Check for updates

Effect of data time interval on real-time flood forecasting

Renji Remesan, Azadeh Ahmadi, Muhammad Ali Shamim and Dawei Han

ABSTRACT

Rainfall–runoff is a complicated nonlinear process and many data mining tools have demonstrated their powerful potential in its modelling but still there are many unsolved problems. This paper addresses a mostly ignored area in hydrological modelling: data time interval for models. Modern data collection and telecommunication technologies can provide us with very high resolution data with extremely fine sampling intervals. We hypothesise that both too large and too small time intervals would be detrimental to a model's performance, which has been illustrated in the case study. It has been found that there is an optimal time interval which is different from the original data time interval (i.e. the measurement time interval). It has been found that the data time interval does have a major impact on the model's performance, which is more prominent for longer lead times than for shorter ones. This is highly relevant to flood forecasting since a flood modeller usually tries to stretch his/her model's lead time as far as possible. If the selection of data time interval is not considered, the model developed will not be performing at its full potential. The application of the Gamma Test and Information Entropy introduced in this paper may help the readers to speed up their data input selection process.

Key words | artificial neural networks, data time interval, flood forecasting, gamma test, information entropy

Renji Remesan (corresponding author)
Muhammad Ali Shamim
Dawei Han
Water and Environmental Management Research Centre,
Department of Civil Engineering,
University of Bristol,
Bristol BS8 1UP,
UK
E-mail: *Renji.Remesan@bristol.ac.uk*

Azadeh Ahmadi
Department of Civil Engineering,
Isfahan University of Technology,
Isfahan,
Iran

INTRODUCTION

Efficient flood forecasting is considered as a challenging field of operational hydrology as rainfall–runoff dynamics is highly nonlinear, time-dependent and spatially varying (Cluckie & Han 2000). Many models have been developed to replicate the rainfall–runoff process (HEC 1990; Duan *et al.* 1992; Michaud & Sorooshian 1994). Although conceptual models and physics-based models provide a deep insight into the physical processes, their calibration requires the collection of a great amount of information regarding the physical properties of the watershed under study and sophisticated mathematical tools for parameter identification (Duan *et al.* 1992; Chang *et al.* 2007). The advent of artificial intelligence techniques to hydrology brought a new dimension to flood modelling (Han *et al.* 2002, 2007*a*,*b*; Bray & Han 2005). Among several artificial intelligence methods artificial neural networks (ANN) hold a vital role and ASCE Task Committee Reports (2000*a*,*b*) have accepted ANN as an efficient forecasting and modelling tool. Over the last decade, the artificial neural network has gained great attention and has evolved as the main branch of artificial intelligence that is now a recognised tool for modelling the underlying complexities in many artificial and physical systems including floods (Abrahart & See 2007; Solomatine & Ostfeld 2008). Unlike traditional conceptual and physics-based models, artificial neural networks are able to mimic flow observations, without any mathematical descriptions of the relevant physical processes. A study by Jain *et al.* (2004) demonstrated that the distributed structure of the ANN was able to capture certain physical properties like infiltration, base flow, delayed and quick surface flow, etc. The success of hydrological forecasting systems depends on accurate

doi: 10.2166/hydro.2010.063

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

predictions in the longer forecast lead time. Multi-step-ahead prediction is a challenging task which attempts to make predictions several time steps into the future. Chang *et al.* (2004) developed a two-step-ahead recurrent neural network for streamflow forecasting. Later they explored three types of multi-step-ahead (MSA) neural networks, viz. multi-input multi-output (MIMO), multi-input single-output (MISO) and serial-propagated structure, for rainfall–runoff modelling using datasets from two watersheds in Taiwan (Chang *et al.* 2007).

However, even with an abundance of studies there are many uncertainties associated with ANN-based modelling, viz. random initialisation, proper model structure, best input combination, training data length, best data time interval for modelling, etc. So far, many of these questions have not been addressed adequately by the hydrological community (Han *et al.* 2007*a*,*b*). Two issues which deserve more attention are the input data selection and the optimal data time interval.

The selection of an appropriate subset of inputs from available input variables to model a system under investigation is a crucial step in model development, particularly for data-driven models like artificial neural networks. The correct choice of model inputs is very important for improving the modelling goal and computational efficiency. Improper input combination and training data length selection could lead to overfitting, which is considered as one of the serious weaknesses associated with data mining models like ANNs. There are two potential tools to deal with this problem for nonlinear systems, which are Mutual Information (MI) based on the Entropy Theory and the Gamma Test. Mutual Information has been used to measure the dependence between output and input variables. MI is capable of measuring dependences based on both linear and nonlinear relationships, making it well suited for use with complex nonlinear systems and can be used as a model input selection criteria (Sharma 2000; Fernando *et al.* 2005). The Gamma Test was proposed in recent years by Agalbjörn (Agalbjörn *et al.* 1997) and a formal proof for the Gamma Test can be found in Evans (2002) and Evans & Jones (2002). It is accomplished by the estimation of noise variance computed from the raw data using efficient, scalable algorithms. This novel technique, the Gamma Test, enables us to quickly evaluate and estimate the best

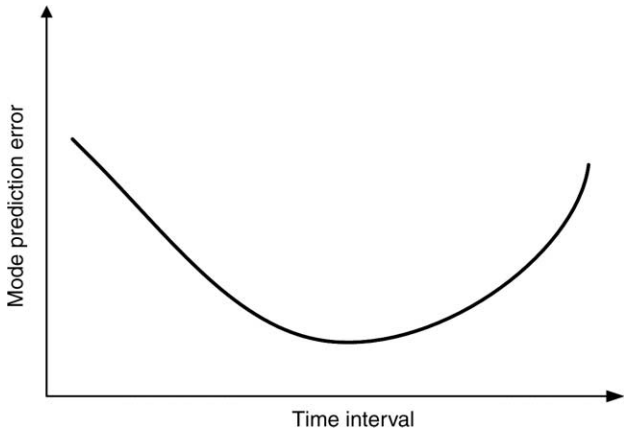

Figure 1 | Hypothetical relation between data time interval and model error.

mean squared error that can be achieved by a smooth model on unseen data for a given selection of inputs, prior to complex and time-consuming model construction. The abilities of GT have been demonstrated in case studies in water level and flow modelling (Durrant 2001; Remesan *et al.* 2009), daily solar radiation prediction (Remesan *et al.* 2008; Moghaddamnia *et al.* 2009) and evapotranspiration estimation (Ghafari *et al.* 2009; Piri *et al.* 2009). This technique can be used to find the best embedded dimensions and time lags for time series analysis. The credibility of the GT was evaluated by cross-correlation analysis and data splitting modelling.

