HYDROLOGICAL PROCESSES
Hydrol. Process. **22**, 827–841 (2008)
Published online 11 June 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/hyp.6644

Fuzzy model identification based on cluster estimation for reservoir inflow forecasting

P. C. Nayak[1]* and K. P. Sudheer[2]

[1] *Deltaic Regional Centre, National Institute of Hydrology, Kakinada 533 003, India*
[2] *Department of Civil Engineering, Indian Institute of Technology Madras, Chennai 600036, India*

Abstract:

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. Since 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 effectiveness of the fuzzy model lies in the identification of the antecedent membership function (MF), which is generally addressed through a fuzzy clustering approach. Most of the applications of fuzzy computing in hydrology seem to have selected the clustering algorithm quite arbitrarily. However, it is apparent that, as the antecedent parameters are based solely on the identified clusters, the method used for clustering should certainly have an impact on the overall performance of the model. This paper presents the results of a study conducted to investigate the impact of choice of clustering algorithm on the overall performance of a fuzzy-based hydrologic model. The research is illustrated through a case study of developing a Takagi–Sugeno fuzzy model for reservoir inflow forecasting in the Narmada basin, India. The model was developed using two popular clustering techniques, namely Gustafson–Kessel (GK) and subtractive clustering (SC), and was extensively evaluated for performance based on various statistical indices. The results show that the model performance is comparable at a 1 h lead forecast. However, it is observed that the GK approach results in a better performance than the SC approach in computing forecasts at higher lead times. The analysis suggest that the GK method clusters the input space based on the actual pattern, since it uses a membership-grade weighted-distance measure as the measure of closeness, whereas the SC method classifies the input space more logically according to the magnitude of flow available in the data set. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS reservoir inflow forecasting; clustering technique; Takagi-Sugeno fuzzy model; GK clustering; subtractive clustering

Received 11 July 2006; Accepted 20 November 2006

INTRODUCTION

Optimal reservoir system management and operation require accurate estimates of the net inflows to the system. It is also required to plan for public safety and to address environmental issues. Mathematical models have been developed for estimating the inflows to the reservoir, based either on physical considerations or on statistical analysis. In most cases, the derivation of such models by first principles (physical, chemical, biological and/or other laws) is expensive, time consuming and involves many unknown parameters and heuristics. The traditional 'mechanistic' approach to hydrologic system modelling is based on a thorough understanding of the nature and behaviour of the actual system, and on a suitable mathematical treatment that leads to development of a model. For incompletely understood hydrologic processes, this approach may become laborious and inefficient. A large amount of process knowledge is qualitative and imprecise and, as such, cannot be readily transformed into traditional mathematical models based on differential and algebraic equations. Hence, methods for data-driven modelling and identification, which do not consider the physics of the process being modelled, are of great interest.

One such data-driven modelling technique is the fuzzy inference system. Fuzzy models can be seen as a logical data-driven modelling approach, which uses if–then rules and logical operators to establish qualitative relationships among the variables in the model. Fuzzy sets serve as a smooth interface between the qualitative variables involved in the rules and the numerical domains of the inputs and outputs of the model. The rule-base nature of the fuzzy model allows the use of information expressed in the form of natural language statements and makes the models transparent to interpretation and analysis. There has been a growing interest, recently, on the fuzzy-rule-based modelling approach in hydrologic systems. Fujita *et al.* (1992) developed a fuzzy model, as suggested by Mamdani (1977), for real-time flood forecasting. See and Openshaw (2000) combined four individual river flow forecast models in a fuzzy framework. A fuzzy conceptual rainfall–runoff model was proposed by Ozelkan and Duckstein (2001). Chang *et al.* (2001) suggested a fusion of a neural network and fuzzy arithmetic in a counter-propagation fuzzy neural network for real-time flood forecasting. Hundecha *et al.*

* Correspondence to: P. C. Nayak, Deltaic Regional Centre, National Institute of Hydrology, Kakinada 533 003, India.
E-mail: nayakpc@yahoo.co.in

(2001) developed fuzzy-rule-based models to simulate the different physical processes involved in the generation of discharge from rainfall, and incorporated them within a conceptual model. Xiong *et al*. (2001) used a Takagi–Sugeno (TS) fuzzy model combined with a fuzzy *c*-means clustering algorithm for flood forecasting. Xiong and O'Connor (2002) developed a fuzzy autoregressive-threshold updating model for real-time river flow forecasting. Chang and Chang (2001) reported the efficient use of a fuzzy model combined with a genetic algorithm for optimal reservoir operation. Hong *et al*. (2002) employed a fuzzy model to identify and predict ground water level fluctuations caused by storm water infiltration. Sen and Altunkaynak (2004) illustrated a fuzzy modelling approach for rainfall–runoff modelling. Nayak *et al*. (2004, 2005a) employed a fuzzy model based on a back-propagation algorithm for river flow forecasting.

The main advantages of fuzzy applications are that fuzzy theory is more logical and scientific in describing the properties of an object. Among the different fuzzy modelling techniques, the TS model (Takagi and Sugeno, 1985) has attracted the most attention. The TS model consists of fuzzy antecedents with fuzzy if–then rules and mathematical functions (usually linear) in the consequent part. The fuzzy sets (antecedent part) partition the input space into a number of fuzzy regions, and the consequent functions describe the system's behaviour in these fuzzy regions. The construction of a TS model is usually done in two steps. In the first step, the fuzzy sets (membership functions (MFs)) in the rule antecedents are determined. This can be done manually using knowledge about the process or by a data-driven technique if knowledge is not available *a priori*. In the second step, the parameters of the consequent functions are estimated. As these consequent functions are usually chosen to be linear in their parameters (in the case of the TS fuzzy model), the standard least-squares error (LSE) method can be applied for parameter estimation. Consequently, the effectiveness of the construction procedure lies in the identification of the antecedent MF, which is a non-linear optimization problem. For this purpose, optimization techniques based on artificial neural networks (ANNs) and genetic algorithms are being employed, but they need severe computational requirements (Jang *et al*., 1997; Jin, 2000; Roubos and Setnes, 2000). Nayak *et al*. (2004, 2005a) used a back-propagation algorithm in identifying the antecedent parameters (MF) while developing a fuzzy model for river flow forecasting. Chang *et al*. (2001) used counter-propagation neural network for automatic fuzzy if–then rules generation and parameter optimization in a TS fuzzy model for stream flow simulation. Although both studies report promising result in streamflow modelling, these models require more computational time, which is unrealistic in the context of real-time inflow forecasting. This problem is more frequent for ANN- and genetic-algorithm-based models when they handle more input vectors with lengthy data sets.

Fuzzy clustering in the Cartesian product-space of the inputs and outputs is another approach that has been quite extensively used to obtain the antecedent MFs (Babuska and Verbruggen, 1997; Babuska, 1998). Attractive features of this approach are the simultaneous identification of the antecedent MFs along with the consequent local linear models, and the implicit regularization (Johansen and Babuska, 2002). Various clustering techniques are reported in the literature, e.g. fuzzy *c*-means, mountain clustering, subtractive clustering (SC), Gustafson–Kessel (GK) (Jang, 1993; Hoppner *et al*., 1999). The choice of clustering technique for antecedent parameter identification seems to be arbitrary in most of the hydrologic applications. For instance, Xiong *et al*. (2001) used fuzzy *c*-means in their application; Chang and Chang (2001) employed SC while developing their reservoir operation model; Hong *et al*. (2002) used the GK method for forecasting ground water levels; Nayak *et al*. (2005b) used SC for river flow forecasting. However, it is apparent that, as the antecedent parameters are based solely on the identified clusters, the method used for clustering should certainly have an impact on the overall performance of the model. This heuristic is not evaluated or confirmed by empirical trials, particularly in the hydrologic modelling studies, except in the recent study by Vernieuwe *et al*. (2005). Consequently, the current study is focused on analysing the relative performance of two popular clustering techniques, i.e. GK and SC, in reservoir inflow forecasting. The choice of these two methods is according to Chiu (1994) and Babuska (1998), who report that these two techniques are the most widely applied clustering technique in fuzzy model identification.

This paper is organized as follows. First, we give a brief overview of the TS fuzzy model. Next, identification of consequent parameters using least-squares estimates is discussed, followed by a brief discussion on identification of the TS fuzzy model, which includes the optimal number of fuzzy partitions, MF parameter estimation using GK and SC techniques. Following this, the development of a fuzzy model for reservoir inflow forecasting is presented. The results are analysed and discussed in the subsequent sections. The conclusions are then presented after the evaluation of the performance by these models in inflow forecasting.

