Analysis of drought and storage for Mula project using ANN and stochastic generation models

Taymoor A. Awchi and D. K. Srivastava

ABSTRACT

A hybrid model for streamflow generation is presented to explore the possibilities of using the multilayer feedforward artificial neural networks (ANNs) as generators of future scenarios, with emphasis on the ability to reproduce the statistics of flows related to drought and storage. The artificial neural network model has two components: deterministic and random. The second part of the model incorporates the uncertainty associated with the hydrological processes. The model is applied to the monthly inflows of Mula irrigation project in Maharashtra, India. A comparison of drought and storage among other statistics was made between the performance of the ANN-based model results and the results of the Thomas–Fiering models. The results show that ANN is a promising alternative modelling approach for flow simulation purposes, with interesting potential in the context of water resources systems management and optimization.

Key words | drought, inflow generation, neural networks, Thomas–Fiering model, stochastic, storage

Taymoor A. Awchi (corresponding author)
Water Resources Engineering Department,
College of Engineering,
University of Mosul,
Mosul,
Iraq
E-mail: *awchi2002@yahoo.co.in*

D. K. Srivastava
Hydrology Department,
Indian Institute of Technology-Roorkee,
247667,
India

NOMENCLATURE

a	Constant
b_τ	Regression coefficient for estimating volume of discharge in the $(\tau+1)$th month from the τth month, which equals r_τ. $S_{\tau+1}/S_\tau$
g_τ	Skewness coefficient for the set $Q_{1,\tau}$, $Q_{2,\tau},\ldots,Q_{N,\tau}$
r_τ	Correlation coefficient between flows in the τth and $(\tau+1)$th months
N	Number of years of record of the series
O_j	Actual output activation at output unit j
$Q_{\nu,\tau}$	Monthly observed inflow for month τ $(\tau=1,\ldots,12)$ and year ν $(\nu=1,\ldots,N)$
$\bar{Q}_i$	Monthly average inflow for month τ
$Q_{\nu,\tau}$	Synthetic values produced by the model
$R_{\nu,\tau}$	Corresponding stochastic component
$R_{\nu,\tau+1}$	Random component
$S_{\tau+1}$	Standard deviation of the normalized and standardized discharge in the $(\tau+1)$th month
T_j	Target output activation at output unit j
$X_{\nu,\tau}$	Normalized inflows, for year ν and month τ
$\bar{X}_\tau, S_\tau$	Sample mean and standard deviation of normalized inflows for month τ, respectively
$Y^d_{\nu,\tau}$	Values produced by ANN scheme
$Y_{\nu,\tau}$, $Y_{\nu,\tau+1}$	Normalized and standardized volumes of generated discharges during the τth and $(\tau+1)$th month, respectively
$\bar{Y}_\tau, \bar{Y}_{\tau+1}$	Mean of normalized and standardized monthly discharges during the τth and $(\tau+1)$th months, respectively
$\varepsilon_{\nu,\tau}$	Random normal deviate with zero mean and unit variance

doi: 10.2166/nh.2009.012

INTRODUCTION

It is well known that water resources systems simulation using only historical records of precipitation, discharge or both, introduces severe restrictions. Loucks *et al.* (1981)

remark that a limited range of designs or alternative strategies result when applying only historical data for simulating the future behaviour of a water resources system, while better operation rules and designs are obtained when they are tested with a variety of generated hydrological scenarios. In general water resources investigations related to the use of water, such as for irrigation, hydropower and urban water supply, may require generation of streamflows at one or more sites for estimation of the required storage capacity of a system of reservoirs. The forecasting of variables, such as rainfall and runoff for operational studies of reservoir systems, is also important. The search for an optimum design of a water resource project therefore often involves finding a method or technique of generating long sequences of flows characteristics of the river in question. The generated sequence can then be used to analyze the performance of the water resources system design.

Several annual and monthly stochastic streamflow generation and forecasting models have been employed for planning of water resources systems. The most extended techniques include models of simple and multiple linear regressions, autoregressive (AR), autoregressive moving average (ARMA), ARMA with exogenous variable (ARMAX) and ARMA and ARMAX with periodic parameters. In all these cases, a linear relationship among the relevant hydrological variables is assumed. Nevertheless, the linear assumption does not always yield the best results and is sometimes found to be inadequate (Raman & Sunilkumar 1995).

Early studies (Brittan 1961; Thomas & Fiering 1962; Fiering 1967) have attempted to describe streamflow sequences by mathematical models. These models can reproduce special features of the historical record such as periodicities and account for the effects of serial correlation. Of these, the most important contribution was made by Thomas & Fiering (1962). They proposed that streamflows could be simulated by a simple linear relationship with previous flows. The Thomas–Fiering model, based on the assumptions that the correlation between months with a lag greater than one is negligible and that the serial correlation is linear, will generate an artificial record of any length. The Thomas–Fiering model has been applied successfully in many studies, e.g. flow data generation (Colston & Wiggert 1970), rainfall data generation (Gangyan *et al.* 2002).

An attractive feature of the Artificial Neural Networks (ANNs) is the ability to extract the relation between inputs and outputs of a process, without the physics being explicitly provided to them. They are able to provide a mapping from one multivariate space to another, given a set of data representing that mapping. Even if the data are noisy, ANNs have been known to identify the relationships between variables. These properties suggest that ANNs may be well-suited to the problems of estimation and prediction in hydrology.

The field of neural networks has a history of some five decades but has found solid application only in the most recent two decades; the field is still developing rapidly. The ANNs have been successfully used in hydrology-related areas, such as forecasting and generation fields e.g. in the fields of flood forecasting, rainfall-runoff prediction, flow prediction, rainfall prediction and multivariate streamflow generation. For a detailed review see ASCE-Task Committee on Application of Artificial Neural Networks in Hydrology (ASCE-TCAANNH 2000).

The present paper aims to explore the possibilities of using feedforward ANNs as generators of future scenarios, with emphasis on the ability to reproduce the statistics related to drought and storage analysis for the Mula irrigation project in Maharashtra, India. The model is applied to generate monthly streamflow series which, in turn, may be applied for real-time operation of the water resources system. The results of ANN are compared with that of the Thomas–Fiering models.

ARTIFICIAL NEURAL NETWORKS STRUCTURE

ANNs were designed to mimic the characteristics of biological neurons in the human brain and nervous system. The network 'learns' by adjusting the interconnections (called weights) between layers. When the network is adequately trained, it is able to generalize relevant output for a set of input data. The learning typically occurs by example through training, where the training algorithm iteratively adjusts the connection weights (synapses).