The data time interval is a major factor affecting the forecast performance of neural network models. The performance of neural network models is highly time-dependent (Avci 2007). Very large and small data time intervals could have negative effects on modelling results. The hypothetical condition for the effect of data time interval on modelling is shown in Figure 1. In this study, an analysis of data time interval on real-time flood forecasting with different lead times was performed to see if the results could validate the hypothetical condition.

THE STUDY AREA

The ANN-based flood forecasting method was performed on the Brue catchment, located in Southwest England. The data for this study was obtained from the Hydrological Radar Experiment (HYREX) at the Brue catchment from

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

Figure 2 | Location map of the study area, Brue Catchments.

1994 to 1995 (hosted at the British Atmospheric Data Centre). The Brue catchment is located in Somerset (51.075°North and 2.58°West) with a drainage area of 135 sq km (Figure 2). It is a predominantly rural catchment of modest relief with spring-fed headwaters rising in the Mendip Hills and Salisbury Plain. The rain gauge network consists of 49 Cassella 0.2 mm tipping bucket type rain gauges. An automatic weather station (AWS) and an automatic soil water station (ASWS) located in the catchment recorded global solar radiation, net radiation and other physical parameters such as wind speed, wet and dry bulb temperatures, and atmospheric pressure at hourly intervals. Eight years (1993–2000) of daily rainfall–runoff data from the Brue catchment have been collected. Observations from 1994 to 1995 were used in the study to avoid missing data in other periods. For this study we used four different sets of data with different data collection time intervals (15 min, 30 min, 60 min and 120 min) and three lead times (2 h, 4 h and 6 h) for real-time flood forecasting.

Gamma test (GT)

This concept has been developed as the Gamma Test by Agalbjörn and his associates (Agalbjörn *et al.* 1997). A formal proof for the Gamma Test can be found in Evans (2002) and Evans & Jones (2002). This novel technique enables us to quickly evaluate and estimate the best mean-squared error that can be achieved by a smooth model on unseen data for a given selection of inputs, prior to model construction. The basic idea is quite distinct from those earlier attempts at nonlinear analysis. Suppose we have a set of data observations of the form

$$\{(x_i, y_i), 1 \le i \le M\} \tag{1}$$

where the inputs $x \in \Re^m$ are vectors confined to some closed bounded set $C \in \Re^m$ and, without loss of generality, the corresponding outputs $y \in \Re$ are scalars, where we presume that the vectors x contain predicatively useful factors influencing the output y. The only assumption made is that the underlying relationship of the system under investigation is of the following form:

$$y = f(x_1 \ldots x_m) + r \tag{2}$$

where f is a smooth function and r is a random variable that represents noise. Without loss of generality it can be assumed that the mean of the distribution of r is zero (since any constant bias can be subsumed into the unknown function f) and that the variance of the noise Var(r) is bounded. The domain of possible models is now restricted to the class of smooth functions which have bounded first partial derivatives. The Gamma statistic (Γ) is the estimate of that part of the variance of the output that cannot be accounted for by a smooth data model.

The Gamma Test is based on $N\,[i, k\,]$, which are the kth $(1 \le k \le p)$ nearest neighbours $x_{N[1,k\,]}(1 \le k \le p)$ for each vector $x_i(1 \le k \le p)$. Specifically; the Gamma Test is derived from the delta function of input vectors:

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M}\left|x_{N(1,k)} - x_i\right|^2 \quad (1 \le k \le p) \tag{3}$$

where $|\ldots|$ denotes Euclidean distance, and the corresponding Gamma function of output values is

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M}\left|y_{N(1,k)} - y_i\right|^2 \quad (1 \le k \le p) \tag{4}$$

where $y_{N(1,k)}$ is the corresponding y value for the kth nearest neighbour of x_i in (3). In order to compute Γ a least-squares fit regression line is constructed for the p points $(\delta_M(k), \gamma_M(k))$:

$$\gamma = A\delta + \Gamma \tag{5}$$

The intercept on the vertical ($\delta = 0$) axis is the Γ value, as can be shown by

$$\gamma_M(k) \to \mathrm{Var}(r) \text{ in probability as } \delta_M(k) \to 0 \tag{6}$$

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

The graphical output of this regression line (5) can provide very useful information. First, it is remarkable that the vertical intercept Γ of the y (or Gamma) axis offers an estimate of the best MSE achievable utilising a modelling technique for unknown smooth functions of continuous variables (Evans & Jones 2002). Second, the gradient A offers an indication of model complexity (a steeper gradient indicates a model of greater complexity).

Another term associated with the Gamma Test is V_{ratio}. A V_{ratio} close to zero indicates that there is a high degree of predictability of the given output y. We can also determine the reliability of the Gamma statistic by running a series of Gamma Tests for increasing M, to establish the size of dataset required to produce a stable asymptote. This is known as an M test. The M-test result would help us to avoid a wasteful attempt of fitting the model beyond the stage where the MSE on the training data is smaller than Var(r), which may lead to overfitting. The M test also helps us to decide how much data we require to build a model with a mean squared error which approximates the estimated noise variance.

ENTROPY THEORY

The concept of information is too broad to be captured completely by a single definition. In information theory, the Shannon entropy or information entropy is a measure of the uncertainty associated with a random variable. It quantifies the information contained in a message, usually in bits or bits/symbol. It is the minimum message length necessary to communicate information. The concept was introduced by Claude E. Shannon (Shannon 1948). Caselton & Husain (1980) introduced the entropy concept into a hydrometric network study. They computed the information transmission, based on the entropy concept, and selected stations with the maximum information transmission. Numerous studies in the hydrogeological literature have employed entropy in the context of model optimisation. Amorocho & Espildora (1973) used entropy to measure the information gained by a hydrologic model. Chapman (1986) studied the application of entropy in various cases involving the use of different assumed distribution functions, different types of flow data and also considered different units of entropy. Harmancioglu & Yevjevich (1987) used entropy to measure the information transmission among stations on the same river. Krstanovic & Singh (1992) investigated information transfer between selected drought or flood sequences, using marginal entropy, joint entropy and transinformation in long-term monthly rainfall series. Yang & Burn (1994) presented an entropy-based methodology for the design of data collection systems. Using marginal entropy, Maruyama & Kawachi (1998) investigated the characteristics of local rainfall in Japan. For any probability distribution, we can define an entropy quantity which has many properties that agree with the intuitive notion of what a measure of information should be. The entropy of a random variable is a measure of the uncertainty of a random variable; it is a measure of the amount of information required on average to describe a random variable. In the case of laboratory or field data, measurements are usually discrete, representing datasets that are limited in time and space. Rather than fitting an analytical function to the data, we can establish a bin specification to construct a probability distribution directly from the data. To calculate entropy for a continuous function, we use the discrete analogue of Equation (7). Let X be a discrete random variable with probability mass function $p(x)$. The entropy $H(X)$ of a discrete random variable X is defined as

$$H(X) = \sum p(x) \log (P(x)) \tag{7}$$

The reduction of the original uncertainty of X, i.e. $H(X)$, due to the knowledge of Y is

$$T(X, Y) = H(X) - H(X|Y) \tag{8}$$

This is called transinformation. It can be viewed as the information transferred by the knowledge of Y into the process to make X better defined and is therefore a goodness measure for the predictor.

