1·9798	3·1732	4·8590	6·6452	8·4633	10·2601
TS_1	48·9494	38·1323	29·4423	21·7899	16·1349	12·0623
TS_2	78·1582	61·2451	48·0674	39·5590	32·9702	25·3178
TS_3	85·8885	73·8262	60·2075	51·1543	43·5538	38·0545
TS_5	92·7367	86·1219	75·7198	66·4073	59·2996	53·3333
TS_{10}	97·1984	94·9676	90·2464	84·9287	79·7925	75·3826
TS_{15}	98·3917	97·1984	95·0195	91·9585	88·6122	85·2918

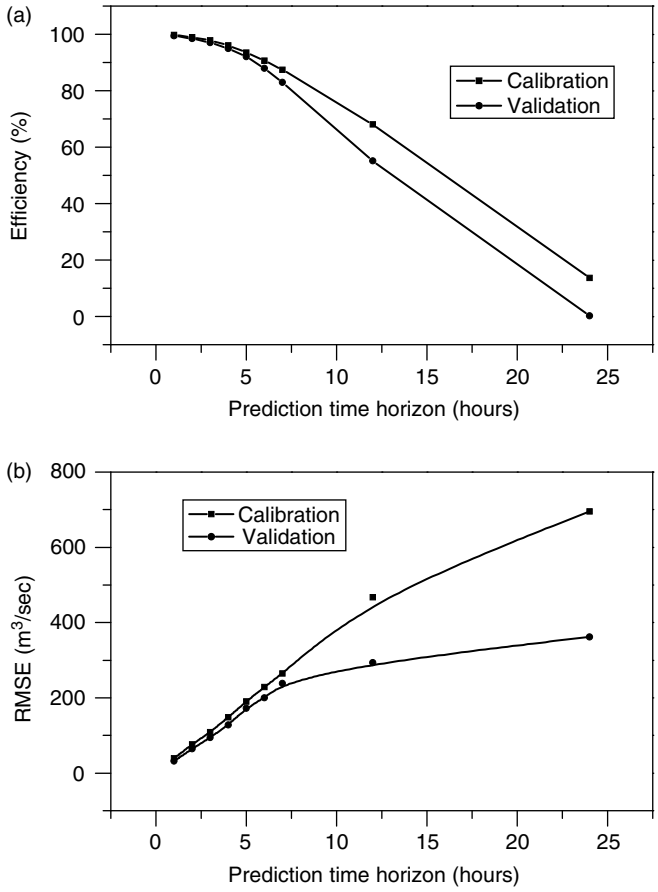

Figure 6. The RMSE and efficiency plot for M-*J*-4 models during calibration and validation: (a) RMSE and (b) efficiency

relative error as opposed to 38·13% (see Table III) by the M-1-3 model. A similar trend is observed for other lead time forecasts as well. The AARE statistic for different models also follows an identical tendency. The M-6-4 model forecast for estimated flow 6 h in advance places 78% of the total test set within 15% deviation from the observed, i.e. 7% less in number compared with M-1-3. It is worth mentioning, however, that the prediction of flow up to 6 h in advance is made by these models with an efficiency of more than 85%, which according to Shamseldin (1997) is highly satisfactory.

The foregoing discussion suggests that a one-step-ahead forecast model if used recursively to forecast flows at larger lead times would result in better performance, as opposed to developing different models. Alternatively, the robustness of fuzzy computing, because of its massively parallel structure, can be used to advantage by having different variables as output in a fuzzy model representing stream flows at different lead times as output in a fuzzy model with the same input vector. This method does not guarantee a better performance, but it appears that it would be a worthwhile exercise.