The architecture of a feedforward neural network (Figure 1) refers to its framework as well as its interconnection scheme. The framework is often specified by the

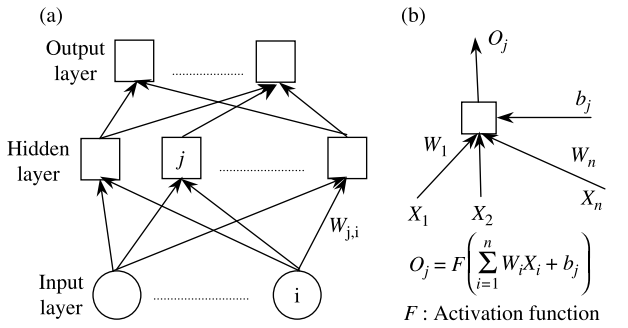

Figure 1 | Schematic diagram for the (a) feedforward neural network structure and (b) an artificial neuron.

number of layers and the number of nodes per layer. The types of layers include the following. The *input* layer contains nodes called *input units*, which encode the data of variables presented to the network for processing. For example, each input unit may be designated by an attribute value possessed by the instance. The *hidden* layer contains nodes called *hidden units*, which are not directly observable. They provide nonlinearities for the network. The *output* layer contains *output units*, which encode possible concepts (or values) to be assigned to the instance under consideration. For example, each output unit represents a class of objects. Input units do not process information, they simply distribute information to other units. The number of hidden layer neurons required is much more difficult to determine since no general methodology is available for its determination. The architecture of the network is therefore finalized after a trial-and-error procedure (Hsu *et al.* 1995).

Initialization of connection weights and threshold values is an important consideration. The closer the initial guess is to the optimum weight space, the faster the training process. However, there is no method of making a good initial guess of the weights, and they are initialized in a random fashion. Small random weights are usually suggested. They are adjustable during network training, but some can be fixed deliberately. When training is completed, all weights should be fixed (Jain & Singh 2003).

The feedforward neural network often trained with the backpropagation training algorithm is a kind of gradient descent technique with backward error (gradient) propagation and is currently the most common approach to training feedforward ANNs. Back propagation is a systematic method for training (calibrating) multilayer neural networks. It uses a set of pairs of input and output values (called patterns). An input pattern is fed into the network to produce an output, which is then compared with the actual output pattern. If there is no difference between the network output and the actual output, then no learning is needed. Otherwise, the weights (which express the contribution of the input nodes to the hidden nodes, and the hidden nodes to the output) are changed backward from the output layer through the hidden layer(s) to the input layer. Since the training makes use of the actual output, the back propagation method is referred to as a *supervised training* method. The typical performance function used for training feedforward neural networks is the mean sum of squares of the network errors:

$$mse = \frac{1}{N}\sum_{j=1}^{N}(T_j - O_j)^2 \tag{1}$$

where T_j is the target output activation and O_j is the actual output activation at output unit j. More details can be found in Haykin (1994) and Fu (1994).

THE THOMAS–FIERING (T–F) MODEL

For the assessment of the ANN model results, the Thomas–Fiering model (Thomas & Fiering 1962) was employed. The method implicitly allows for the non-stationarity observed in monthly discharge data. In its simplest form, the method consists of the use of twelve linear regression equations. The model may be written as:

$$Y_{\nu,\tau+1} = \bar{Y}_{\tau+1} + b_\tau(Y_{\nu,\tau} - \bar{Y}_\tau) + R_{\nu,\tau+1} \tag{2}$$

and

$$R_{\nu,\tau+1} = \varepsilon_{\nu,\tau} S_{\tau+1}\sqrt{1 - r_\tau^2} \tag{3}$$

where $Y_{\nu,\tau}$ and $Y_{\nu,\tau+1}$ are normalized and standardized volumes of generated discharges during the τth and $(\tau+1)$th month, respectively; $\bar{Y}_\tau$ and $\bar{Y}_{\tau+1}$ are mean of normalized and standardized monthly discharges during the τth and $(\tau+1)$th months, respectively; b_τ is the regression coefficient for estimating volume of discharge in the

$(\tau+1)$th month from the τth month, which equals r_τ. $S_{\tau+1}/S_\tau$; $R_{\nu,\tau+1}$ is a random component; $\varepsilon_{\nu,\tau}$ is a random normal deviate with zero mean and unit variance; $S_{\tau+1}$ is the standard deviation of the normalized and standardized discharge in the $(\tau+1)$th month; and r_τ is the correlation coefficient between flows in the τth and $(\tau+1)$th months.

Where the Thomas–Fiering model is fitted to monthly streamflows, values in the generated sequence are sometimes negative. It has been recommended that they should be retained and used to derive the subsequent values in the sequence. Once the generated sequence is complete, however, all negative values in the generated sequence are replaced by zero (Clarke 1973).

It is sometimes necessary to generate synthetic discharge sequences for stream having no discharge during dry season. The Thomas–Fiering model may be generalized by the inclusion of more terms of higher lag on the left-hand sides of the equations defining the model, so that they become multiple regression equations. Clarke (1973) described a model to deal with such a case.

DATA PROCESSING

A series of observed monthly inflows to the Mula irrigation reservoir for the period of June 1972–May 1990 (i.e. 18 years) is employed in this study (a water year starts from the month of June and is referred to as month 1 in this study). A preliminary exploration of these observed data showed that the skewness coefficients are biased. A transformation to reduce the skewness to zero is therefore needed (Salas *et al.* 1980). The skewness of the observed data is reduced using log-transformation which can be written as:

$$X_{\nu,\tau} = \log(Q_{\nu,\tau} + c_\tau \bar{Q}_\tau) \tag{4}$$

$$c_\tau = \frac{a}{g_\tau^2} \tag{5}$$

where $Q_{\nu,\tau}$ is monthly observed inflow for month τ ($\tau = 1, \ldots, 12$) and year ν ($\nu = 1, \ldots, N$); N is number of years of record of the series; $\bar{Q}_\tau$ is monthly average inflow for month τ, a is a constant; g_τ is the skewness coefficient for the set $Q_{1,\tau}, Q_{2,\tau}, \ldots, Q_{N,\tau}$; and $X_{\nu,\tau}$ are the normalized inflows, for year ν and month τ.

Raman & Sunilkumar (1995) and Salas *et al.* (1985) used a log-transformation in order to reduce the skewness coefficient; the equation they used contained a constant value of c_τ (i.e. the same value for every month). Ochoa-Rivera *et al.* (2002) employed a variable value of c_τ (i.e. different values for each month) to achieve an optimal skewness reduction. For the Mula irrigation project, variable c_τ values are adopted. Preliminary trial and error data tests showed that using a value of 0.85 for a resulted in the best performance of the model. The data were then standardized on a monthly basis utilized for the data generation by the Thomas–Fiering model.

Before training the neural network, the inputs and targets should be scaled so that they always fall within a specified range. For the Mula irrigation project, the data were scaled in the range of $[-1, +1]$. This step helps the neural network training to be more efficient (Demuth & Beale 2001).