ANN RAINFALL–RUNOFF MODEL

There are many ways to implement ANNs in rainfall–runoff modelling. In general ANN architecture is a multiple-layer perceptron (MLP), which can have many layers where a layer represents a set of parallel processing units (or nodes).

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

In this study we established a three-layer feedforward neural network (one input layer, one hidden layer and one output layer). This topology has proved its ability in modelling many real-world functional problems. The FFBP (Feedforward Backpropagation) is the most popular ANN training method in water resources literature. In this study, the FFBPs were trained using the Levenberg–Marquardt (LM) optimisation technique. This LM optimisation technique is more powerful than the conventional gradient descent techniques (El-Bakyr 2003; Cigizoglu & Kisi 2005).

In training, it is typical to choose a performance function, which has the form of a sum of squares. In this case, the gradient can be written as

$$g = 2J(x)^{\mathrm{T}}e(x) \tag{9}$$

where $\boldsymbol{J}$ (Jacobian) contains the first derivatives of the network error with respect to the weights and biases, and $\boldsymbol{e}$ is a vector of network errors. The Jacobian matrix can be computed using a standard backpropagation technique. The basic implementation of this algorithm can be written as

$$x_{k+1} = x_k - \left[J^{\mathrm{T}}(x_k)J(x_k)\right]^{-1}J^{\mathrm{T}}(x_k)e(x_k) \tag{10}$$

where $\boldsymbol{k}$ is an integer. This is the Gauss–Newton method of approximating the Hessian matrix. If the matrix $J^{\mathrm{T}}(x_k)J(x_k)$ is not invertible, the following formula could be used to solve that issue:

$$x_{k+1} = x_k - \left[J^{\mathrm{T}}(x_k)J(x_k) + \mu_k I\right]^{-1}J^{\mathrm{T}}(x_k)e(x_k) \tag{11}$$

where μ_k is an adaptive value. When the scalar μ_k is zero, this is just Newton's method, using the approximate Hessian matrix. When μ_k is large, this becomes gradient descent with a small step size. Newton's method is faster and more accurate near an error minimum, so the aim is to shift toward Newton's method as quickly as possible.

In this study, a multi-input single-output (MISO) neural network architecture was adopted since it has been popularly used as a neural network architecture for multi-step-ahead forecasting (Chang *et al.* 2007). We have constructed three independent networks to forecast $\hat{y}(t+2)$, $\hat{y}(t+4)$ and $\hat{y}(t+6)$, respectively. Even though we require $\boldsymbol{n}$ networks for $\boldsymbol{n}$-step-ahead forecasting, MISO networks were considered better than a multi-input multi-output (MIMO) scheme with reduced training time and increased accuracy (Han *et al.* 2007*a*,*b*). For a MISO network, the number of parameters and weights between hidden and output layers is much less than that of MIMO; thus the complexity is less. The selection of hidden neurons is another tricky part in ANN modelling as it relates to the complexity of the system being modelled and there are several ways of doing it, such as the geometric average between input and output vector dimensions (Maren *et al.* 1990), the same as the number of inputs (Mechaqrane & Zouak 2004), twice the input layer dimension plus one (Hecht-Nielsen 1990), etc. In this study, the Hecht-Nielsen (1990) approach has been adopted according to our past experience.

Evaluation of the model's predictive abilities for different data time intervals was compared with statistical terms such as the root-mean-square error (RMSE) between the observed and forecasted values, the coefficient of efficiency (Nash & Sutcliffe 1970), mean bias error (MBE), coefficient of correlation (R), slope and the mean absolute error (MAE). For a reliable comparison, the same data points were used for all the time intervals during validation.

RESULTS AND DISCUSSIONS

There were two aspects in this study to address the model data input selection and the impact of data time interval. The modelling results are explained as follows.

Input data selection for ANN model

The three-layer feedforward neural network with one hidden layer is used in this study and the number of hidden neurons was twice the number of input vectors plus one. To construct an ANN structure for the catchment, its streamflow data and the rainfall information were selected as the input vector and the best combination of antecedent values of this information were selected using the Gamma Test. The Gamma Test analysis can provide vital information which would help the modelling development. There are 2^{n-1} possible combinations of inputs; from which the best one can be determined by observing the Gamma statistic values. In this study, the Gamma Test analysis was

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

Table 1 | The gamma test results on the rainfall–runoff data of the Brue catchment

Parameters	Data time interval 15 min	30 min	60 min	120 min
Input combination	[4, 4]*	[2, 2]*	[3, 3]*	[4, 3]*
	$Q(t)\ldots Q(t-3)$ and $P(t)\ldots P(t-3)$	$Q(t)$, $Q(t-1)$, $P(t)$ and $P(t-1)$	$Q(t)\ldots Q(t-2)$ and $P(t)\ldots P(t-2)$	$Q(t)\ldots Q(t-3)$ and $P(t)\ldots P(t-2)$
Gamma (Γ)	0.00447	0.00061	0.00813	0.00843
Gradient (A)	0.05873	0.18535	0.08403	0.08168
Standard error	0.00022	0.00054	0.00089	0.00173
V_{ratio}	0.01789	0.00244	0.03253	0.03375
Near neighbours	10	10	10	10
M	65,524	35,031	17,515	8,756
Mask	11111111	11001100	11101110	11101111

*[a, b] where, a denotes the number of antecedent runoff datasets used for the M test and b denotes the corresponding number of antecedent rainfall data.

performed for different datasets with the data collection intervals of. 15 min, 30 min, 60 min and 120 min. Different combinations of inputs evaluated in this study are shown in Table 1. The least Gamma value was observed when we used the input data combination equivalent to [4, 4] for the 15 min dataset (high frequency data), i.e. four antecedent runoff and four antecedent rainfall as model input. The network input include information $Q(t)$, $Q(t-1)$, $Q(t-2)$, $Q(t-3)$, $P(t)$, $P(t-1)$, $P(t-2)$ and $P(t-3)$ for multi-step-ahead forecasting when we used 15 min data. The Gamma statistic (Γ) and Standard Error (SE) variation with unique data points of 15 min data, obtained from the M test analysis are shown in Figure 3(a). The test produced an asymptotic convergence of the Gamma statistic to