THE TAKAGI–SUGENO FUZZY MODEL

Various types of fuzzy rule-based model are reported in the literature (e.g. Mamdani and Assilian, 1975; Tsukamoto, 1979; Takagi and Sugeno, 1985), and each of them is characterized by their consequent function only. In the TS fuzzy model, the rule consequents are typically taken to be either crisp numbers or linear functions of the inputs. The first-order TS model can be 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 vector (k-dimensional) of independent variables that have

DOI: 10.1002/hyp

a casual relationship with y. Assume that n example pairs $[\boldsymbol{x}, y]$ are available for parameter estimation. Considering m rules, the mathematical functioning of the TS model is

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

where $\boldsymbol{x} \in \Re^k$ is the input variables (*antecedent*) and $y_i \in \Re$ is the output (*consequent*) of the ith rule $\boldsymbol{R}_i$. The number of rules is denoted by m and A is the antecedent fuzzy set (MF) of the ith rule, such that

$$A_i(x) : \Re^k \to [0, 1] \quad (2)$$

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

$$A_i(x) = \prod_{j=1}^{k} \mu_{ij}(x) \quad (3)$$

For the lth input $\boldsymbol{x}_l$, the total output $y(l)$ of the model is computed by aggregating the individual rule's contributions:

$$y(l) = \sum_{i=1}^{m} u_{li} y_i(l) \quad (4)$$

where u_{li} is the normalized degree of fulfilment of the antecedent clause of rule $\boldsymbol{R}_i$:

$$u_{li} = \frac{A_i(x_l)}{\sum_{i=1}^{m} A_i(x_l)} \quad (5)$$

Consequent parameter identification

The TS model is identified in two steps. First, the fuzzy antecedents A_i in the rules are determined. The next section describes how this can be done using fuzzy clustering. In the second step, the rule antecedents are kept fixed, and LSE estimation from the data is applied to determine the consequent parameters a_i^{T} and b_i of the rules. In the following, the LSE estimation approach is presented.

Let $\boldsymbol{X}_{\mathrm{e}}$ denote the matrix $[\boldsymbol{X}, 1]$ with rows $[\boldsymbol{x}_l^{\mathrm{T}}, 1]$. The activation of each rule $R_i, i = 1, 2, \ldots, m$, is gathered in Γ_i, which is a diagonal matrix in $\Re^{\mathrm{k}\times\mathrm{k}}$ having the normalized degree of fulfilment u_{li} as its lth diagonal element. Further, denote $\boldsymbol{X}'$ the matrix in $\Re^{\mathrm{n}\times\mathrm{mn}}$ composed from matrices obtained by multiplying the matrices Γ_i and $\mathbf{X}_{\mathrm{e}}$, such that

$$\boldsymbol{X}' = [\Gamma_1 \boldsymbol{X}_{\mathrm{e}}, \Gamma_2 \boldsymbol{X}_{\mathrm{e}}, \ldots, \Gamma_m \boldsymbol{X}_{\mathrm{e}}] \quad (6)$$

Denote $\boldsymbol{\theta}'$ the vector in $\Re^{\mathrm{m}\times\mathrm{n}+1}$ given by

$$\boldsymbol{\theta}' = [\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}}, \ldots, \theta_m^{\mathrm{T}}]^{\mathrm{T}} \quad (7)$$

where $\theta_i^{\mathrm{T}} = [a_i^{\mathrm{T}}, b_i]$ for $1 \le i \le m$. The model in Equation (4) can now be written as a regression model:

$$\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{\theta}' + e \quad (8)$$

where e is the approximation error. From this, the least-squares solution to the consequent parameter estimation problem can be written as

$$\boldsymbol{\theta}' = [(\boldsymbol{X}')^{\mathrm{T}}\boldsymbol{X}']^{-1}(\boldsymbol{X}')^{\mathrm{T}}\boldsymbol{y} \quad (9)$$

Antecedent parameter identification by clustering

In the previous section, how the consequent part of the TS model can be identified by the least-squares method when the antecedent MFs are given is discussed. As stated earlier, the bottleneck of the TS model development is the data-driven identification of the antecedent part that requires non-linear optimization. Hence, for this purpose, heuristic approaches like fuzzy clustering methods are often applied. Antecedent parameter identification methods based on fuzzy clustering originate from data analysis and pattern recognition, where the concept of graded membership is used to represent the degree to which a given object, represented as a vector of features, is similar to some prototypical object. The degree of similarity can be calculated using a suitable distance measure. Based on the similarity, feature vectors can be clustered such that the vectors within a cluster are as similar (close) as possible, and vectors from different clusters are as dissimilar as possible. The objective of the clustering is to partition the identification data into a number of clusters. Various clustering algorithms can be used, depending on the assumed structure of the identification data and the model one wants to obtain (Hoppner *et al*., 1999).

Fuzzy clustering is applied to discover regions in the product space of the input and output variables in which the system can be approximated locally by simple (such as linear) sub-models (Babuska, 1998). The number of clusters c determines the number of rules in the fuzzy model obtained. The fuzzy sets in the antecedent of the rules are obtained from the partition matrix by projection onto the antecedent variables. The point-wise fuzzy sets obtained are approximated by suitable parametric functions. The consequent parameters for each rule can then be obtained by the LSE method.

Fuzzy partitioning

From the available input–output data pairs, the regression matrix $\boldsymbol{X}$ and the output vector $\boldsymbol{y}$ are constructed:

$$\boldsymbol{X}^{\mathrm{T}} = [x_1, x_2 \ldots, x_n] \qquad \boldsymbol{y}^{\mathrm{T}} = [y_1, \ldots, y_n] \quad (10)$$

where n 4 k is the number of samples used for identification. The antecedent fuzzy set A_i in Equation (1) is determined by means of fuzzy clustering in the product space of the system's inputs and outputs. Hence, the data set $Z \in R^{(k+1)\times n}$ to be clustered is represented as a $(k + 1) \times n$ data matrix composed from $\boldsymbol{X}$ and $\boldsymbol{y}$:

$$Z^{\mathrm{T}} = [\boldsymbol{X}, \boldsymbol{y}] \quad (11)$$

where each column z_j, $j = 1, 2, \ldots, n$, of Z contains an input–output data pair: $z_j = [x_j^{\mathrm{T}}, y_j]^{\mathrm{T}}$. When clustering is

DOI: 10.1002/hyp

applied to the modelling and identification of the dynamic systems, the column Z contains samples of time signals, and the rows are typically the input and output variables observed in the system.

Given Z and an estimated number of clusters m, fuzzy clustering partitions Z into m fuzzy clusters. A fuzzy partition can be represented as an $n \times m$ matrix U, whose elements $u_{ji} \in [0, 1]$ represented the membership degree of z_j in cluster i. Hence, the ith column of U contains values of the ith MF in the fuzzy partition, which is taken to be a point-wise representation of the antecedent fuzzy set A_i of the ith rule in Equation (1). The sum of each row of U is constrained to unity, but the distribution of the membership among the m fuzzy subsets is not constrained. Also, there can be no empty clusters and no cluster may contain all the objects. This means that the membership degrees in the partition matrix U are normalized; and for the identification data, the membership values u_{ji}, correspond to the normalized degree of fulfilment of the rule antecedent in Equation (5). Thus, the n membership values in the ith column u_i of the fuzzy partition matrix correspond to those in the diagonal matrix Γ_i used in Equation (6) to construct the regression matrix $\boldsymbol{X}'$ for the least-squares parameter estimation problem in Equation (9). Thus $\Gamma_i = \text{diag}(u_i)$, where $\text{diag}(u_i)$ denotes a diagonal matrix with the jth element u_{ji} of the vector u_i as the jth diagonal element.

Gustafson–Kessel algorithm

Gustafson and Kessel (1979) extended the standard fuzzy c-means algorithm by employing an adaptive distance norm, in order to detect clusters of different geometrical shapes in one data set. In adaptive norm clustering, each cluster has its own norm-inducing matrix D_i, which is obtained from the covariance of the clusters. The distance of a data point z_k to a cluster centre u_i is given by the inner product norm:

$$d_{ki}^2 = (z_k - u_i)^{\mathrm{T}} D_i (z_k - u_i) \tag{12}$$

where $\boldsymbol{V} = [u_i, u_2, \ldots, u_m]$ is a vector of cluster prototypes $u_i \in R^{(n+1)}$ that have to be determined. The GK fuzzy clustering algorithm determines $\boldsymbol{V}$ based on the minimization of

$$J(X; U, V) = \sum_{i=1}^{m} \sum_{k=1}^{n} (u_{ki})^{\phi} d_{ki}^2 \tag{13}$$

where $\phi \in (1, \infty)$ is a weighting exponent that determines the fuzziness of the clusters. The minimization of Equation (13) represents a non-linear optimization problem, which is solved in an iterative manner (Takagi and Sugeno, 1985). The cluster algorithm stops when a predetermined stopping criterion is fulfilled. The algorithm is described in Appendix A.

Subtractive clustering algorithm

The SC 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 the grid points defined on the data space. As a result, clusters are elected from the system training data according to their potential. The algorithm considers each data point ($\{x_1, x_2, \ldots, x_n\}$ in k-dimensional space) as a potential cluster centre, and estimates the 'potential' of the data point x_i as

$$P_i = \sum_{j=1}^{n} e^{-\alpha \|x_i - x_j\|^2} \tag{14}$$

The 'potential' is a measure of the distance between a pattern and the rest, and it takes higher values when more neighbours exist. In Equation (14), $\alpha = 4/r_{\mathrm{a}}^2$, where r_{a} is effectively the radius defining a neighbourhood; data points outside this radius have 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. 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

$$p_i = p_i - p_1^* e^{-\beta \|x_i - x_1^*\|^2} \tag{15}$$

where $\beta = 4/r_{\mathrm{b}}^2$ and r_{b} is a positive constant. After reduction is done, the data point with the highest potential is selected as cluster x_2^*. Next, it 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 $P_k^*/P_1^* < \varepsilon$. The choice of ε is an important factor affecting the results: if ε is too large, then too few data points will be accepted as cluster centres; if ε is too small, then too many cluster centres will be generated. 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

$$\mu_i = e^{-\alpha \|y - y_i^*\|^2} \tag{16}$$

where α is the constant defined by Equation (14). The output vector may be computed using Equation (4). Equations (16) and (4) provide a simple and direct way to translate a set of cluster centres into a TS fuzzy model. Readers are referred to Chiu (1994) for more details about SC.