Streamflow data generation using neural network model with random component

The proposed scheme is a mixed deterministic stochastic model for hydrological synthetic data generation, in terms of monthly series of reservoir inflows using an ANN-based model. The model consists of two components: the same stochastic part of Thomas–Fiering model $R_{\tau+1}$ (Equation (3)) and the deterministic component $Y^d_{\nu,\tau}$ which is represented by the ANN architecture. The two components are gathered over the normalized and standardized series. Consequently, the neural network component needs to be de-scaled first. The final form of the model can be resumed as the sum of both the components, and is given by:

$$Q_{\nu,\tau} = f\left(Y^d_{\nu,\tau} + R_{\nu,\tau}\right) \tag{6}$$

where $Q_{\nu,\tau}$ is the synthetic values produced by the model; $Y^d_{\nu,\tau}$ are the values produced by the ANN; and $R_{\nu,\tau}$ is the corresponding stochastic component given by Equation (3). Function f represents the inverse of the pre-processing operations, that is,

$$X_{\nu,\tau} = \left(Y^d_{\nu,\tau} + R_{\nu,\tau}\right)S_\tau + \bar{X}_\tau \tag{7}$$

$$Q_{\nu,\tau} = 10^{X_{\nu,\tau}} - c_\tau \bar{Q}_\tau \tag{8}$$

where $\bar{X}_\tau$ and S_τ are the sample mean and standard deviation of normalized inflows for month τ, respectively.

Two neural network based models are prepared. The first model (ANN1) generates inflows for the present month utilizing the inflow of the past month, and the second model (ANN2) generates the inflow of the present month utilizing inflows of two previous months.

A three-layer feedforward architecture is adopted, and tan-sigmoid functions were used as an activation function for which output falls in the range of $[-1, +1]$. In addition, linear activation functions for the output layer are employed. The total data were divided into two sets; training set (contained two-thirds of the total patterns) and validation set (contained one-third of the total data patterns). For the time series under consideration, the number of input and output nodes of the network are set in such a way that the target values to be predicted are the immediate next month's inflow. The number of hidden layer nodes was decided by a trial and error procedure, starting with a small number of nodes then increasing this number. Each change of nodes number was followed by network training until no significant improvement in network performance was detected, then that number of nodes was fixed.

The Levenberg-Marquardt training algorithm (Demuth & Beale 2001) is adopted in the present study, which is one of the fastest convergence algorithms. It is highly recommended for the small and medium networks, which contain several hundreds of connections. For designing and training the neural network models, the Neural Network Toolbox, Version 4.0 is employed. This toolbox is one of the MATLAB Version 6.5 Release 13 software toolboxes. Fifty inflow series were generated, each one of a length of 90 years of monthly discharges.

Evaluation of the performance of models

Having estimated the model parameters, diagnostic checks are necessary in order to see whether the model has been mis-specified and to search for model improvements. Diagnostic checks may include model implementation as well as testing the robustness of the model. For instance, the model may be implemented according to its intended utilization, such as data augmentation, generation or forecasting, and examine how well the model performs. Robustness tests may be applied to determine if the model preserves properties not explicitly parameterized in the model, e.g. Hurst coefficient and drought characteristics (Salas *et al.* 1985). A technique to be used in comparing the statistical characteristics derived from the generated series and historical data is presented by Salas *et al.* (1980). This technique is adopted in the present study and is explained later.

Storage statistics

To calculate the reservoir storage capacity to be guaranteed during the period of 90 years, the well-known sequent peak method (Loucks *et al.* 1981) is employed. The storage capacities are determined for different release values (taken as the threshold release varying between 10% and 90% of the mean discharge). The statistics for both the annual and monthly series are computed.

Drought statistics

For each of the 50 generated inflow series, a given percentage of the mean discharge is taken as a threshold so that each group of consecutive values below it defines a single drought run with its length and the sum (total volume below threshold). Consider the time series $X_1, \ldots, X_N$ and a constant demand level y (crossing level) as shown in Figure 2. A negative run occurs when X_t is consecutively less than y during one or more time intervals. Negative runs are related to the drought characteristics. For instance, a negative run of length 4 and run magnitude equal to d are shown. In general, several runs result in a time series of given demand level and sample size. The run length

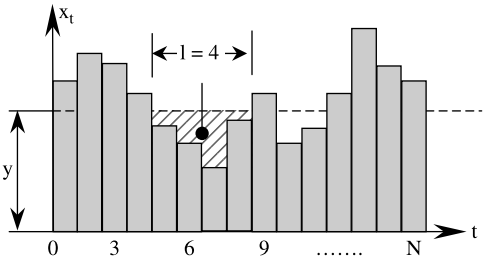

Figure 2 | Definitions of drought length and magnitudes.

characteristics may be used for comparison with the corresponding characteristics derived from the mathematical models fitted to the historical series (Salas *et al.* 1980). The basic statistics have been computed for all the series including annual and monthly series. In addition, maximum drought statistics also were calculated.

RESULTS AND DISCUSSION

Both the T–F and ANN-based models were employed to generate synthetic inflow data to the Mula reservoir. The historical inflow data for Mula reservoir contained 18 years of monthly discharges. In the present study 50 inflow series have been generated, each of 90 years length of monthly discharges. The comparison procedure included verification of the statistics (means, standard deviations, skewness coefficients and percent occurrence of zero inflows).

All four statistics mentioned above were computed for each of the 50 generated series, then the monthly means $\bar{u}_g$ and the standard deviations $s(u_g)$ for each of the above statistical characteristics computed are estimated to calculate the range limits for each statistical characteristics preservation which can be written as $[\bar{u}_g - k\ s(u_g);\ \bar{u}_g + ks(u_g)]$ where k can be equal to 1.0, 1.5 or 2.0 depending on the strictness of the test. Alternatively, k could be taken as the standard normal variate of a given significance level such as $k = 1.96$ for the 5% level (Salas *et al.* 1980). The monthly means, standard deviations, skewness coefficients and percent occurrence of zero inflows are computed for the historical data and are compared with that of the generated data.

Thomas–Fiering (T–F) models

The Thomas–Fiering lag-one model (T–F1) which utilizes one previous month's inflow ($Y_{\nu,\tau}$) to calculate the $(\tau + 1)$ month's inflow is applied for the Mula project inflow data to generate 50 inflow series. The results of basic statistical characteristics (monthly values of mean, standard deviation, coefficient of skewness and percent occurrence of zero inflows) are shown in Figure 3. The results show that the model preserves the statistical characteristics of the historical data since the historical data plot lies within the plots of maximum and minimum limits, calculated using the generated inflow data by the model.