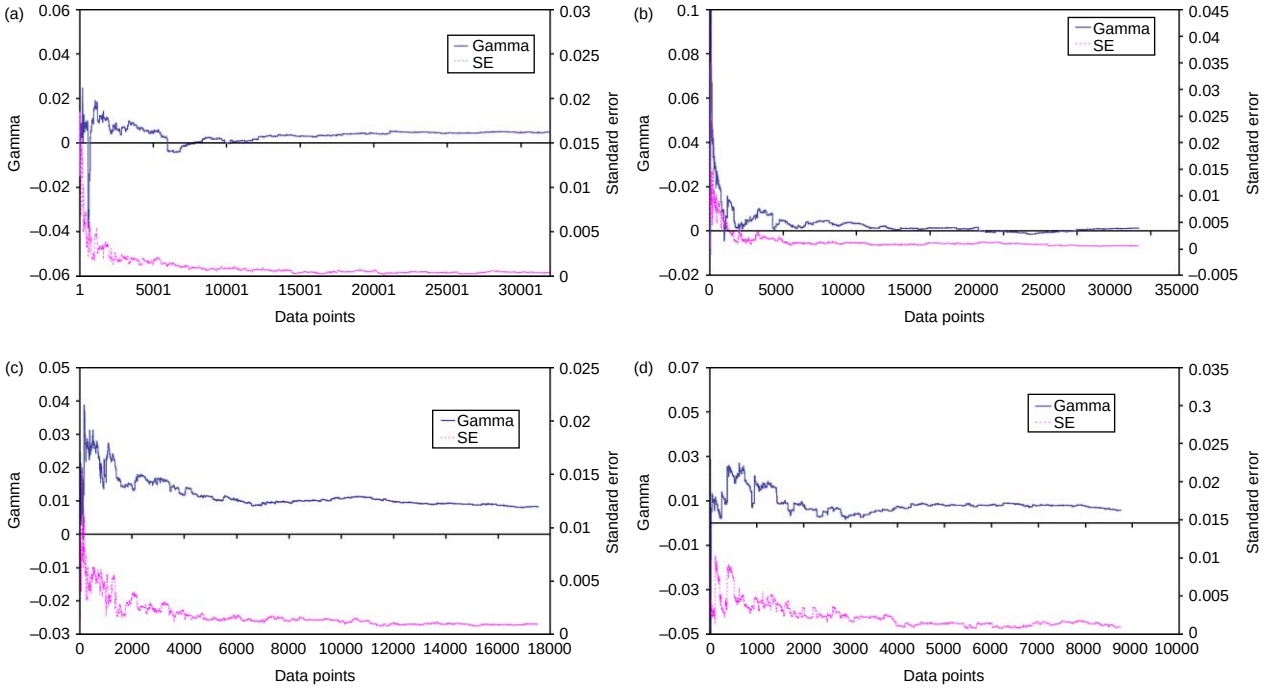

Figure 3 | Gamma and Standard Error (SE) variations for different data frequencies: (a) 15 min data, (b) 30 min data, (c) 60 min data and (d) 120 min data.

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

a value of 0.00273 at around 16,004 data points (i.e. $M = 16{,}004$). The standard error (SE) corresponding to $M = 16{,}004$ was relatively small at ~0.00041. As a result, $M = 16{,}004$ could be used effectively to construct a reliable model and we used 16,004 data points for the ANN training. Similarly, the Gamma Test-based analysis have identified [2, 2], [3, 3] and [4, 3] input combinations as the best ones with the least Gamma statistic value for the datasets with time intervals of 30 min, 60 min and 120 min. The M test details can be found in Table 1.

It is interesting to note that the input combination with only four vectors (viz. $Q(t)$, $Q(t-1)$, $P(t)$ and $P(t-1)$) was identified as the best for 30 min data whereas for other time intervals the best input combinations consist of more than four input vectors. The M test analysis on the 30 min data is shown in Figure 3(b). It was identified that $M = 13{,}830$ was the best length for the training data with the least value of the Gamma statistic 0.00064 and corresponding SE as 0.00085. The embedding 11001100 model (four input and one output set of I/O pairs) was identified as the best structure for the 30 min data with a low noise level (Γ value) and rapid fall-off SE value. It shows that, for the 30 min scenario, a reliable nonlinear predictive model could be built using around 13,830 data points and the remaining data out of the total 35,031 data points could be used as the validation dataset. The M test results for 60 min and 120 min data are shown in Figure 3(c, d). The training data length for 60 min and 120 min datasets are identified as 6,628 and 2,893 with the corresponding Gamma values as 0.00856 and 0.00159, respectively.

Entropy Theory (ET) is also used to identify the best embedded input combination. The entropy information is an indicating factor of the best input combination. The combination with higher entropy information can be considered as the one with a potentially better predictability. The variation of entropy information with different major input combinations is shown in Figure 4. It can be found that the 15 min data are the best for the modelling and also we can observe that the entropy information for 30 min data is close to that of the 15 min data for most of the combinations. The entropy theory identifies the best combination for different data time intervals (15 min, 30 min, 60 min and 120 min) as [4, 4], [3, 3], [4, 4] and [4, 4] while the corresponding findings by the Gamma Test

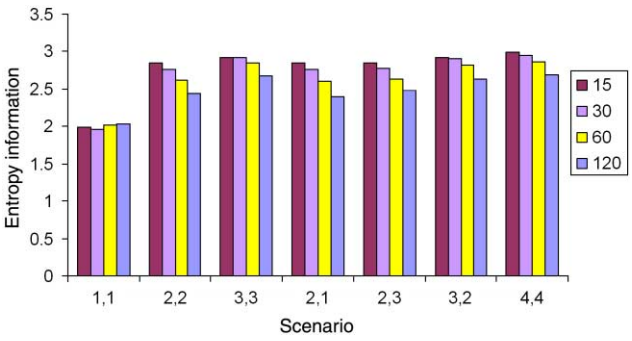

Figure 4 | Entropy information variation corresponding to major input combinations.

are [4, 4], [2, 2], [3, 3] and [4, 3]. Although there are differences in the identified combinations, it is interesting to note that both techniques found the smallest number of input vectors for 30 min datasets. The variation of the Gamma statistic value for different input combinations is shown in Figure 5. This figure illustrates the 30 min data are the best for modelling as the Gamma statistic value is the smallest compared with other datasets. In this study we used the Gamma Test results [4, 4], [2, 2], [3, 3] and [4, 3] combinations for modelling because of the smaller input vectors compared with the entropy theory's results.