APPLICATION OF TAKAGI–SUGENO MODEL TO RESERVOIR INFLOW FORECASTING

The above-discussed methods of fuzzy model identification have been employed to develop a river flow forecasting model for the Narmada River basin, India. The Narmada River originates at Amarkantak at an elevation of 1057 m above mean sea level in the Shahdol district of Madhya Pradesh State in India. The river travels a distance of 1312 km before it enters the Gulf of Cambay in the Arabian Sea near Bharuch City in Gujarat State. The Narmada basin extends over an area of 98 796 km^2 and lies between longitudes 72°32′E and 81°45′E and between

DOI: 10.1002/hyp

latitudes 21°20′N and 23°45′N. Three reservoirs are situated in the upstream of Hoshangabad city, namely Bargi, Barna and Tawa. The releases from these reservoirs and the flow from the intermediate catchment cause flooding at Hoshangabad City during the monsoon season. The release from Bargi Reservoir may increase the inundation at Hoshangabad City. Therefore, efficient operation of Bargi Reservoir is essential, which warrants accurate forecasts of flow into the reservoir. Flood forecasts at Mandla gauging site are sufficient to operate the reservoir efficiently during the monsoon season, since it is located at the tail end of the reservoir. This provided an impetus to develop an inflow forecasting model for the basin up to the Mandla gauging site. The catchment area up to Mandla is 13 120 km^2 and is shown in Figure 1. The rainfall and runoff data available for the monsoon season during the years 1989 to 1993 on an 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 Mohegaon, Hridyanagar and Dindori, are also used in the study.

Takagi–Sugeno model development

The development of the TS fuzzy model constitutes three steps: (1) selection of input–output variables; (2) selection of model structure and estimation of its parameters; (3) validation of the model identified. The goal is to identify a TS fuzzy model for the reservoir inflow dynamics:

$$y(k+t) = f\begin{pmatrix} y(k), y(k-1), \ldots, y(k-n_y+1) \\ u(k-n_{k1}), \ldots, u(k-n_{k1}-n_{k2}+1) \end{pmatrix} \quad (17)$$

where $y(k)$ is the historic inflow time-series, $u(k)$ is the time-series of any exogenous influencing variable (in the current study, rainfall and upstream runoff values), t is the desired forecast lead time, and $f(.)$ is a fuzzy model of the TS type. In Equation (17), n_y, n_{k1} and n_{k2} represent the lag of the corresponding variables.

One of the major concerns in fuzzy modelling is the identification of the appropriate input vector that represents the transformation of rainfall into runoff for a particular basin. It may constitute rainfall and/or runoff values at different lag times. However, how many antecedent rainfall/runoff values should be included in the vector is not known *a priori*. When modelling a hydrologic system, usually a large number of input variables are dealt with. To obtain a simple and transparent, yet accurate and reliable, model, the most important variables have to be determined. In fuzzy modelling, no rigorous criteria for input selection exist. Hence, a statistical approach suggested by Sudheer *et al*. (2002) to identify the appropriate input vector that can best represent the process is employed in this study. 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-correlation, autocorrelation, and partial autocorrelation

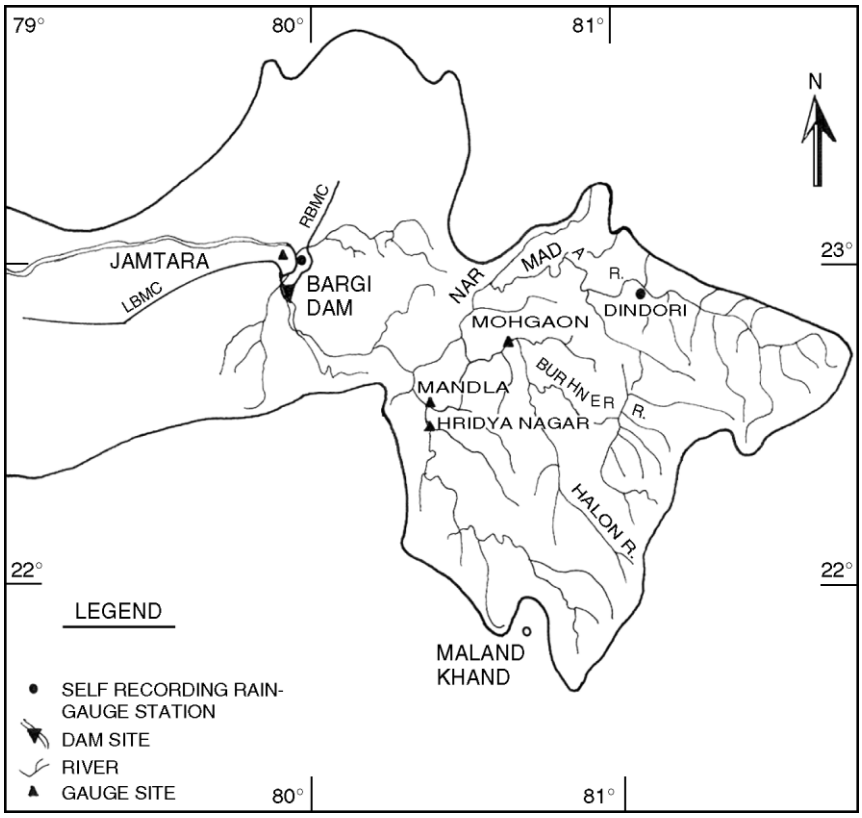

Figure 1. Basin map of River Narmada up to Bargi Reservoir

DOI: 10.1002/hyp

between the variables in question, and the current study employed this methodology 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, then 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 fall between 0 and 1 as the activation function warrants. The total available data has been divided into two sets, i.e. 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.

Calibration of the Takagi–Sugeno fuzzy models

Identification of TS fuzzy models combined with the clustering techniques is carried out using the MATLAB fuzzy logic toolbox. The models are trained using data for three years (1989–1991) and validated on the rest of the data (1992–1993). The input vector identified included a total of 16 variables, which are detailed in Table I, in which Q represents the runoff, R is the rainfall and t indicates the prediction time horizon. The output of the model is the flow at Mandala. The nomenclature used in the present study for the TS fuzzy model integrated with GK clustering and with SC techniques are the TS-GK model and the TS-SC model respectively.

As the TS-GK model is sensitive to the number of clusters, the optimal number of clusters has been identified through a trial-and-error procedure. This is achieved by varying number of cluster from 2 to 10, and developing a TS-GK model each time. In all experiments reported in this paper, the fuzziness parameter employed was $\phi = 2$, and a termination tolerance of $\delta = 0{\cdot}001$ was used. During calibration, the performance of the model was monitored using the efficiency index given in Appendix B.

In the case of the TS-SC model, the radius of influence r_a of the cluster centre is fixed by various trials. The value of r_a is varied from 0·10 to 1·0 with a step size of 0·05 at each stage. In this study, the effectiveness of the fuzzy models in forecasting flows at a higher forecast time horizon is also evaluated by developing models that forecasts flows up to 12 h in advance.

Table I. Input vector for the identified model

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

Model evaluation

To evaluate the model performance, different evaluation measures are considered and the resulting hydrographs from these models are analysed statistically. Global error statistics, which include the root-mean-square error (RMSE) between the computed and observed runoff, the coefficient of correlation (CORR), and the model efficiency (EFF), etc., provide relevant information on overall performance, but they do not provide specific information in the flood period, which in a flood forecasting context is of critical importance. Therefore, peak flow criteria (PFC) and low flow criteria (LFC) are also employed, in addition to the global measures, as a performance indicator for high flow and low flow periods. In addition to these indices, certain event-specific evaluation measures are also employed to evaluate the model performance, such as the percentage error in volume under the hydrograph (termed the index of volumetric fit (IVF)) and time difference to peak flow. Finally, the error distribution at different threshold levels for both the models is also compared. The equations for computing all these evaluation measures are presented in Appendix B.

RESULTS AND DISCUSSION

Model calibration

The variation in efficiency along with the number of cluster centres in the TS-GK model is presented in Figure 2. It is observed from Figure 2 that the efficiency increases as the number of clusters increases during calibration data set, and the trend is reverse for the validation data set. It is noted that the slope of the learning curve changes after five clusters. Judging from the trial-and-error result, five clusters seems to be a suitable choice and is selected for further analysis.

Figure 3 depicts the variation of efficiency along with the cluster radius r_a in the case of TS-SC model. It is observed from the experiment that the normalized firing strength of the fuzzy model is zero up to r_a of 0·2. From Figure 3 it is evident that model performance, increases up to a cluster radius of 0·4, but thereafter there is a quick deterioration observed in performance. Therefore, it is decided that a cluster radius of 0·4 which resulted in five fuzzy rules may be appropriate for inflow forecasting, and this is selected for further analysis.