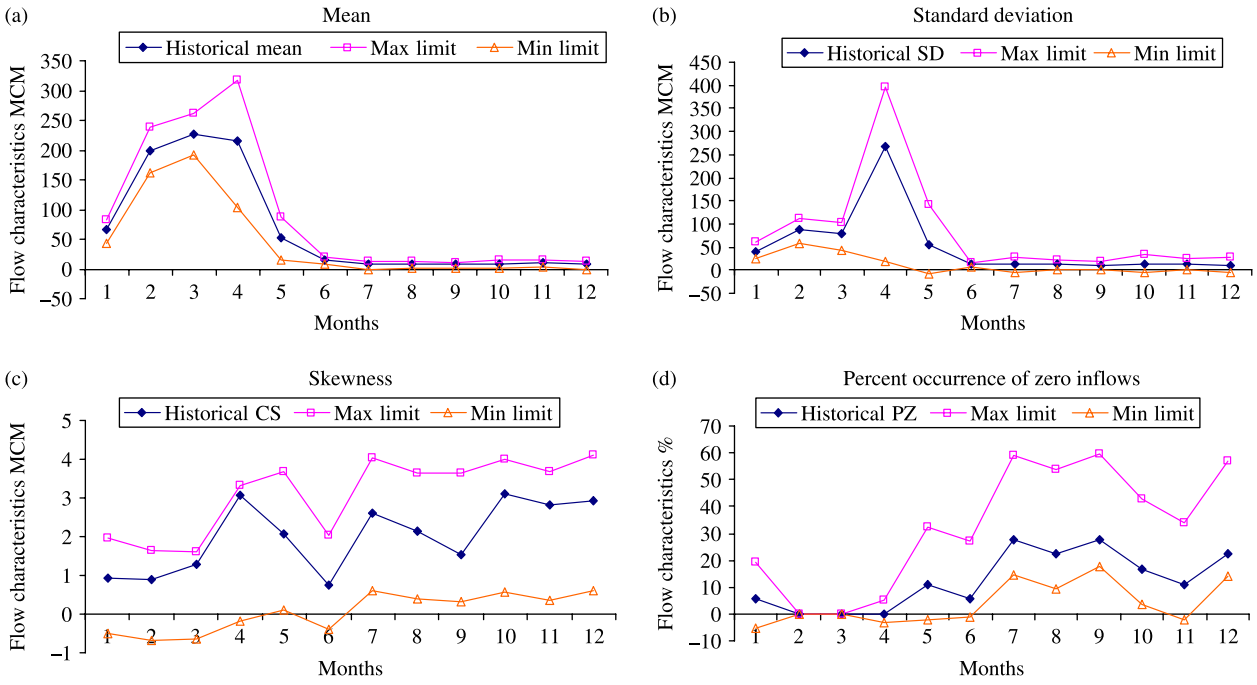

Figure 3 | Historical statistical characteristics and limits calculated from T–F1 model.

Table 1 | The MSE values calculated for the results of T–F1 and T–F2 models

Model	Mean	Standard deviation	Coefficient of skewness	Percent occurrence of zero inflows
T–F1	5.67	322.27	0.42	50.08
T–F2	6.94	366.54	0.36	90.30

Table 2 | MSE of the basic statistical characteristics calculated for ANN models with different inputs

Inputs	Hidden nodes	Means	Standard deviation	Coefficient of skewness	Percent occurrence of zero inflows
$Z_{\nu,\tau-1}$	3	3.76	180.30	0.38	61.73*
$Z_{\nu,\tau-1}$, b_1	7	2.17	222.21	0.42	42.63
$Z_{\nu,\tau-1}$, r	3	4.63	293.76	0.48	64.15
$Z_{\nu,\tau-1}$, $Z_{\nu,\tau-2}$	5	6.07	107.43	0.33	38.31†
$Z_{\nu,\tau-1}$, $Z_{\nu,\tau-2}$, mr	7	4.98	127.62	0.43	39.43

*Best results for ANN1 model.
†Best results for ANN2 model.

The lag-two (T–F2) model utilizes two previous months' inflows ($Y_{\nu,\tau}$ and $Y_{\nu,\tau-1}$) to generate the present month's inflow. The generated data using this model are verified by its statistical characteristics (Table 1) which proved that this model also preserves the statistical characteristics of the historical data.

The Mean Squared Error (MSE) values for the monthly basic statistical characteristics shown in Table 1 reveal that the MSE values in the case of T–F1 model are generally less than its values with the T–F2 model. This suggests that the T–F1 model suits the historical inflow data considered in the present study better than the T–F2

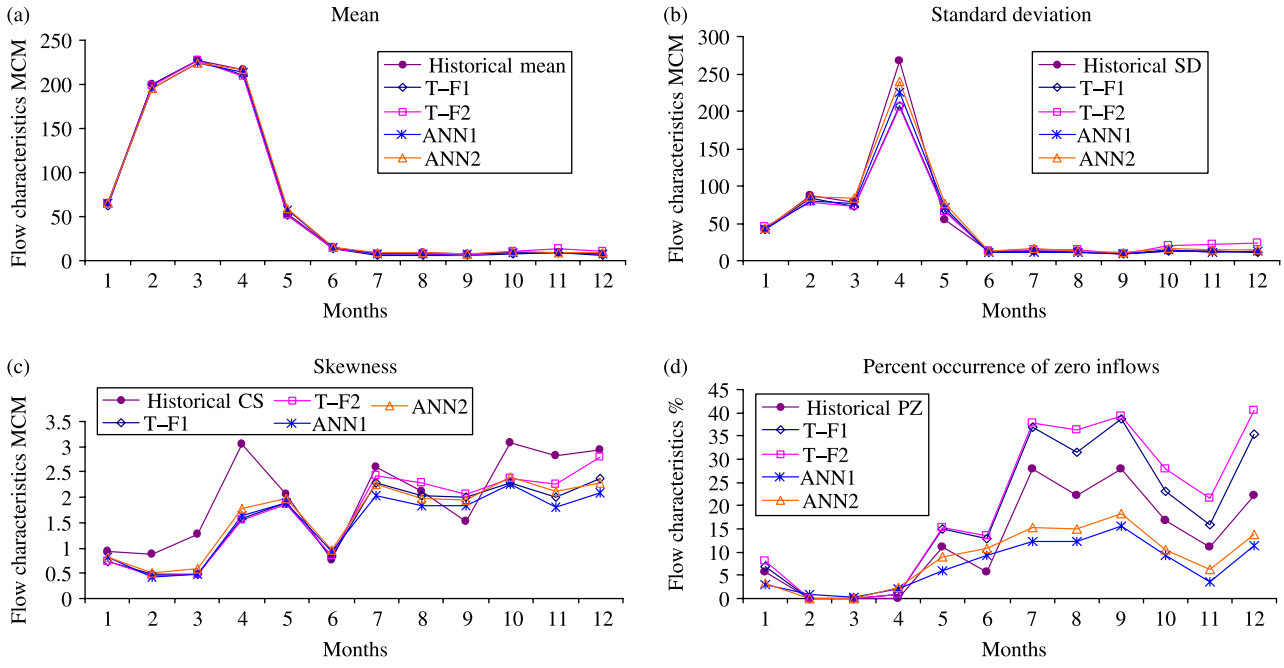

Figure 4 | Historical statistical characteristics and limits calculated from inflow data generated by T–F1, T–F2, ANN1 and ANN2 models.