To check the authenticity of GT analysis, we performed a cross-correlation analysis (Tayfur & Guldal 2006) between the target runoff dataset $Q(t)$ and different lag time series of precipitation and runoff using daily rainfall–runoff data. A study by Remesan *et al.* (2009) identified three-step antecedent runoff values ($Q(t-1)$, $Q(t-2)$, $Q(t-3)$), one-step antecedent rainfall ($P(t-1)$) and current rainfall information ($P(t)$) are the best for daily rainfall–runoff modelling, with a training data length of

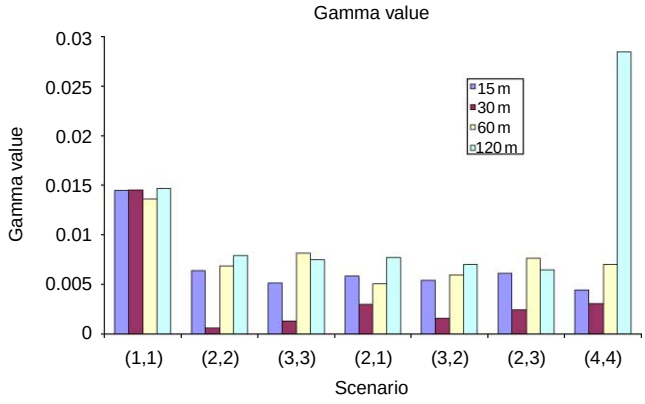

Figure 5 | Variation of Gamma value corresponding to different input combinations.

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

Table 2 | Cross-correlations of $Q(t)$ with different time lags of precipitation ($P(t-i)$) and runoff data ($Q(t-i)$). Bold in table shows selected antecedent rainfall and runoff series for modelling

Time lags	Correlation coefficient	
	$Q(t)$ vs $P(t-i)$	$Q(t)$ vs $Q(t-i)$
Zero-day ($i=0$)	**0.2918**	1.0000
One-day ($i=1$)	**0.1636**	**0.2674**
Two-day ($i=2$)	0.0764	**0.1888**
Three-day ($i=3$)	0.0533	**0.1101**
Four-day ($i=4$)	0.0312	0.0589

1,056 data points. The cross-correlation analysis between the target runoff dataset $Q(t)$ and different lag time series of precipitation and runoff data (viz. $Q(t-1)$, $Q(t-2)$, $Q(t-3)$, $Q(t-4)$, $P(t-1)$, $P(t-2)$, $P(t-3)$ and $P(t-4)$) were performed to see if the results matched the GT findings. The analysis results are shown in Table 2. From the table, up to a time lag of 3 d, the cross-correlations are higher for the runoff information, whereas, for the precipitation, the cross-correlation is much smaller just after a time lag of 1 d. It indicates that the runoff time series with a higher time lag than 3 d and precipitation time series after a time lag of 1 d wouldn't possess any significant effect on the target runoff data, $Q(t)$, as the cross-correlation values are close to zero. These cross-correlation results are matched with the results obtained from the Gamma Test.

To confirm the reliability of the GT in identifying the training data length, a data partitioning approach was adopted (Tayfur & Guldal 2006). Different scenarios of data partitioning into training and testing periods were tried in order to determine the optimal length of training data required for modelling without overfitting during training. Figures 6 and 7 show different partitioning scenarios and the related CORR and RMSE values for each scenario during training and validation. As per Figure 7 the best RMSE value of 0.429 m^3/s was obtained when 1,500 data points were used in the training. With the training results alone, one can observe that employing more than 750 data points in the training period would result in satisfactory runoff predictions with higher CORR values. At the same time we can observe a reduction in statistical term values (higher RMSE and lower value of CORR) in the validation phase after 1,100 training data length. It is an indication of overfitting with better statistical values during training and

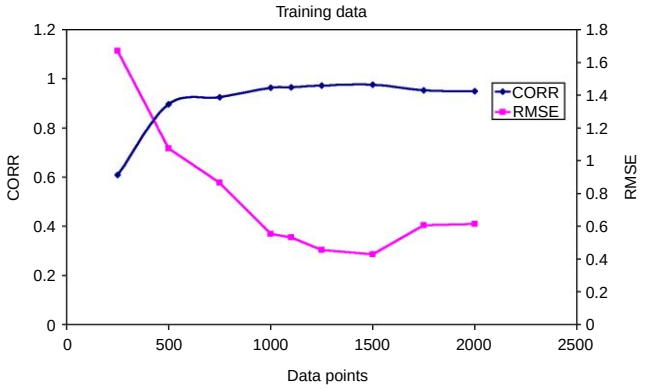

Figure 6 | Variation of CORR and RMSE in data splitting analysis in training data range.

poor values in the validation phase. As stated above, according to these figures, the optimum value of training data length is in the 1,000–1,100 range, whereas the GT identified the optimum length of the training data as 1,056. The data partitioning approach can give an indicative idea of the optimum data length while the GT can provide a more accurate estimate of the optimum data length.

Effect of data time interval on modelling

After selection of the final ANN structure with the Gamma Test (GT), modelling performance was evaluated with the statistical indices for the data collected at different frequencies from the Brue catchment (i.e. 15 min, 30 min, 60 min and 120 min data) and 2- to 6-h lead time forecasts. The scatter plots of 2-h-ahead, 4-h-ahead and 6-h-ahead forecasted versus observed discharges for the training data with a sampling time interval of 30 min are shown in Figure 8(a–c).

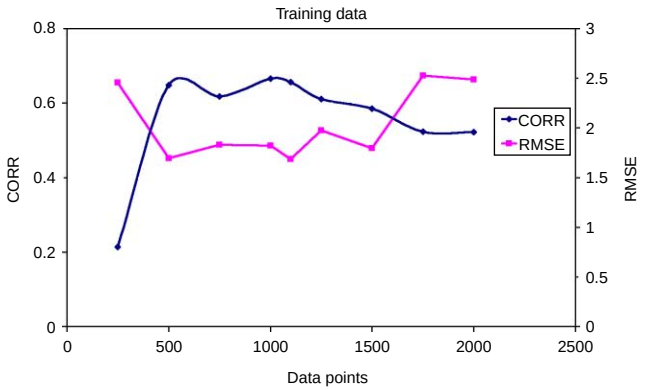

Figure 7 | Variation of CORR and RMSE in data splitting analysis in validation data range.