Evaluation of selected models

The values of global evaluation measures during the calibration and validation periods for both the models at 1 h lead forecast time are summarized in Table II.

DOI: 10.1002/hyp

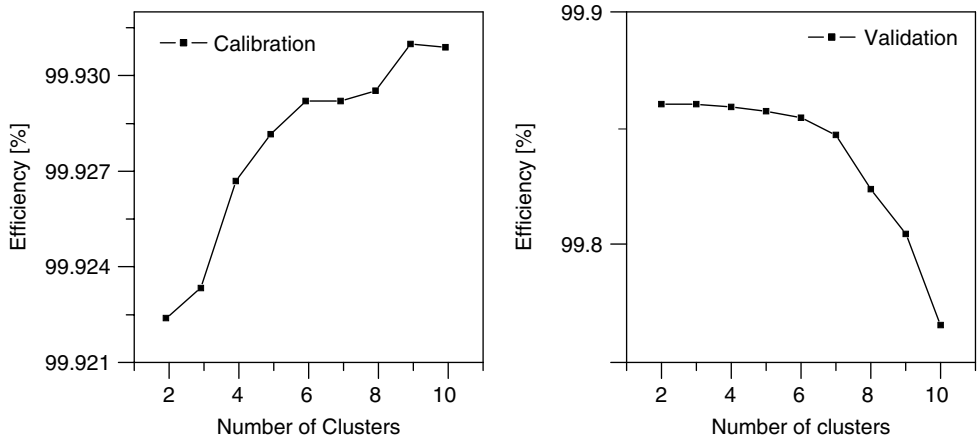

Figure 2. Variation of efficiency along with the number of clusters in the case of the TS-GK model

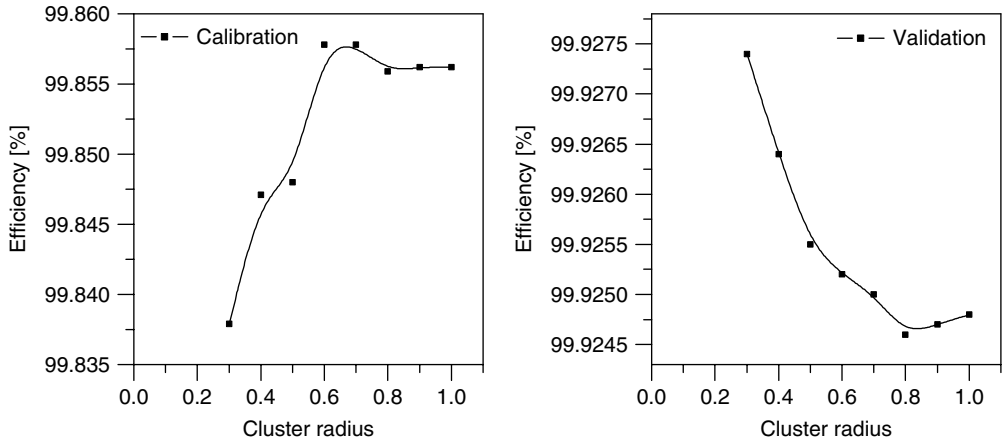

Figure 3. Variation of efficiency along with cluster radius for TS-SC model

The results presented in Table II indicate that the performance of both models is comparable. The RMSE value varies from 33 to 41 $m^3 s^{-1}$, indicating a very good performance compared with the mean discharge of 592·98 $m^3 s^{-1}$. Both the models have comparable efficiency index (>99%), which indicates a very satisfactory model performance (Shamseldin, 1997). A similar argument also holds well with other performance indices.

Figure 4 presents the scatter plots of observed and computed inflows during the validation period for the 1 h ahead forecast. These plots give a clear indication of the relative skill of each of the models across the full range of flows, and it is interesting to note that most of the flows tend to fall close to the 45° line, thus showing a good agreement between observed and forecasted flows. The results clearly illustrate that both the clustering approaches are quite competent in forecasting inflow to the reservoir at 1 h lead time. Both the models are further evaluated for their effectiveness to forecast flows at higher lead times, such as 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12 h in advance. The variations of RMSE and efficiency for different lead forecasts are presented in Figure 5. It is observed that the RMSE index is somewhat linearly increasing along the prediction time horizon during calibration and validation data set for both the models. However, it is evident from Figure 5 that, as the lead time increases, the model accuracy is better for the TS-GK model than the TS-SC model. A similar observation holds well with the efficiency statistics also (see Figure 5).

Table II. Statistical indices for 1 h ahead forecast

Statistical index[a]	TS-GK model		TS-SC model	
	Calibration	Validation	Calibration	Validation
CORR	0·9991	0·9968	0·9985	0·9968
Efficiency	99·80	99·34	99·69	99·34
RMSE	33·79	35·02	41·81	35·22
NRMSE	0·1281	0·1883	0·1585	0·1894
SSE	4552400	4734900	6969600	4790000
SEE	33·86	35·09	41·89	35·30
NSEE	0·1284	0·1887	0·1589	0·1898
PEMF	−0·9800	−1·5398	−1·9921	−1·8223
AARE	0·0173	0·0213	0·0173	0·0222
MBE	0·0823	0·9682	0·3835	0·6457
NMBE	0·0003	0·0052	0·0015	0·0035
NS	0·0452	0·0811	0·0559	0·0816
MAE	7·0997	7·0571	7·4625	6·7179

[a] NRMSE, NSEE and NMBE are the normalized values of RMSE, SEE and MBE respectively.

DOI: 10.1002/hyp

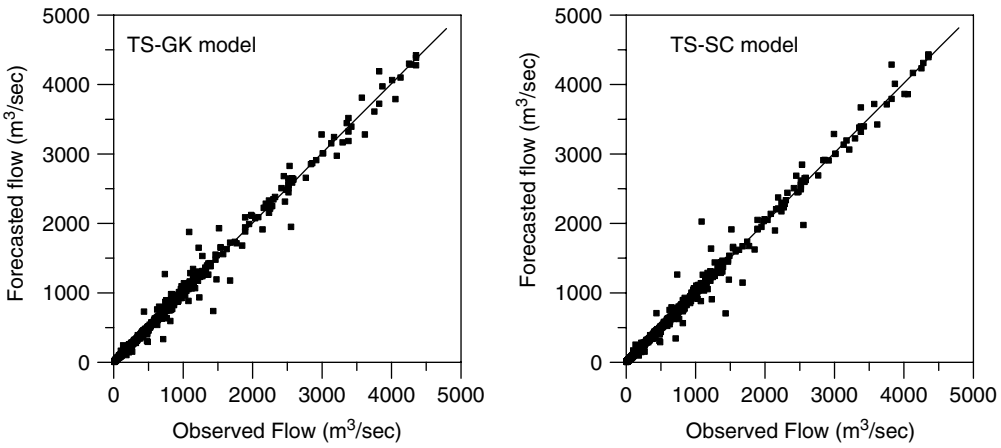

Figure 4. Scatter plots of observed and forecasted reservoir inflow at 1 h lead forecast for TS-GK and TS-SC models

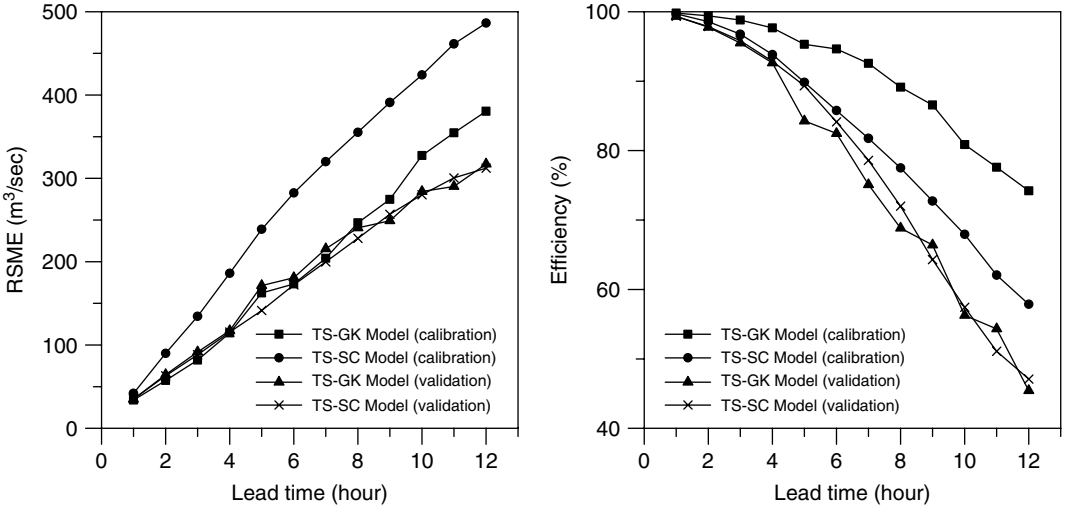

Figure 5. RMSE and efficiency versus prediction time horizon

To examine the model prediction accuracy for the peak discharge during the validation period, the computed values of percentage error in peak flow estimation (PEMF) are presented in Figure 6. PEMF is the ratio between the observed and the forecasted maximum peak flow minus unity, expressed as a percentage. PEMF statistics closer to zero indicate the best fit. The PEMF plot (Figure 6) indicates that the TS-GK model is better in peak flow prediction. It may be noted that the TS-GK model performs better than the TS-SC model in peak flow estimation irrespective of the forecast lead time. A further analysis of the results indicates that larger prediction errors are associated with the high inflow period (peak flow of 4354·46 m^3 s^{-1}) around 11 September 1992. However, a close examination of the predicted values during this period suggests that the TS-GK model is able to predict the high inflows more accurately than the TS-SC model.