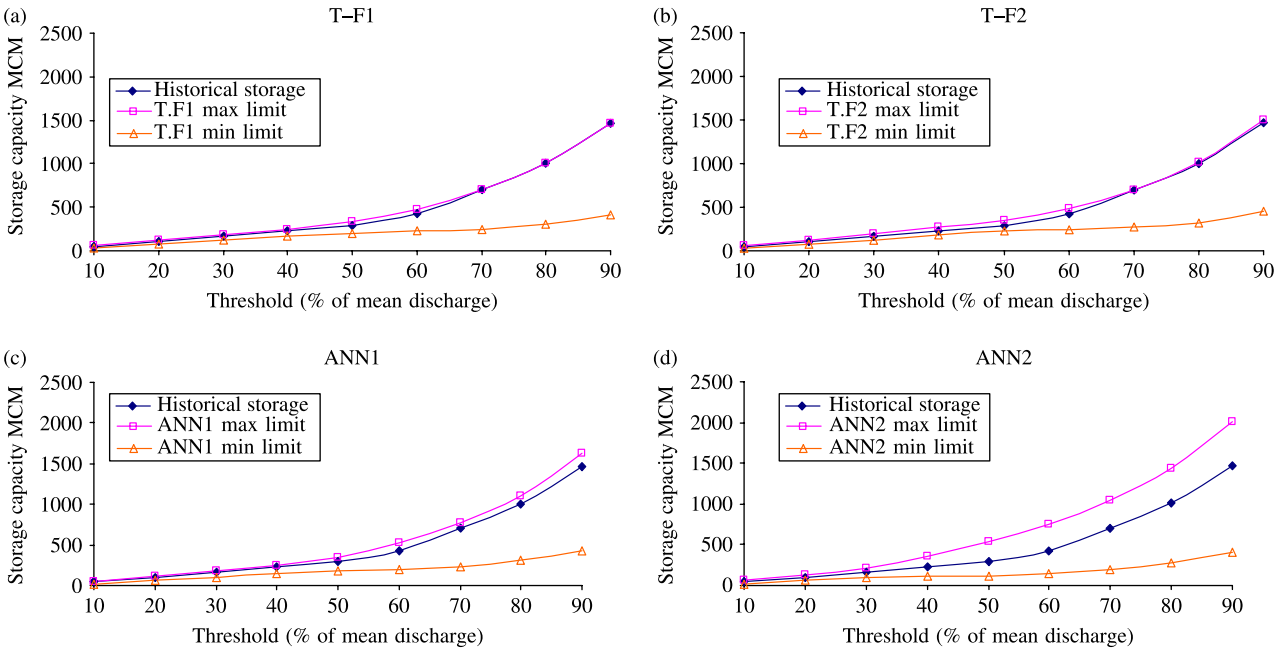

Figure 5 | Monthly historical storage capacity and their limits calculated from the models.

model. It also confirms that the increase of model lag does not necessarily enhance its performance (Salas *et al.* 1985).

ANN-based models

Two ANN-based models, ANN1 and ANN2, were prepared. The first model utilizes one previous month's inflow data ($Z_{\nu,\tau}$) and the second model utilizes two previous months' ($Z_{\nu,\tau}$ and $Z_{\nu,\tau-1}$) inflow data as an input. The output from the models is the generated present month's inflow value. Models with other inputs are also tested; results are presented later.

To find the best number of nodes in the hidden layer, the models are operated using a different number of nodes, starting with a small number then increased until no improvement in the model results is detected. That number of nodes is then fixed as no well-defined algorithm exists for determining the optimal number of hidden nodes (French *et al.* 1992). In the preliminary operating stage of the model, different numbers of training epochs (25, 50, 75, 100, 125, 150 and 200) are employed. Results demonstrated that the model gives best results with 75 epochs. This is confirmed by the study of French *et al.* (1992) who reported that increasing the number of training epochs alone, with no change in neural network structure, improves performance on the training data but does not necessarily improve performance on validation data. The number of epochs is therefore fixed for the subsequent training procedures.

In order to check the performance of the ANN models, maximum and minimum limits are calculated to compare with the historical data statistical characteristics. The results show that both the models (ANN1 and ANN2) preserve the basic statistical characteristics of the historical inflow data. Table 2 shows MSE values which are calculated from the results of ANN model with different inputs. In this table, MSE of basic statistical characteristics for the ANN1 model with inputs of

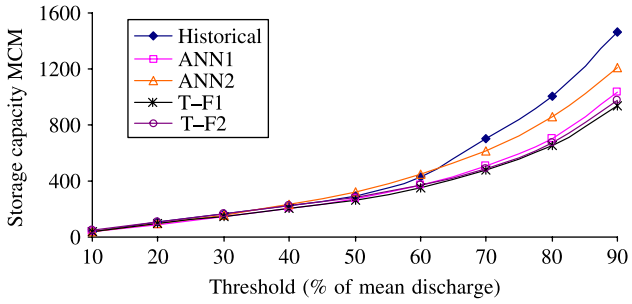

Figure 6 | Monthly storage capacity calculated using historical inflow, and data generated by ANN1, ANN2, T–F1 and T–F2 models.

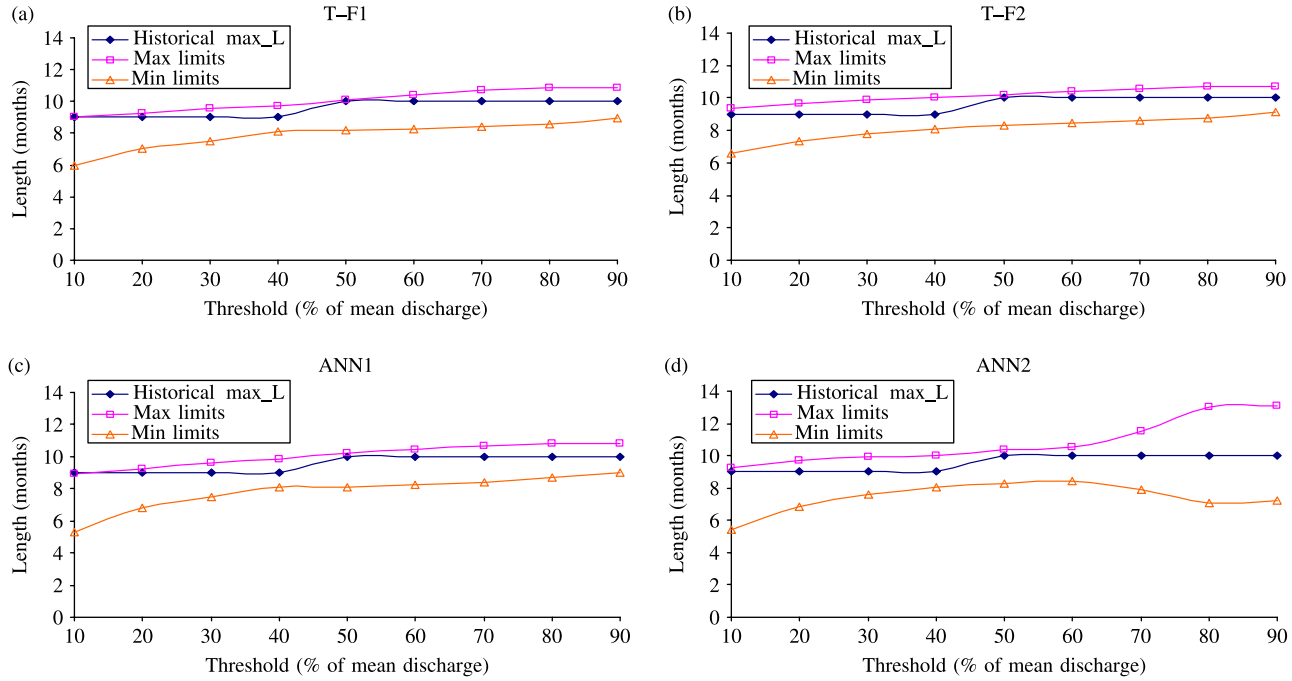

Figure 7 | Lengths of maximum monthly droughts of historical data and limits calculated from data generated by the models.