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

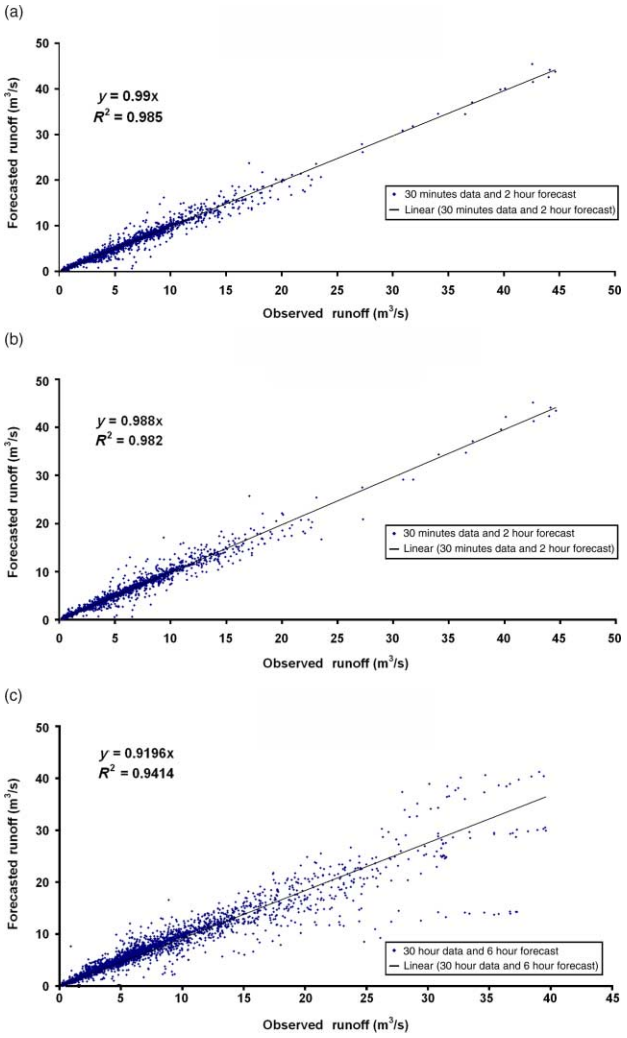

Figure 8 | Scatter plots of observed and forecasted runoff in the study area during the training period using data time interval of 30 min. (a) 2 h forecast, (b) 4 h forecast and (c) 6 h forecast.

The hydrographs of 4-h-ahead forecasted data versus the observed values of discharges of training data with the 30 min data are shown in Figure 9 and the 6-h-ahead forecasted hydrographs shown in Figure 10.

The values of the performance indices for the MSA forecast are presented in Table 3. The linear correlation coefficient and slope indices between the observed and computed runoff were consistent and values near one for 30 min data during the calibration as well as the validation period. One can note that the forecasted values have good correlations with the observed values for the 30 min data even for a 6-h-ahead forecast. The Nash coefficient of efficiency of the model, which determines the capability of

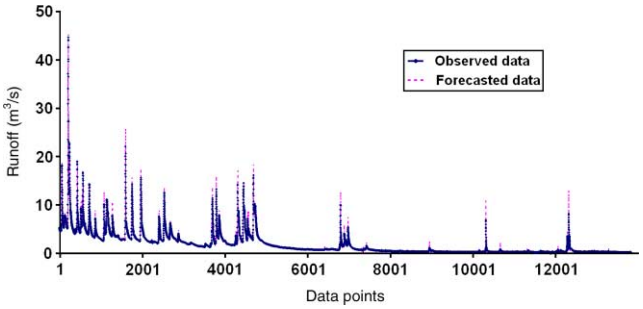

Figure 9 | Observed and 4-h-ahead forecasted runoff of the study area using data with 30 min sampling frequency.

the model in predicting runoff values, is also shown in Table 3. The Nash efficiency is more than 93% during the calibration and validation periods for 30 min data set in all lead time predictions, which is at an acceptable level for runoff modelling (Shamseldin 1997). Followed by the 30 min datasets, better predictions can be observed for the 60 min datasets, though the efficiency values were less than 93% for 6-h lead-time predictions in both validation and training data. The RMSE statistic values indicate the 30 min data are better for modelling as shown from both validation and training processes as the corresponding RMSE values are low. The studied ANN model forecast the flows with very close RMSE values for 30 min time steps; which were 0.3437 m^3/s, 0.3646 m^3/s and 0.385 m^3/s, respectively, with lead-times of 2 h, 4 h and 6 h, respectively, for the training data. The corresponding RMSE values for the validation data were 0.5513 m^3/s, 0.7915 m^3/s and 0.8922 m^3/s for the lead times 2 h, 4 h and 6 h, respectively. The variation of the Nash efficiency for different data frequencies along the forecast range is presented in Figure 11, from which it is clear that the model performance is superior for 30 min time

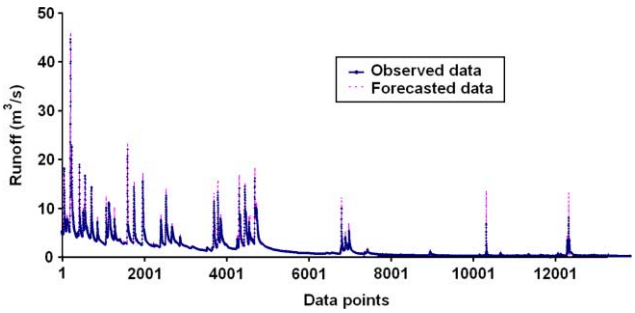

Figure 10 | Observed and 6-h-ahead forecasted runoff of the study area using data with 30 min sampling frequency.

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

Table 3 | Statistical indices of the modelling results corresponding to each data time interval at the study area

	Training			Validation		
Data time interval	**2 h forecast**	**4 h forecast**	**6 h forecast**	**2 h forecast**	**4 h forecast**	**6 h forecast**
RMSE (m^3/s)						
15 min	0.3704	0.7939	1.356	0.6242	1.114	1.8019
30 min	0.3437	0.3646	0.3850	0.5513	0.7915	0.8922
60 min	0.2884	0.5842	0.8800	0.6680	0.9683	1.4588
120 min	0.3547	0.5618	0.9426	0.5726	1.2919	1.397
MAE(m^3/s)						
15 min	0.0355	0.0840	0.1063	0.1355	0.5114	3.3850
30 min	0.0350	0.0415	0.03613	0.0458	0.0776	0.0646
60 min	0.0387	0.1138	0.1451	0.0520	0.1474	0.1815
120 min	0.0456	0.0833	0.1793	0.0799	0.1365	0.2415
MBE (m^3/s)						
15 min	−0.0103	−0.00092	−0.0241	−0.0463	−0.1542	−0.8486
30 min	0.00173	0.00136	0.00061	−0.0319	−0.0480	−0.0437
60 min	−0.00044	−0.00266	0.00041	−0.0439	−0.0589	−0.1058
120 min	−0.00316	−0.00452	−0.0019	−0.0014	0.0046	−0.0006
R and slope						
15 min	0.99 (0.99)	0.93(0.96)	0.83(0.91)	0.98(0.95)	0.9(0.86)	0.81(0.83)
30 min	0.98(0.99)	0.98(0.99)	0.98(0.99)	0.98(0.96)	0.96(0.93)	0.92(0.94)
60 min	0.99(0.99)	0.95(0.97)	0.89(0.93)	0.97(0.92)	0.93(0.90)	0.84(0.93)
120 min	0.99(0.99)	0.96(0.97)	0.89(0.93)	0.97(0.98)	0.87(0.96)	0.89(0.94)
Efficiency						
15 min	0.99	0.94	0.88	0.98	0.88	0.69
30 min	0.99	0.98	0.98	0.98	0.95	0.94
60 min	0.99	0.96	0.9	0.97	0.93	0.84
120 min	0.99	0.96	0.9	0.97	0.87	0.89