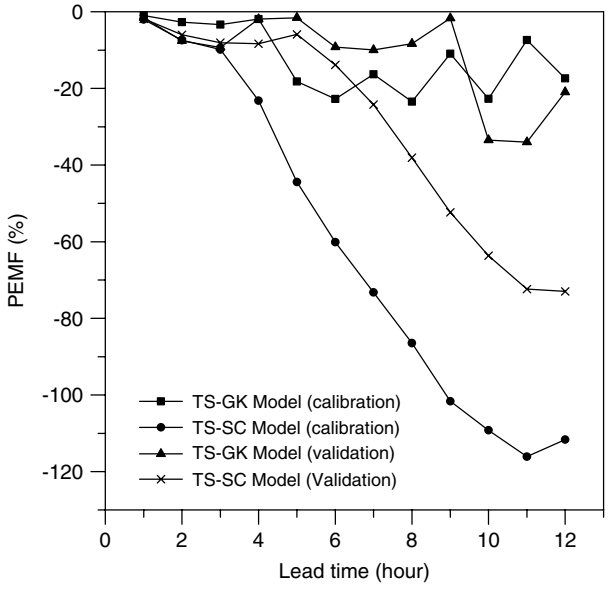

Figure 6. PEMF plots for different lead times

In order to evaluate the prediction accuracy during high flow and low flow periods, PFC and LFC are estimated. The PFC provides a more accurate measure of the model performance than RMSE for the flood period, and LFC is a better performance indicator for the low flow period (Coulibaly *et al.*, 2001). The PFC and LFC statistics for various models are shown in Figure 7. It may be noted

DOI: 10.1002/hyp

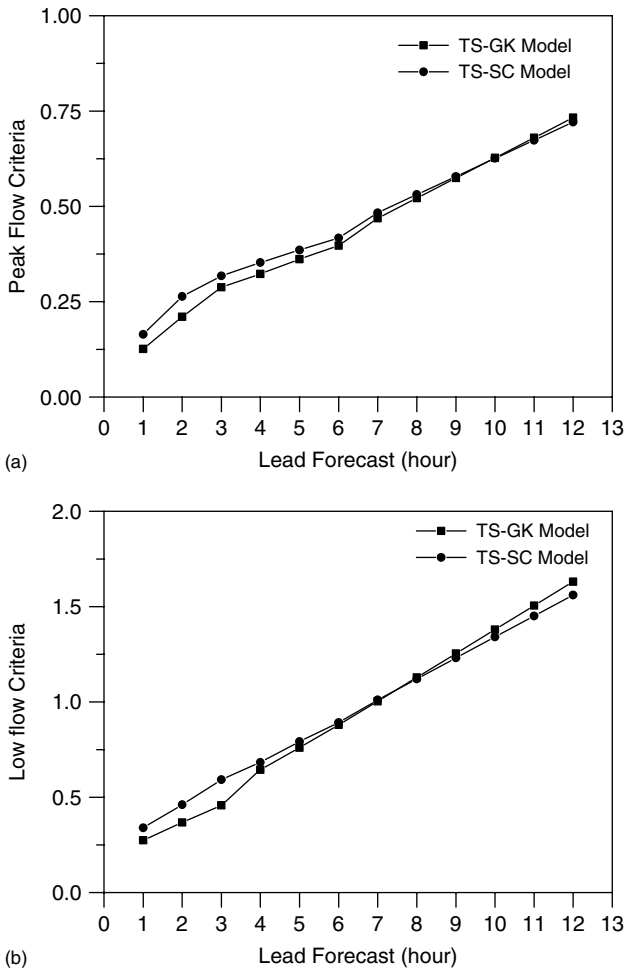

Figure 7. Plots showing (a) PFC and (b) LFC statistics during validation period

that PFC and LFC equal to zero represents a perfect fit. Figure 7 clearly indicates that the TS-GK model is better than the TS-SC model in peak flow as well as low flow forecasting. It may be noted that as the lead-forecast horizon increases, so the PFC and LFC statistics also increase. Note that the LFC value becomes unity after 6 h ahead forecast for both the models, indicating that both the models are unable to predict low flow beyond 6 h.

It may be noted that, in the inflow forecasting context, global statistics as well as combination with event-based performance criteria provide a better insight into the modelling performance of individual solutions. Consequently, the performance of the TS fuzzy models is further evaluated using various event-specific evaluation measures. The indices employed are (i) IVF, (ii) relative error and (iii) time difference to peak flow criteria. Qualitative comparisons between the results forecasted by the two methods are presented in Table III for six typical hydrographs during the validation period. The IVF statistic measures the percentage error in volume under the observed and forecasted hydrographs; positive values indicate overestimation and negative values indicate underestimation. The IVF performance ranges from $-0{\cdot}15\%$ to $+15{\cdot}89\%$, which indicates that the forecasts by both the methods are consistently good. It is also observed that overestimation is quite prominent for low flow events, which clearly suggests that both the models have some difficulty in forecasting low flows. This observation confirms the earlier discussion on LFC statistics. From Table III, it is also observed that TS-GK model performance is better according to the index, i.e. the relative error of peak (for different peak flows). It is to be noted that that the TS-SC fails to preserve the time to peak characteristics of a hydrograph when the lead time increases. The results, in general, indicate that TS-GK performs better than the TS-SC model in terms of the global evaluation criteria, as well as event-specific evaluation criteria. It is also observed that both the models are unable to predict low flow satisfactorily. Figure 8 shows the observed and forecasted hydrographs (by both the models) corresponding to an event that has a peak discharge of $4354{\cdot}46\ \mathrm{m^3\ s^{-1}}$ for a 1 h lead forecast. It may be noted that the TS-GK model forecasts are better than those of the TS-SC model are.

As the global evaluation measures do not give any idea about the distribution of errors, the distribution of error at different threshold error levels (TL as a percentage) and the average absolute relative error (AARE) statistics are computed for both the models, which is expressed in $x\%$ level (TL) as $\mathrm{TS}_x = (Y_x/n) \times 100$, where Y_x is the number of computed streamflows (out of n total computed) for which the absolute relative error is less than $x\%$ for the model. The distribution of error is plotted for both the models for different lead forecasts in Figure 9, which provides an indication of performance in terms of the distribution of residual errors at different forecast lead times. It is observed from Figure 9 that the distribution of errors is similar for both the models at 1 h lead forecast. However, as the forecast lead time increases, so the TS-GK model tends to have a high frequency with low errors. The results indicate that the TS-GK model consistently has the smallest bias compared with the TS-SC model.

From the foregoing discussions it is evident that the TS-GK model performs better than the TS-SC model, especially at higher forecast lead times. The results indicate that the procedure employed for estimating the antecedent parameters of a fuzzy model certainly has an impact over the model performance. It may be noted that the SC algorithm for clustering data points is basically based on a transformation of Euclidian distance as the measure of closeness between the data points, which turns out to be geometry based. However, the GK clustering considers the covariance matrix between the clusters along with the adaptive distance norm as a measure of closeness and, hence, produces more representative sample points as cluster prototypes. In other words, the GK method allows each cluster to adapt the distance norm to the local topological structure of the data, since it is weighted by the MF grade. An advantage of the GK cluster is that it can

DOI: 10.1002/hyp

Table III. Comparison of index of volumetric fit, relative error and time difference to peak flow: validation period

Observed flow ($m^3 s^{-1}$)	Lead forecast (h)	TS-GK model			TS-SC model		
		IVF (%)	Relative error (%)	Time difference to peak flow	IVF (%)	Relative error (%)	Time difference to peak flow
4354.46	1	0·36	1·54	0	0·54	1·82	0
	2	1·12	5·74	0	0·96	6·01	0
	3	−1·97	7	1	0·66	8·09	3
	4	−6·43	6·2	1	−0·77	8·32	2
	5	−2·51	5·34	0	−3·66	5·92	1
	6	−4·39	3·62	1	−7·92	4·08	0
3379.29	1	0·88	4·11	1	0·23	8·59	1
	2	1·27	10·31	0	0·41	19·43	0
	3	−0·37	−1·96	0	0·32	20·6	1
	4	−1·25	15·48	1	−0·02	14·97	1
	5	6·27	12·87	0	−0·68	15·81	2
	6	6·03	−2·22	1	−1·65	20·34	1
2569.37	1	0·6	2·96	2	0·56	3·14	2
	2	−0·15	6·75	0	1·44	13·42	2
	3	−1·92	2·53	2	2·26	0·69	2
	4	−3·89	7·81	2	2·89	6·64	1
	5	3·7	8·51	0	3·26	12·8	0
	6	1·78	8·08	0	3·5	9·99	1
1427.91	1	0·54	1·11	2	−0·39	2·5	0
	2	1·9	18·19	2	−1·5	25·51	1
	3	4·12	18·97	0	3	41·18	0
	4	2·5	17·71	0	−3·16	26·25	1
	5	5·57	14·23	1	−3·82	17·52	2
	6	6·9	16·74	2	−4·15	29·59	3
975.76	1	2·13	6·22	0	1·7	9·38	0
	2	4·11	14·12	1	2·97	16·78	1
	3	8·29	20·97	0	5·6	25·92	0
	4	13	33·05	0	8·85	51·61	0
	5	11·7	34·93	2	12·36	49·8	1
	6	13·7	23·97	2	15·89	32·06	2
420.62	1	1·34	3·94	1	1·11	5·29	1
	2	2·67	7·54	2	2·4	8·73	2
	3	5·81	10·41	2	3·49	18·53	2
	4	6·76	12·25	3	4·43	23·91	3
	5	5·42	12·84	4	4·67	18·2	4
	6	5·51	13·44	1	4·24	23·83	4

detect clusters of different shapes and orientation in the data set.