$(Z_{\nu,\tau-1}, Z_{\nu,\tau-1}, b_\tau$ and $Z_{\nu,\tau-1}, r_\tau)$ and the ANN2 model with inputs $(Z_{\nu,\tau-1}, Z_{\nu,\tau-2}$ and $Z_{\nu,\tau-1}, Z_{\nu,\tau-2}, mr_\tau)$ are shown.

Table 2 shows that the best MSE values are recorded when the ANN2 model is employed. The table reveals that it is not necessary for the results to be enhanced by increasing the number of nodes in the hidden layer. As in the case of the ANN1 model, when $(Z_{\nu,\tau-1})$, and b_1 are used as inputs, the model needed 7 hidden nodes to

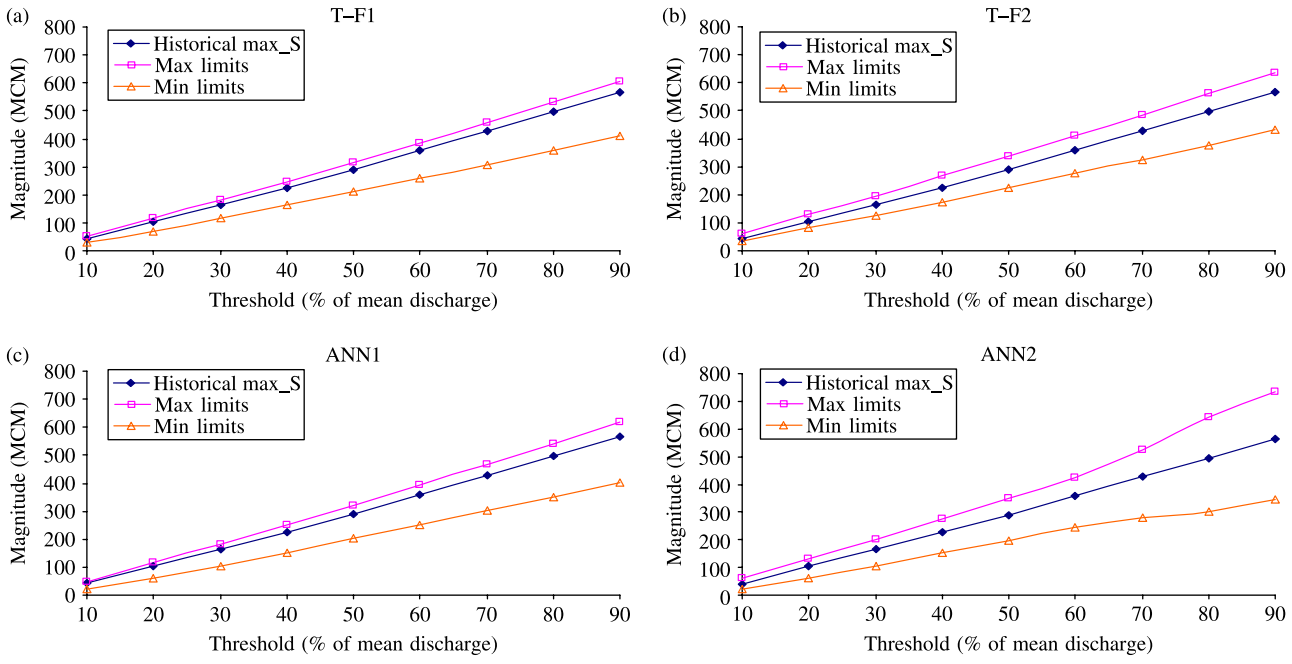

Figure 8 | Magnitudes of maximum monthly droughts calculated for historical data, and their limits calculated using data generated by the models.

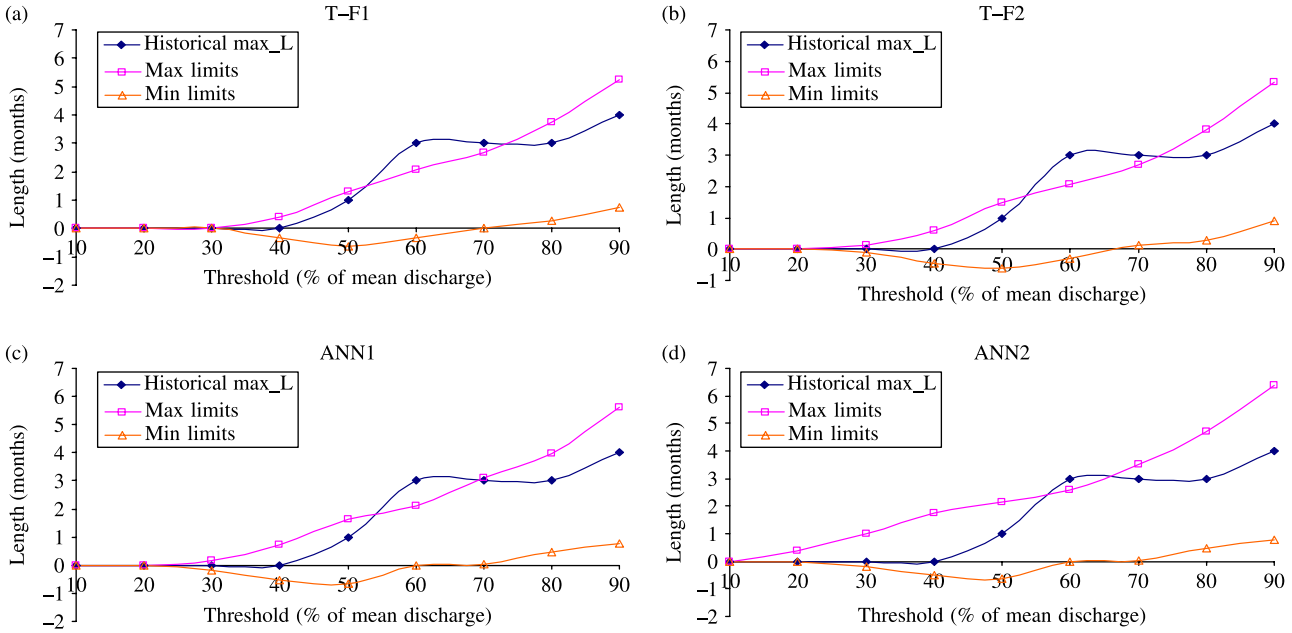

Figure 9 | Lengths of maximum annual droughts calculated for historical data and limits calculated from data generated by the models.

produce the best results. In comparison, when ($Z_{\nu,\tau-1}$) or ($Z_{\nu,\tau-1}$, r) are used as inputs to the model, the model needed only 3 hidden nodes to give the best results. On the other hand, and with ANN2 model when ($Z_{\nu,\tau-1}$ and $Z_{\nu,\tau-2}$) are used as inputs, best results are produced when 5 hidden nodes are used. This is attributed to the over-design of the model when a higher number of nodes are considered. The model structure should resemble the convolution of the relation between the input and output data used in the model. In the above-mentioned cases, it is