step to other time steps at all lead times. The RMSE variation corresponding to different data time intervals and lead times for training and validation data is shown in Figures 12 and 13. From Figure 9 we can note that the minimum value of RMSE for a 2 h forecast is observed corresponding to 60 min time steps which means for low lead times 60 min time step data are equally good as that of 30 min. As was expected, all the data time intervals have shown an increase in RMSE with longer lead times. A steep slope of the RMSE was observed between the 15 min and 30 min data in both training and validation phase for long lead times like 4 h and 6 h forecasts; while the corresponding slope was flat for short lead time (2 h prediction) in both cases. This indicates that the model data time interval

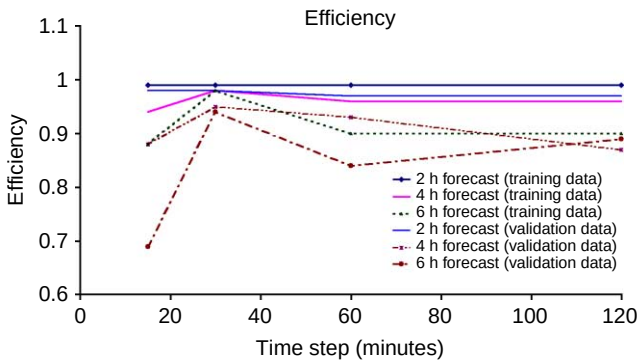

Figure 11 | Variation of modelling efficiency with data time interval and forecasting lead time steps.

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

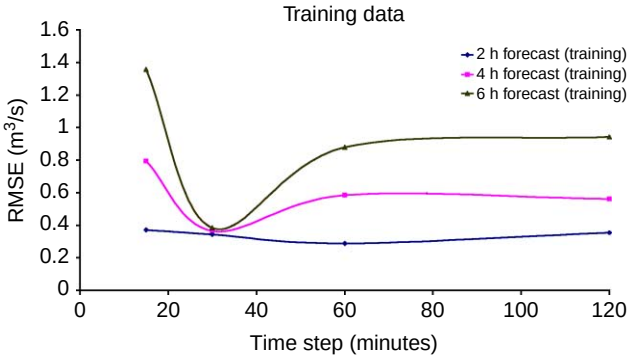

Figure 12 | Variation of RMSE with data time interval and forecasting lead time steps for the training data.

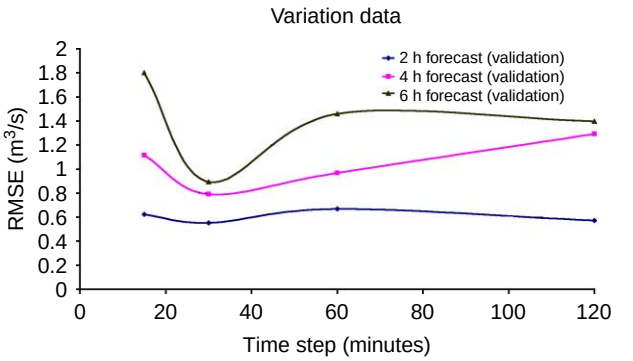

Figure 13 | Variation of RMSE with data frequency and forecasting lead time steps for the validation data.

definitely had an influence on the model's prediction and this case study confirms the dependence of the data time interval on the long lead-time prediction. It is also observed that, for a short lead time, the curve is relatively flat. Thus the influence of the data time interval hasn't got much effect on short lead-time prediction results.

CONCLUSIONS

Despite a large number of published papers on hydrological ANN applications, there are still many unsolved problems. Modern data collection and telecommunication technologies can provide high resolution data with very fine sampling intervals (such as 15 min or even less). However, there is little research on the optimal data time interval for modelling. We hypothesised that either too large or too small time intervals were detrimental to a model's performance. There should be an optimal time interval for a particular catchment in terms of hydrological modelling. So far, there is no published research about the optimal time interval for ANN models (and other data mining models). This study has demonstrated that the data time interval does have a major impact on the model's performance. For the Brue catchment, it has been found that a 30 min interval is the optimal. It is interesting to note that the significance of the time interval influence is more prominent for longer lead times than shorter ones. This is highly relevant to flood forecasting since a flood modeller usually tries to stretch his/her model's lead time as far as possible. If the selection of the data time interval is not considered, the model developed will not be performing at its full potential. Clearly, more research is needed to explore the relationship between optimal time intervals and catchment characteristics (e.g. catchment concentration time). Therefore we hope this study will stimulate further study of this problem in different catchments and with various data mining models so that some generalisation or pattern on the optimal time interval could be found. The application of the Gamma Test and Information Entropy in this paper will help readers to speed up their data input selection process.

REFERENCES

Abrahart, R. J. & See, L. M. 2007 Neural network modelling of non-linear hydrological relationships. *Hydrol. Earth Syst. Sci.* **11** (5), 1563–1579.

Agalbjörn, S., KonHar, N. & Jones, A. J. 1997 A note on the gamma test. *Neural Comput. Appl.* **5** (3), 131–133.

Amorocho, J. & Espildora, B. 1973 Entropy in the assessment of uncertainty of hydrologic systems and models. *Water Resour. Res.* **9** (6), 1511–1522.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 2000 Artificial neural networks in hydrology I: preliminary concepts. *J. Hydrol. Eng.* **5** (2), 115–123.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 2000 Artificial neural networks in hydrology II: hydrologic applications. *J. Hydrol. Eng.* **5** (2), 124–137.

Avci, E. 2007 Forecasting daily and sessional returns of the ISE-100 index with neural network models. *Doğuş Üniversitesi Dergisi* **8** (2), 128–142.

Bray, M. & Han, D. 2005 Identification of support vector machines for runoff modelling. *J. Hydroinf.* **6** (4), 265–280.