The coordinates of the cluster centres identified by both the models in the current study are presented in Table IV. Note that only a single dimension corresponding to each variable is presented in Table IV. It can be seen from Table IV that the SC method does not consider the rainfall variability in the data set while selecting the clusters, whereas GK method considers it in approximately two different ranges. It may also be noted that the GK method clusters the data according to the actual magnitude of data available in the data set. It was observed in the time-series data of inflows that the frequency of inflow was very high within the range 150–170 $m^3 s^{-1}$, and the GK method classifies this range into sub-domains effectively. However, the SC method classifies the input space in a more logical way by arranging the data in the order of magnitude. In order to reinforce the observations and findings, two MF plots are presented, in the Figures 10 and 11, for 1 h antecedent flow at the Mandla gauging site. The SC algorithm with Euclidean distance favours hyperspherically shaped clusters of equal size, as is evident from Figure 10. This has the undesirable effect of splitting large clusters, as well as elongated clusters, under some circumstances. Indeed, it has been noted that most clusters in real data sets are neither well isolated nor have a spherical shape. It is apparent from Figure 11 that the GK clustering algorithm classifies the flow series into a different domain of ranges than the SC method does. This results in effective linear separable sub-domains by the GK method, compared with the SC method, and hence gives a better performance.

This can be seen from Table V, wherein the values of the consequent parameters of the fuzzy model are presented for the TS-SC and TS-GK models. It can be observed from Table V that, for the input variable $Q(t-1)$ at the Mandla gauging site, the weights (magnitude of the consequent parameter) associated with the TS-SC model are of similar order for all five rules. In contrast, the TS-GK model gives a clearly distinct range of weights for the same variable in the five rules. Similarly, the input variable $Q(t-13)$ at Dindori plays an

DOI: 10.1002/hyp

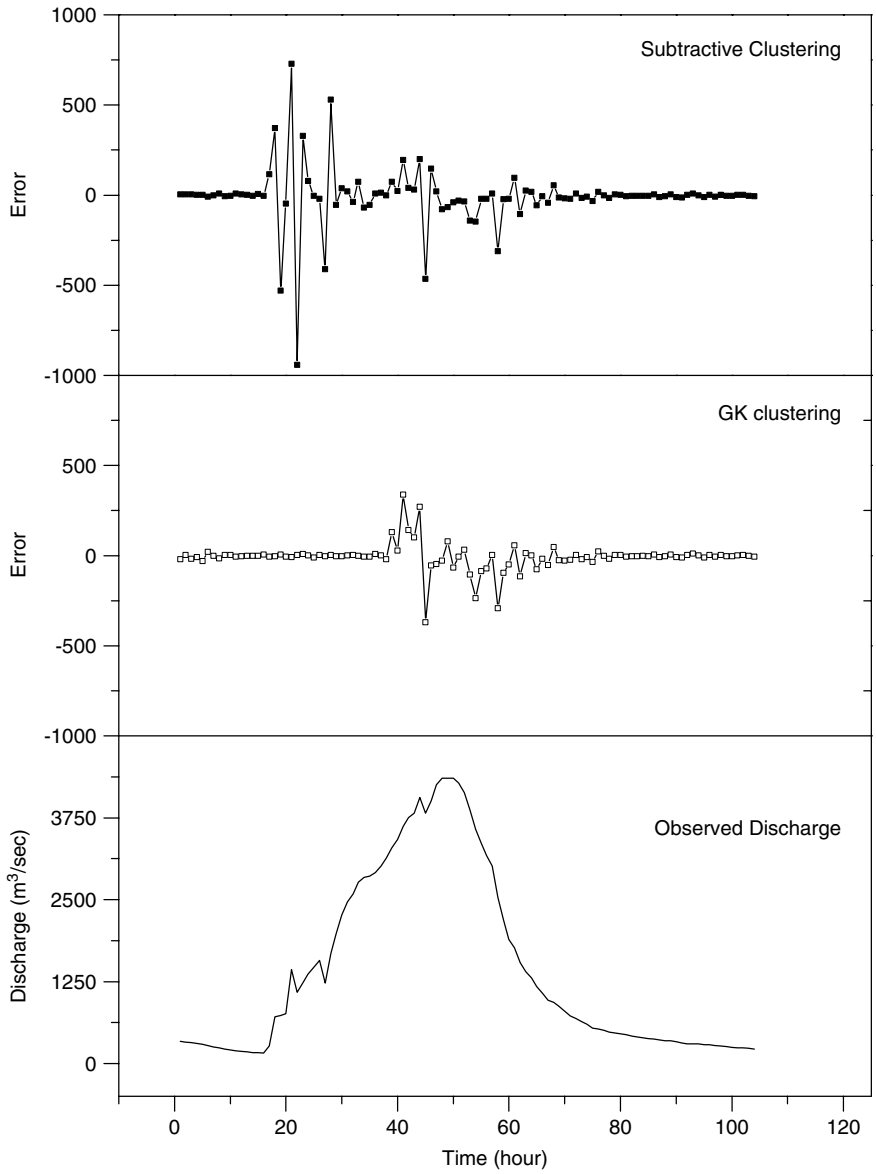

Figure 8. Plot showing the error in flow estimation by the TS-SC and TS-GK models for a hydrograph corresponding to the peak flow (4354·46 m^3 s^{-1}) during 1 h lead forecast (validation period)

Table IV. Coordinates of cluster centres from SC and GK methods (the values are in the transformed-standardized-scaled domain; the values in parentheses represent the value of the variable in flow domain in m^3 s^{-1})

	Rainfall	Flow at Mohegaon	Flow at Hridaynagar	Flow at Dindori	Flow at Mandla
GK	0·0371 (0·0002)	0·3817 (30·24)	0·3691 (43·43)	0·1350 (37·30)	0·2740 (26·33)
	0·2237 (0·0018)	0·5875 (190·06)	0·5734 (200·11)	0·2743 (76·47)	0·4837 (152·82)
	0·2345 (0·0020)	0·5929 (200·52)	0·5765 (204·80)	0·2981 (86·45)	0·4897 (160·71)
	0·2271 (0·0018)	0·5775 (174·69)	0·5638 (186·25)	0·2555 (69·41)	0·4712 (137·61)
	0·2376 (0·0021)	0·5949 (204·15)	0·5787 (208·20)	0·3203 (96·93)	0·4941 (166·75)
SC	0·0000 (0·0001)	0·5284 (112·53)	0·5106 (125·12)	0·1984 (51·72)	0·4086 (81·41)
	0·0000 (0·0001)	0·3178 (17·06)	0·3150 (28·98)	0·0947 (30·31)	0·2142 (15·95)
	0·0000 (0·0001)	0·4333 (48·01)	0·3964 (53·26)	0·1605 (42·54)	0·3089 (35·28)
	0·0000 (0·0001)	0·6046 (222·67)	0·5783 (207·58)	0·2955 (85·30)	0·4895 (160·43)
	0·0000 (0·0001)	0·6534 (344·75)	0·6178 (279·92)	0·3661 (122·73)	0·5968 (394·56)

important role in the first two rules of the TS-SC model, since for the other three rules the consequent parameter value is negligibly small. However, in the TS-GK model, the contribution of this input variable in producing the output is important in all five rules. In general, it is seen that the consequent parameters of the TS-GK model in each rule are distinct compared with the TS-SC model for all variables. This result reinforces the earlier conclusion