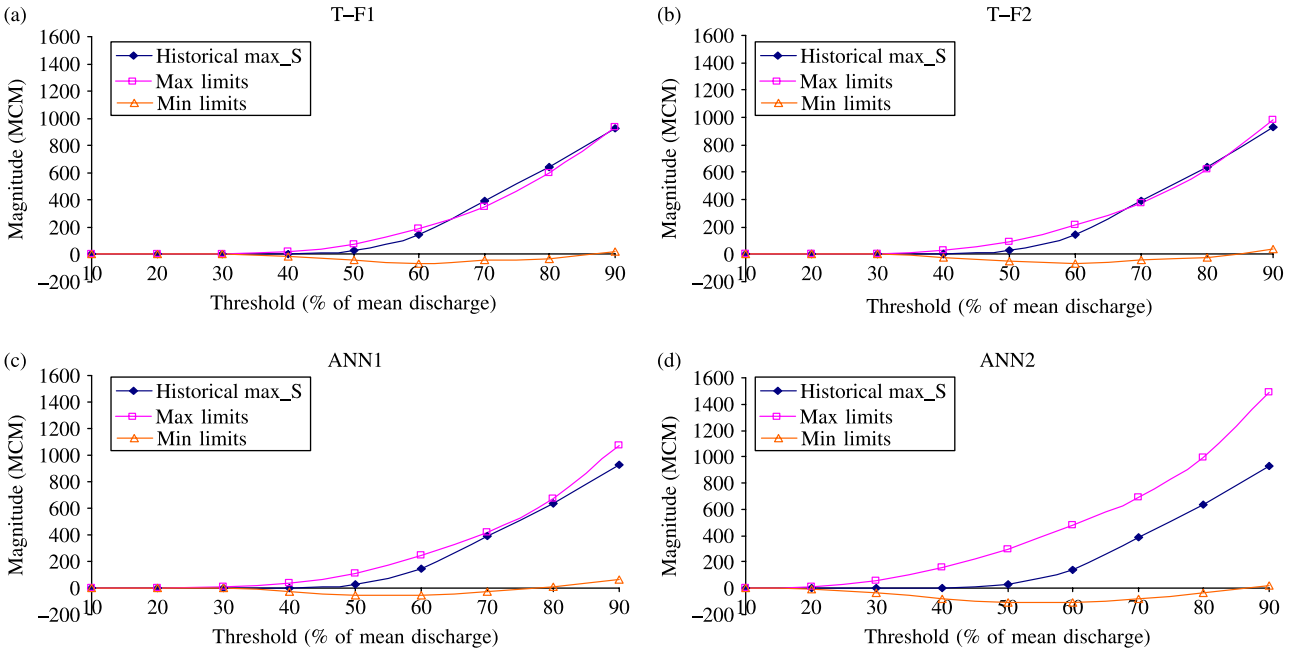

Figure 10 | Magnitudes of maximum annual droughts calculated for historical data and their limits calculated from data generated by the models.

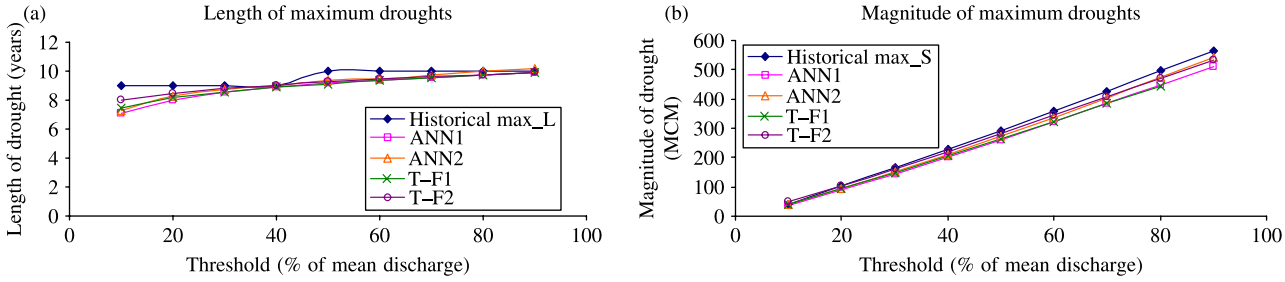

Figure 11 | Length and magnitude of maximum monthly droughts calculated for historical data and data generated from the models.

not desirable to over-design the neural network using more hidden nodes than necessary (as the neural network incorporates more the underlying characteristics of the process behaviour when a large number of hidden nodes are utilized) or to employ a large number of training epochs (which may not necessarily improve neural network performance on independent data).

The results of the neural networks based model (Table 2) and the Thomas–Fiering models (Table 1) for data generation have been compared; results show that the best MSE values are obtained when the ANN2 model is employed. This reveals that for the Mula project data, utilizing two previous months inflow values as input along with present month inflow value as output allows the model to interpret the relation between the input and output better than when using only one previous month's inflow value as the input. The former leads to a better distribution of connection weights between the nodes in the layers of the ANN model.

A comparison is carried out to check the performance of the models considered in the present study using the basic statistical characteristics. The results shown in Figure 4 reveal that the monthly means and standard deviation values obtained using generated data matches perfectly with the monthly means and standard deviations of the historical inflows. Moreover, the ANN2 model shows better results than the other models. Considering the coefficient of skewness values, it is shown that the results of all the models are harmonizing with the historical monthly coefficient of skewness values. For the case of monthly percent occurrence of zero inflows for various months, the results reveal that, generally, the T–F models present values higher than that of the historical data. This occurred by exchanging the generated negative inflows with zero values, which increases the percent of monthly zero inflows. In contrast, the ANN models show values less than the historical ones and this can be attributed to the quantity of historical data (only 18 values for each month, insufficient in this case).

For the T–F models, Figure 4 reveals that the matching in case of percent occurrence of zero inflows is quite perfect for the monsoon months and the distinctions are higher in non-monsoon months. This reveals that the generated data contains more zero inflows than the historical data. This is attributed to the fact that the T–F model generates negative values which should be corrected to zeros (Clarke 1973).

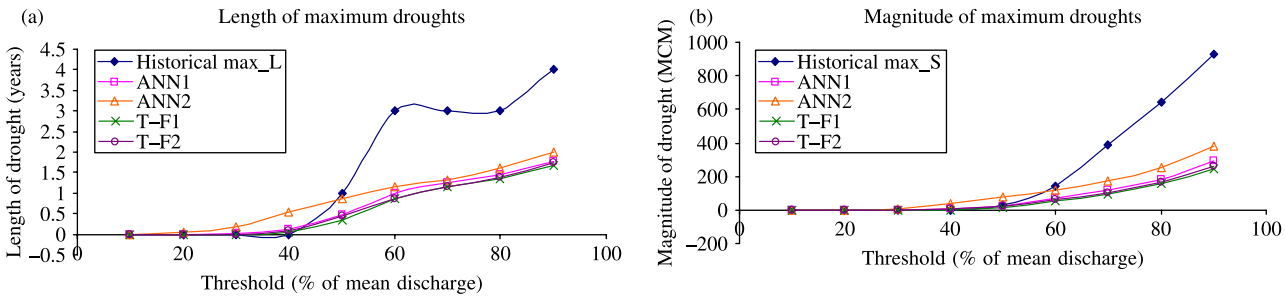

Figure 12 | Lengths and magnitude of maximum annual droughts calculated for historical data and data generated from the models.