SUMMARY AND CONCLUSIONS

The fuzzy computing approach presented in this paper for river flow forecasting has furnished very promising results. For continuous river flow series, a very good agreement has been obtained between the forecast

Table IV. The AARE and threshold statistics of M-J-4 models

	Prediction time horizon								
	1 h	2 h	3 h	4 h	5 h	6 h	7 h	12 h	24 h
AARE	2·1030	3·7649	5·4645	7·6363	9·9126	12·2728	14·4127	24·7037	50·4375
TS_1	45·1604	23·6631	19·1711	13·8770	10·6684	8·6096	6·4706	3·8503	2·0856
TS_2	71·1497	44·4118	34·5455	25·6684	20·0535	16·5241	13·9037	7·4332	4·3048
TS_3	82·8610	60·3476	47·8075	37·2727	30·6684	25·2941	21·123	11·5508	6·4439
TS_5	93·0214	79·8663	66·8717	54·1711	45·7219	39·7059	34·8663	19·6791	10·9091
TS_{10}	97·5668	94·2246	87·7273	78·9037	71·4973	62·7540	58·369	40·3743	20·8556
TS_{15}	98·7433	97·2995	93·7968	89·3316	83·3957	77·9144	72·6471	59·7861	30·0267

and the observed values, especially in correspondence to the flood peaks. The values of three performance evaluation criteria, i.e. the coefficient of efficiency, the RMSE and the coefficient of correlation, are very good and consistent for flows forecast 1 h in advance. The value of the relative error of the peak, which is a useful index in simulating events such as floods, is within reasonable limits for the fuzzy model. As a result, the fuzzy approach seems to be well suited to exploit the information to model the non-linear dynamics present in the data without requiring any previously defined mathematical model of the phenomenon. The very short computer time required for single forecast (always a fraction of a second when using a normal Pentium processor) does not lead to any constraints for the use of the method for real time flood forecasting. The results indicate that simple models involving fewer parameters to be evaluated, and relying on simple mathematical procedures (e.g. the ordinary least squares solution), are good in discharge forecasting. This study suggests that simpler models with only runoff values in the input vector for continuous river-flow simulation can surpass their complex counterparts that use a number of influencing variables as the input in performance. There is a strong justification, therefore, for the general claim that increasing the model complexity, thereby increasing the number of parameters, does not necessarily enhance the model performance. The results show that the use of a fuzzy model for one-step-ahead forecasts may allow an extension of the lead-time up to which a reliable flood forecast may be issued, by providing a quick prediction based solely on forecast values. More substantial improvement certainly should be pursued through further research to improve the forecasts at greater lead times.

REFERENCES

Amorocho J, Brandstetter A. 1971. A critique of current methods of hydrologic systems investigations. *Eos (Transactions of the American Geophysical Union)* **45**: 307–321.

Anders U, Korn O. 1999. Model selection in neural networks. *Neural Networks* **12**: 309–323.

Beven KJ. 1996a. A discussion of distributed hydrological modelling. In *Distributed Hydrological Modelling*, Abbott MB, Refsgarrd JC (eds). Kluwer: Dordecht; 255–278.

Beven KJ. 1996b. Response to comments on A discussion of distributed hydrological modelling by JC Refsgarrd *et al*. In *Distributed Hydrological Modelling*, Abbott MB, Refsgarrd JC (eds). Kluwer: Dordecht; 289–296.

Campolo M, Andreussi P, Soldati A. 1999. River flood forecasting with neural network model. *Water Resources Research* **35**(4): 1191–1197.

Chiu S. 1994. Fuzzy model identification based on cluster estimation. *Journal of Intelligent and Fuzzy Systems* **2**(3): 267–278.

Fujita M, Zhu M-L, Nakoa T, Ishi C. 1992. An application of fuzzy set theory to runoff prediction. *Proceedings of the Sixth IAHR International Symposium on Stochastic Hydraulics*, Taipei, Taiwan; 727–734.

Garrick M, Cunnane C, Nash JE. 1978. A criterion of efficiency for rainfall–runoff models. *Journal of Hydrology* **36**: 375–381.

Hsu K, Gupta VH, Soorooshian S. 1995. Artificial neural network modeling of the rainfall–runoff process. *Water Resources Research* **31**(10): 2517–2530.