DOI: 10.1002/hyp

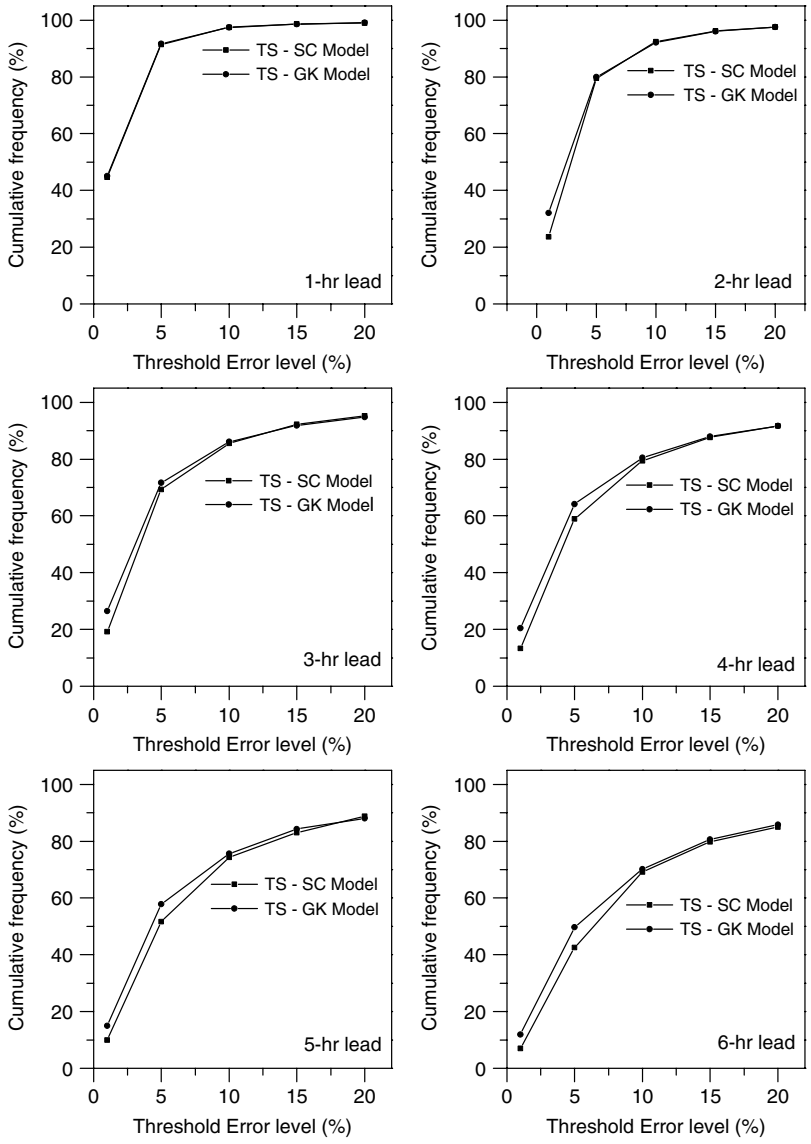

Figure 9. Distribution of error plots during validation period

that the TS-GK model results in effective linear separable sub-domains and, hence, results in a better performance than the TS-SC model does.

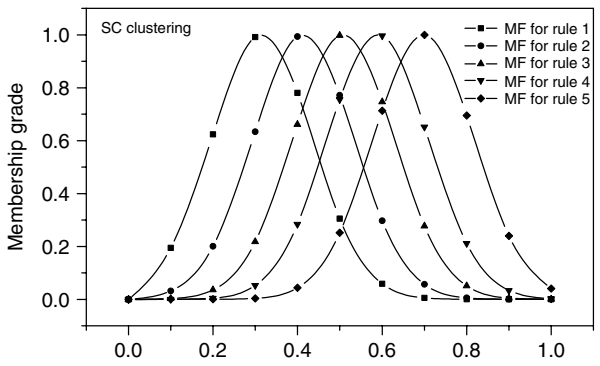

Figure 10. Membership grade from SC method for Mandla 1 h antecedent flow for five rules

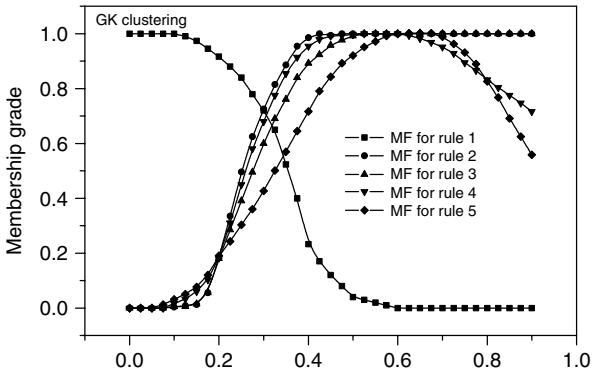

Figure 11. Membership grade from GK method for Mandla 1 h antecedent flow for five rules

SUMMARY AND CONCLUSIONS

This paper discusses the results of a research study conducted to investigate the effect of choice of clustering algorithm in the performance of a fuzzy-based reservoir inflow forecast model. The primary motivation for this

DOI: 10.1002/hyp

Table V. The values of consequent parameters for both TS-SC and TS-GK models for comparison

Rule	Rainfall		Mohegaon		Hridayanagar	Dindori		Mandala					Const.
	$R(t-18)$	$R(t-17)$	$R(t-16)$	$Q(t-6)$	$Q(t-5)$	$Q(t-4)$	$Q(t-1)$	$Q(t-11)$	$Q(t-12)$	$Q(t-13)$	$Q(t-6)$	$Q(t-5)$	$Q(t-4)$	$Q(t-3)$	$Q(t-2)$	$Q(t-1)$	
GK clustering																	
1	−0·0008	0·0013	−0·0011	0·022	−0·071	0·048	0·002	−0·014	0·020	−0·0051	0·0213	−0·047	−0·003	0·097	−1·125	2·062	−0·002
2	−0·0014	0·0094	0·0006	−0·768	0·917	−0·090	−0·037	−0·067	0·189	−0·1284	0·0130	−0·447	0·898	0·938	−2·909	2·478	0·002
3	0·0073	−0·0087	−0·0011	0·040	0·173	−0·190	0·053	0·001	−0·115	0·1484	−0·0177	0·243	−0·566	−0·934	2·091	0·097	−0·015
4	0·0008	−0·0066	0·0026	0·696	−0·924	0·191	0·041	0·082	−0·134	0·0766	0·0031	0·378	−1·007	−0·663	1·426	0·896	−0·018
5	−0·0041	0·0078	0·0061	0·164	−0·996	0·932	−0·001	0·008	0·023	−0·0138	0·2121	−0·140	0·765	−0·378	−3·247	3·796	−0·089
Subtractive clustering																	
1	0·0008	0·0006	0·0010	−0·143	0·142	0·023	0·010	−0·012	0·008	0·0127	0·0003	0·008	0·016	−0·087	−0·553	1·576	−0·002
2	0·0015	−0·0016	0·0006	−0·038	0·062	0·003	0·020	0·016	−0·030	0·0269	0·0244	0·022	−0·142	−0·168	−0·243	1·458	−0·008
3	0·0000	0·0001	0·0004	−0·088	0·029	0·073	0·012	0·014	−0·014	0·0009	0·0619	−0·223	0·165	−0·102	−0·295	1·354	0·001
4	−0·0007	0·0007	−0·0018	0·036	−0·091	0·035	−0·018	0·000	0·016	0·0017	−0·0036	0·039	−0·169	0·321	−1·453	2·294	0·004
5	0·0007	−0·0003	0·0008	−0·003	−0·026	0·035	0·010	−0·012	0·013	0·0079	−0·0139	−0·001	−0·125	−0·140	−0·012	1·268	−0·001

study came from the observation that little discussion has been provided in research papers about the impact of antecedent parameter estimation procedure on the model performance, despite a huge amount of studies. The research is illustrated through a case study of developing a TS fuzzy model for reservoir inflow forecasting in the Narmada basin, India. The hourly data on rainfall and runoff available for the basin have been employed to develop fuzzy models that forecast flows up to 12 h in advance. The model was developed using two popular clustering techniques, namely GK and SC, and was extensively evaluated for performance based on various statistical indices. The choice of these two clustering algorithms for evaluation is based on some reported studies that these two algorithms are popularly used. The performance evaluation indices included global evaluation measures, event-specific evaluation measures and distribution of forecast errors. The analysis suggests that the choice of the clustering algorithm may not have a significant impact on the model performance if the forecast is required at 1 h in advance. However, the performance at higher lead times very much depends on the clustering algorithm, and in the current study it is observed that the GK method is consistently better than the SC method at providing reasonable forecasts up to 6 h in terms of most of the performance indices considered. It is observed that the GK method clusters the data according to the actual magnitude of data available in the data set, whereas the SC method classifies the input space in a more logical way by arranging the data in the order of magnitude, and is plausibly the reason for a better performance by the GK-based fuzzy model.

REFERENCES

Babuska R. 1998. *Fuzzy Modeling for Control*. Kluwer: Boston, MA.

Babuska R, Verbruggen HB. 1997. Constructing fuzzy models by product space clustering. In *Fuzzy Model Identification: Selected Approaches*, Hellendoorn H, Driankov D (eds). Springer: Berlin; 53–90.

Chang F-J, Hu H-F, Chen Y-C. 2001. Counterpropagation fuzzy-neural network for streamflow reconstruction. *Hydrological Processes* **15**: 219–232.

Chang L-C, Chang F-J. 2001. Intelligent control for modelling of real-time reservoir operation. *Hydrological Processes* **15**: 1621–1634.

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

Coulibaly P, Anctil F, Bobee B. 2001. Multivariate reservoir inflow forecasting using temporal neural network. *Journal of Hydrologic Engineering* **6**: 367–376.

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

Gustafson DE, Kessel WC. 1979. Fuzzy clustering with a fuzzy covariance matrix. In *Proceedings of IEEE CDC*, San Diego, CA; 761–766.

Hong Y-S, Rosen MR, Reeves RR. 2002. Dynamic fuzzy modeling of storm water infiltration in urban fractured aquifers. *Journal of Hydrologic Engineering* **7**: 380–391.

Hoppner F, Klawonn F, Kruse R, Runkler T. 1999. *Fuzzy Cluster Analysis and Methods for Classification, Data Analysis and Image Recognition*. Wiley.