This procedure therefore increases the number of zeros; hence, higher percents of zero inflows in the generated data are detected.

Storage capacity analysis

The results presented in Figure 5 show that in all four models, the historical data based storage capacities lie between the maximum and minimum limits calculated using the data generated by the models. This reveals that all the models preserve the historical storage capacities. In addition, as a comparison between these models, the ANN2 model shows the best results as the historical storage capacity line lies almost in the middle of the maximum and the minimum plots.

Comparing the storage capacity characteristics of the historical and the generated data with the models, Figure 6 reveals that the storage capacities are very well reproduced when the threshold is taken as 50% of the mean discharge or less. The differences between the historical and generated data based storage capacities then increase gradually. Also, the figure reveals that the best results are obtained when the ANN2 model is employed. Considering the T–F models, it is shown that there is no significant enhancement for the T–F2 model over the T–F1 model. This corroborates the fitness of the T–F1 model for the data considered in this study. Also, the results support the use of ANN2 since it shows perfect improved storage capacity compared to the ANN1 model.

Drought analysis

The drought analysis is applied for both the monthly and annual inflow values obtained by employing the ANN based and T–F models. In the present study, two main characteristics of the drought statistics are considered, i.e. the maximum length of drought period and maximum drought magnitude.

The monthly drought analysis showed preservation of historical data drought analysis for all the ANN based and T–F models (Figures 7 and 8) for maximum drought period length and maximum drought magnitude, respectively. On the other hand, annual drought analysis did not show full preservation of the features (Figure 9) of maximum drought period length. The analysis estimated the same maximum drought period length for demands of 60%, 70% and 80% of the mean discharge. For this, two points (i.e. demands of 60% and 70%) are over the maximum limit. This is attributed to the length of the historical data quantity which is only for 18 years and is not sufficient for such an analysis. The annual maximum drought magnitude preservation showed excepted values except in for demands of 70% and 80% of the mean discharge for the T–F models, but it is well preserved by the ANN2 model as shown in Figure 10.

A more comprehensive comparison of the results of the maximum drought length and the maximum drought magnitude (monthly and annual), using the historical and the data generated by the models, is shown in Figures 11 and 12. For the monthly drought analysis, the models perfectly reproduce the historical drought characteristics. For annual analysis, the models reproduced the historical data drought characteristics up to 50% threshold. Beyond this value, the difference between the historical and model-generated data drought characteristics increase with the threshold value, attributed to the same justification as above.

CONCLUSION

A hybrid model as an effective tool for generation of monthly streamflow series is presented and applied to the Mula irrigation project, Maharashtra, India. The model basically consists of a deterministic component defined in terms of a three-layer feedforward neural network, in addition to a stochastic component which includes a white noise module. Historical monthly inflows were used to train and test the ability of the model for inflow data generation. For comparison, the Thomas–Fiering model was applied. The Thomas–Fiering model is employed in both its common T–F1 form as well as the modified T–F2 form. The latter utilizes the two previous months' inflows as inputs to generate the present month's inflow.

The proposed hybrid model is operated and tested using two neural networks architectures, i.e. ANN1 and ANN2, which utilize one and two previous months' inflows as inputs, respectively. The present month's inflow is the output from the model in both cases. The results show that

the ANN2 model performs better compared to the ANN1 and Thomas–Fiering models. The results showed that the Thomas–Fiering and neural network models preserve the historical statistics and that the ANN2 model showed the best overall results. The storage capacity analysis showed that the best preserving of the historical data based storage capacities are obtained when the ANN2 model is used. The monthly analysis of drought length and magnitude showed good preservation for the historical data characteristics. On the other hand, annual drought analysis did not show results as good as the monthly analysis, which may be attributed to the short length of the historical data. The results revealed that the proposed hybrid model performs better when compared with the stochastic Thomas–Fiering Model. It can be concluded that the proposed hybrid model is a promising alternative to be considered for future applications, competing with other linear stochastic autoregressive models.

REFERENCES

ASCE-TCAANNH 2000 Artificial neural networks in hydrology. I: preliminary concepts. *J. Hydrol. Eng.* **5**(2), 115–123.

Brittan, M. R. 1961 *Probability Analysis Applied to the Development of Synthetic Hydrology for the Colorado River*. Boulder Colorado: Bureau of Economic Research, Report No. 4.

Clarke, R. T. 1973 *Mathematical Models in Hydrology*. Rome: FAO. Irrigation and Drainage Paper No. 19.

Colston, N. V. & Wiggert, J. M. 1970 A technique of generating a synthetic flow record to estimate the variability of dependable flows for a fixed reservoir capacity. *Water Resour. Res.* **6**(1), 310–315.

Demuth, H. & Beale, M. 2001 *Neural Network Toolbox for Use with MATLAB*. The MathWorks Inc. User Guide Version 4. Available from http://www.mathworks.com

Fiering, M. B. 1967 *Streamflow Synthesis*. Harvard University Press, Cambridge, MA.

French, M. N., Krajewski, W. F. & Cuykendall, R. R. 1992 Rainfall forecasting in space and time using a neural network. *J. Hydrol.* **137**, 1–31.

Fu, L. 1994 *Neural Networks in Computer Intelligence*. McGraw-Hill Inc, New York.

Gangyan, Z., Goel, N. K. & Bhatt, V. K. 2002 Stochastic modeling of the sediment load of the Upper Yangtze River (China). *Hydrol. Sci. J.* **47**(S), S93–S105.

Haykin, S. 1994 *Neural Networks: a Comprehensive Foundation*. MacMillan, New York.

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

Jain, S. K. & Singh, V. P. 2003 Applications of neural networks to water resources. In: (eds) Singh, V. P. & Yadava, R. N., Water resources system operation, *Proceedings of the International Conference on Water and Environment (WE-2003)*, Dec. 15–18, Bhopal, India.

Loucks, D. P., Stedinger, J. R. & Haith, D. A. 1981 *Water Resources Systems Planning and Analysis*. Prentice-Hall Inc, New Jersey.

Ochoa-Rivera, J., Garcia-Bartual, R. & Andreu, J. 2002 Multivariate synthetic streamflow generation using a hybrid model based on artificial neural networks. *Hydrol. Earth Syst. Sci.* **6**(4), 641–654.

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

Salas, J. D., Delleur, J. W., Yevjevich, V. & Lane, W. L. 1980 *Applied Modeling of Hydrologic Time Series*. Water Resources Publications, Littleton, CO, p. 484.

Salas, J. D., Tabios, G. Q. & Bartolini, P. 1985 Approaches to multivariate modeling of water resources time series. *Water Resour. Bull.* **21**(4), 683–708.

Thomas, H. A. & Fiering, M. B. 1962 Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. (ed, Maass, A., *et al.*), *Design of Water Resources Systems*. Harvard University Press, Cambridge, MA, pp. 459–493.

First received 30 January 2008; accepted in revised form 17 October 2008