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest

Caselton, W. F. & Husain, T. 1980 Hydrological networks: information transmission. *J. Water Res. Plann. Manage.* **106** (WR 2), 503–520.

Chang, L. C., Chang, F. J. & Chiang, Y. M. 2004 A two-step ahead recurrent neural network for streamflow forecasting. *Hydrol. Process.* **18**, 81–92.

Chang, F. J., Chiang, Y. M. & Chang, L. C. 2007 Multi-step-ahead neural networks for flood forecasting. *Hydrol. Sci. J.* **52** (1), 114–130.

Chapman, T. G. 1986 Entropy as a measure of hydrologic data uncertainty and model performance. *J. Hydrol.* **85**, 111–126.

Cigizoglu, H. K. & Kisi, O. 2005 Flow prediction by three back propagation techniques using k-fold partitioning of neural network training data. *Nordic Hydrol.* **36** (1), 1–16.

Cluckie, I. D. & Han, D. 2000 Fluvial flood forecasting. *Water Environ. Manage.* **14** (4), 270–276.

Duan, Q., Sorooshian, S. & Gupta, V. K. 1992 Effective and efficient global optimization for conceptual rainfall–runoff models. *Water Resour. Res.* **28** (4), 1015–1031.

Durrant, P. J. 2001 *win Gamma: A non-linear data analysis and modelling tool with applications to flood prediction*. PhD thesis, Department of Computer Science, Cardiff University, Wales, UK.

El-Bakyr, M. Y. 2003 Feed forward neural networks modeling for K-P interactions. *Chaos Solut. Fractals* **18** (3), 995–1000.

Evans, D. 2002 *Data Derived Estimates of Noise Using Near Neighbour Asymptotics*. PhD Thesis, Department of Computer Science, University of Cardiff, UK.

Evans, D. & Jones, A. J. 2002 A proof of the gamma test. *Proc. R. Soc. Ser. A* **458** (2027), 2759–2799.

Fernando, T. M. K. G., Maier, H. R., Dandy, G. C. & May, R. 2005 Efficient selection of inputs for artificial neural network models. In: Zerger, A. & Argent, R. M (eds) *MODSIM 2005 International Congress on Modelling and Simulation, December*. Modelling and Simulation Society of Australia and New Zealand, Canberra, Australia. pp. 1806–1812. Available at: http://www.mssanz.org.au/modsim05/papers/fernando.pdf

Ghafari, G. M., Moghaddamnia, A., Piri, J., Amin, S. & Han, D. 2009 Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. *Adv. Water Res.* **32** (1), 88–97. (doi:10.1016/j.advwatres.2008.10.005).

Han, D., Cluckie, I. D., Karbassioun, D., Lawry, J. & Krauskopf, B. 2002 River flow modelling using fuzzy decision trees. *Water Res. Manage.* **16** (6), 431–445.

Han, D., Chan, L. & Zhu, N. 2007*a* Flood forecasting using support vector machines. *J. Hydroinf.* **9** (4), 267–276.

Han, D., Kwong, T. & Li, S. 2007*b* Uncertainties in real-time flood forecasting with neural networks. *Hydrol. Process.* **21**, 223–228.

Harmancioglu, N. & Yevjevich, V. 1987 Transfer of hydrologic information among river points. *J. Hydrol.* **91** (1/2), 103–118.

HEC (Hydrological Engineering Center) 1990 *HEC-1 Flood Hydrograph Package. Program User's Manual*. US Army Corps of Engineers, Davis, CA.

Hecht-Nielsen, R. 1990 *Neurocomputing*. Addison-Wesley, Reading, MA.

Jain, A., Sudheer, K. P. & Srinivasulu, S. 2004 Identification of physical processes inherent in artificial neural network rainfall runoff models. *Hydrol. Process.* **18**, 571–581.

Krstanovic, P. F. & Singh, V. P. 1992 Evaluation of rainfall networks using entropy I. *Water Res. Manage.* **6**, 279–293.

Maren, A. J., Harston, C. T. & Pap, R. M. 1990 *Handbook of Neural Computing Applications*. Academic, New York.

Maruyama, T. & Kawachi, T. 1998 Evaluation of rainfall characteristics using entropy. *J. Rainwater Catch. Syst.* **4** (1), 7–10.

Mechaqrane, A. & Zouak, M. 2004 A comparison of linear and neural network ARX models applied to a prediction of the indoor temperature of a building. *Neural Comput. Appl.* **13**, 32–37.

Michaud, J. & Sorooshian, S. 1994 Comparison of simple versus complex distributed runoff models on a midsized semiarid watershed. *Water Resour. Res.* **30** (3), 593–605.

Moghaddamnia, A., Remesan, R., Hassanpour Kashani, M., Mohammadi, M. & Han, D. 2009 Comparison of LLR, MLP, Elman, NNARX and ANFIS Models with a case study in solar radiation estimation. *J. Atmos. Solar-Terrestrial Phys.* **71** (8–9), 975–982.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models: 1. A discussion of principles. *J. Hydrol.* **10**, 282–290.

Piri, J., Amin, S., Moghaddamnia, A., Keshavarz, A., Han, D. & Remesan, R. 2009 Daily pan evaporation modelling in a hot and dry climate. *ASCE J. Hydrol. Eng.* **14** (8), 803–811.

Remesan, R., Shamim, M. A. & Han, D. 2008 Model data selection using gamma test for daily solar radiation estimation. *Hydrol. Process.* **22**, 4301–4309 (doi:10.1002/hyp.7044).

Remesan, R., Shamim, M. A. & Han, D. 2009 Runoff prediction using an integrated hybrid modelling scheme. *J. Hydrol.* **372** (1–4), 48–60 (doi:10.1016/j.jhydrol.2009.03.034).

Shamseldin, A. Y. 1997 Application of a neural network technique to rainfall runoff modeling. *J. Hydrol.* **199**, 272–294.

Shannon, C. E. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423.

Sharma, A. 2000 Seasonal to inter annual rainfall probabilistic forecasts for improved water supply management: part 1–a strategy for system predictor identification. *J. Hydrol.* **239**, 232–239.

Solomatine, D. P. & Ostfeld, A. 2008 Data-driven modelling: some past experiences and new approaches. *J. Hydroinf.* **10** (1), 3–22.

Tayfur, G. & Guldal, V. 2006 Artificial neural networks for estimating daily total suspended sediment in natural streams. *Nordic Hydrol.* **37** (1), 69–79.

Yang, Y. & Burn, D. H. 1994 An entropy approach to data collection network design. *J. Hydrol.* **157**, 307–324.

First received 11 August 2008; accepted in revised form 22 June 2009. Available online 2 February 2010

Downloaded from http://iwaponline.com/jh/article-pdf/12/4/396/476727/396.pdf
by guest