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

Jang J-SR. 1993. ANFIS: adaptive network based fuzzy inference system. *IEEE Transactions on Systems, Man and Cybernetics* **23**: 665–683.
Jang J-SR, Sun C-T, Mizutani E. 1997. *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Prentice Hall: Upper Saddle River, NJ.
Jin Y. 2000. Fuzzy modeling of high-dimensional systems. *IEEE Transactions on Fuzzy Systems* **8**: 212–221.
Johansen TA, Babuska R. 2002. On multi-objective identification of Takagi–Sugeno fuzzy model parameters. In *Preprints 15th IFAC Word Congress*, Barcelona, Spain.
Mamdani EH. 1977. Application of fuzzy logic to approximate reasoning using linguistic synthesis. *IEEE Transactions on Computers* **26**: 1182–1191.
Mamdani EH, Assilian S. 1975. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man–Machine Studies* **7**: 1–13.
Nayak PC, Sudheer KP, Rangan DM, Ramasastri KS. 2004. A neuro-fuzzy computing technique for modeling hydrological time series. *Journal of Hydrology* **291**: 52–66.
Nayak PC, Sudheer KP, Rangan DM, Ramasastri KS. 2005a. Short term flood forecasting with a neurofuzzy model. *Water Resources Research* **41**: W04004.
Nayak PC, Sudheer KP, Ramasastri KS. 2005b. Fuzzy computing based rainfall–runoff model for real time flood forecasting. *Hydrological Processes* **19**: 955–968.
Ozelkan EC, Duckstein L. 2001. Fuzzy conceptual rainfall–runoff models. *Journal of Hydrology* **253**: 41–68.
Roubos JA, Setnes M. 2000. Compact fuzzy models through complexity reduction and evolutionary optimization. In *Proceedings of IEEE International Conference on Fuzzy Systems*, San Antonio, USA; 762–767.
See L, Openshaw S. 2000. Applying soft computing approaches to river level forecasting. *Hydrological Sciences Journal* **44**: 763–779.
Sen Z, Altunkaynak A. 2004. Fuzzy awakening in rainfall–runoff modelling. *Nordic Hydrology* **35**: 31–43.
Shamseldin AY. 1997. Application of a neural network technique to rainfall–runoff modelling. *Journal of Hydrology* **199**: 272–294.
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**: 671–686.
Takagi T, Sugeno M. 1985. Fuzzy identification of systems and its application to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics* **15**: 116–132.
Tsukamoto Y. 1979. An approach to fuzzy reasoning method. In *Advances in Fuzzy Set Theory and Application*, Gupta MM, Ragade RK, Yager RR (eds). North-Holland: Amsterdam; 137–149.
Vernieuwe H, Georgieva O, Baets B, Pauwels VRN, Verhoest NEC, Troch FP. 2005. Comparison of data-driven Takagi–Sugeno models of rainfall–discharge dynamics. *Journal of Hydrology* **302**: 173–186.
Xiong LH, O'Connor KM. 2002. Comparison of four updating models for real-time river flow forecasting. *Hydrological Sciences Journal* **47**: 621–639.
Xiong LH, Shamseldin AY, O'Connor KM. 2001. A nonlinear combination of the forecasts of rainfall–runoff models by the first order Takagi–Sugeno fuzzy system. *Journal of Hydrology* **245**: 196–217.
Yager R, Filev D. 1994. Generation of fuzzy rules by mountain clustering. *Journal of Intelligent & Fuzzy Systems* **2**: 209–219.

APPENDIX A: GUSTAFSON–KESSEL CLUSTERING ALGORITHM

For a given data $\mathbf{Z}$, an initially overestimated number of clusters $1 < m < n$, the fuzziness parameter $\phi > 1$, the rule contribution threshold $\rho\%$, and the termination tolerance $\delta > 0$, initialize $U^{(o)}$ randomly.

Repeat for $l = 1, 2 \ldots$

Step 1. Compute cluster prototypes:

$$u_i^{(l)} = \frac{\sum_{k=1}^{n} (u_{ki}^{(l-1)})^{\phi} z_k}{\sum_{k=1}^{N} (u_{ki}^{(l-1)})^{\phi}} \quad 1 \le i \le m$$

Step 2. Compute the cluster covariance matrices:

$$\boldsymbol{F}_i = \frac{\sum_{k=1}^{n} (u_{ki}^{(l-1)})^{\phi} (\boldsymbol{z}_k - u_i^{(l)})(\boldsymbol{z}_k - u_i^{l})^{\mathrm{T}}}{\sum_{k=1}^{n} (u_{ki}^{(l-1)})^{\phi}}$$

$$1 \le i \le m$$

Step 3. Compute distance to cluster prototype:

$$d_{ki}^2 = (z_k - u_i^{(l)})^{\mathrm{T}} D_i (z_k - u_i^{(l)})$$

$$1 \le i \le m, \quad 1 \le k \le n$$

where the $D_i = \det[(\boldsymbol{F}_i)^{1/(n+1)} \boldsymbol{F}_i^{-1}]$

Step 4. Update the partition matrix. For $1 \le i \le m$, $1 \le k \le n$:

if $d_{ki} > 0$

$$u_{ki}^{(l)} = \frac{1}{\sum_{j=1}^{m} (d_{ki}/d_{kj})^{2/(\phi-1)}}$$

if $d_{ki} = 0$

$$u_{ki}^{(l)} = 1$$

until $\|U^{(l)} - U^{(l-1)}\| < \delta$.

APPENDIX B: EQUATIONS FOR PERFORMANCE EVALUATIONS INDICES

Global evaluation measures

1. Correlation CORR $= \dfrac{\sum_{t=1}^{n} (y_t^{\mathrm{o}} - \overline{y^{\mathrm{o}}})(y_t^{\mathrm{c}} - \overline{y^{\mathrm{c}}})}{\sqrt{\sum_{t=1}^{n} (y_t^{\mathrm{o}} - \overline{y^{\mathrm{o}}})^2} \sqrt{\sum_{t=1}^{n} (y_t^{\mathrm{c}} - \overline{y^{\mathrm{c}}})^2}}$

2. Coefficient of efficiency $R^2 = 1 - \dfrac{\sum_{t=1}^{n} (y_t^{\mathrm{o}} - y_t^{\mathrm{c}})^2}{\sum_{t=1}^{n} (y_t^{\mathrm{o}} - \overline{y^{\mathrm{o}}})^2}$

3. Root-mean-square error RMSE $= \sqrt{\dfrac{\sum_{t=1}^{n} (y_t^{\mathrm{o}} - y_t^{\mathrm{c}})^2}{n}}$

4. Sum squared error SSE $= \sum_{t=1}^{n} (y_t^{\mathrm{o}} - y_t^{\mathrm{c}})^2$

DOI: 10.1002/hyp

5. Standard error estimates $\mathrm{SEE} = \sqrt{\dfrac{\sum_{t=1}^{n}(y_t^{\mathrm{o}} - y_t^{\mathrm{c}})^2}{v}}$

6. Mean bias error $\mathrm{MBE} = \dfrac{1}{n}\sum_{t=1}^{n}(y_t^{\mathrm{c}} - y_t^{\mathrm{o}})$

7. Noise to signal ratio $\mathrm{NS} = \dfrac{\mathrm{SEE}}{\sigma_{y_n}}$

8. Mean absolute error $\mathrm{MAE} = \dfrac{\sum_{t=1}^{n}|(y_t^{\mathrm{o}} - y_t^{\mathrm{c}})|}{n}$

9. Relative error $R_{\mathrm{e}} = \left|\dfrac{y_t{}^{\mathrm{o}} - y_t{}^{\mathrm{c}}}{y_t{}^{c}}\right| \times 100$

Note: y_i^{o} and y_i^{c} respectively are the observed and computed flow values at time t, $\overline{y^{\mathrm{o}}}$ and $\overline{y^{\mathrm{c}}}$ are the mean of the observed and computed flow values corresponding to n patterns, ν is the number of degrees of freedom, and σ_y is the standard deviation of the observed flow. Normalized values of these statistics are obtained by dividing the value by the observed mean.

Event-specific evaluation measures

They PFC and LFC are computed as follows:

$$\mathrm{PFC} = \frac{\left\{\sum_{t=1}^{T_{\mathrm{p}}}[(Q_t^{\mathrm{o}} - Q_t^{\mathrm{c}})^2(Q_t^{\mathrm{o}})^2]\right\}^{1/4}}{\left[\sum_{t=1}^{T_{\mathrm{p}}}(Q_t^{\mathrm{o}})^2\right]^{1/2}}$$

$$\mathrm{LFC} = \frac{\left\{\sum_{t=1}^{T_{\mathrm{l}}}[(Q_t^{\mathrm{o}} - Q_t^{\mathrm{c}})^2(Q_t^{\mathrm{o}})^2]\right\}^{1/4}}{\left[\sum_{t=1}^{T_{\mathrm{p}}}(Q_t^{\mathrm{o}})^2\right]^{1/2}}$$

in which T_{p} is the number of peak flows greater than the one-third of the mean peak flow observed; T_{l} is the number of low flows lower than the one-third of the mean low flow observed; $Q_t{}^{\mathrm{o}}$ and $Q_t{}^{\mathrm{c}}$ are respectively the observed and computed flows for the time period t.

DOI: 10.1002/hyp