Hundecha Y, Bardossy A, Theisen H-W. 2001. Development of a fuzzy logic based rainfall–runoff model. *Hydrological Sciences Journal* **46**(3): 363–377.

Ikeda S, Ochiai M, Sawaragi Y. 1976. Sequential GMDH algorithm and its applications to river flow prediction. *IEEE Transactions of System Management and Cybernetics* **6**(7): 473–479.

Jain A, Ormsbee LE. 2002. Evaluation of short-term water demand forecast modeling techniques: conventional methods versus AI. *Journal of the American Water Works Association* **94**(7): 64–72.

Jain A, Varshney AK, Joshi UC. 2001. Short-term water demand forecast modeling at IIT Kanpur using artificial neural networks. *Water Resources Management* **15**(5): 299–321.
Kruse R, Gebhardt J, Klawonn F. 1994. *Foundations of Fuzzy Systems*. Wiley: New York.
Legates DR, McCabe Jr. GJ. 1999. Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35**(1): 233–241.
Mamdani EH, Assilian S. 1975. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man–Machine Studies* **7**(1): 1–13.
Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models: 1. A discussion of principles. *Journal of Hydrology* **10**: 282–290.
Porporato A, Ridolfi L. 2001. Multivariate nonlinear prediction of river flows. *Journal of Hydrology* **248**: 109–122.
Schulz K, Huwe B. 1997. Water flow modeling in the unsaturated zone with imprecise parameters using a fuzzy approach. *Journal of Hydrology* **201**: 211–229.
Schulz K, Huwe B, Peiffer S. 1999. Parameter uncertainty in chemical equilibrium calculations using fuzzy set theory. *Journal of Hydrology* **217**: 119–134.
See L, Openshaw S. 1999. Applying soft computing approaches to river level forecasting. *Hydrological Sciences Journal* **44**(5): 763–779.
Shamseldin AY. 1997. Application of a neural network technique to rainfall–runoff modeling. *Journal of Hydrology* **199**: 272–294.
Stuber M, Gemmar P, Greving M. 2000. Machine supported development of fuzzy-flood forecast systems. *European Conference on Advances in Flood Research*, Potsdam, PIK Report 65, Bronstert A, Bismuth C, Menzel L (eds), Reprint of Proceedings, Vol. 2; 504–515.
Sudheer KP, Jain SK. 2003. Radial basis function neural networks for modeling stage discharge relationship. *Journal of Hydrological Engineering American Society of Civil Engineers* **8**(3): 161–164.
Sudheer KP, Gosain AK, Ramasastri KS. 2002. A data-driven algorithm for constructing artificial neural network rainfall–runoff models. *Hydrological Processes* **16**: 1325–1330.
Sudheer KP, Nayak PC, Ramasastri KS. 2003. Improving peak flow estimates in artificial neural network river flow models. *Hydrological Processes* **17**(1): 677–686.
Takagi T, 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.
Theodoridis S, Koutroumbas K. 1999. *Pattern Recognition*. Academic Press: New York; 482–483.
Tokar AS, Markus M. 2000. Precipitation–runoff modeling using artificial neural network and conceptual models. *Journal of Hydrologic Engineering, American Society of Civil Engineers* **5**(2): 156–161.
Yager R, Filev D. 1994. Generation of fuzzy rules by mountain clustering. *Journal of Intelligent and Fuzzy Systems* **2**(3): 209–219.
Zadeh LA. 1965. Fuzzy sets. *Information and Control* **8**(3): 338–353.
Zhu M-L, Fujita M. 1994. Comparison between fuzzy reasoning and neural network method to forecast runoff discharge. *Journal of Hydroscience and Hydraulic Engineering* **12**(2): 131–141.
Zhu M-L, Fujita M, Hashimoto N, Kudo M. 1994. Long lead time forecast of runoff using fuzzy reasoning method. *Journal of the Japanese Society of Hydrology and Water Resources* **7**(2): 83–89.

