Towards Defining Evaluation Measures for Neural Network Forecasting Models

Dr Pauline Kneale; Dr Linda See; Mr Andrew Smith
School of Geography, University of Leeds LS2 9JT UK
P.Kneale@geog.leeds.ac.uk

Abstract. There is diversity in the use of global goodness-of-fit statistics to determine how well models forecast flood hydrographs. This paper compares the results from nine evaluation measures and two forecasting models. The evaluation measures comprise global goodness-of-fit measures as recommended by Legates and McCabe (1999) and Smith (2000) and more flood specific measures which gauge the ability of the models to predict operational alarm levels and the rising limb of the hydrograph. The models used in this particular study are artificial neural networks trained with backpropagation (BPNNs) and a Time Delay Neural Network (TDNN). Networks were trained to forecast stage on the River Tyne, Northumbria, for lead times ranging from 2 to 6 hours with minimal data. The training data set consisted of continuous data for one winter period and validation was undertaken using data from winter periods for three other years. The results showed that the TDNNs performed better than the BPNNs although the evaluation measures indicate that the performance of these particular models for operational purposes is not yet sufficient. Issues regarding the creation of appropriate training data sets and sampling procedures as well as the hydrological conditions of the River Tyne need to be investigated further. A consistent set of operational evaluation measures and benchmark data sets must be established before comparison of neural network and physical models can be facilitated. The combination of evaluation measures provides a good picture of the overall forecasting performance of the models, and suggests that operators should consider a range of measures.

1. INTRODUCTION

Climate trends in Northern Europe suggest that weather patterns are moving towards wetter and stormier conditions. Damage due to flooding in the past five years has been estimated at several billion pounds in the UK, while on an international scale, flooding poses life-threatening problems particularly in tropical and subtropical regions. Although global warming is now recognised as a real threat, there has been little success in establishing protocols for the reduction of greenhouse gases. Therefore, an increase in the occurrence of flooding is likely to continue into the future. Building new flood defences to contain higher flows, defending vulnerable sites and domestic and commercial insurance claims all have serious financial implications. For these reasons the timely forecasting of floods is becoming more crucial for both flood defence and catchment management purposes.

There are several integrated flood forecasting systems available. These systems generally comprise several interacting components, which usually include a rainfall-runoff forecasting model, a database for the management of historical and real-time data, a dissemination module which issues warnings to hydrologists monitoring the catchment, and a component which continually evaluates and updates the forecasting outputs (Nemec, 1986). The forecasting systems currently in operation within the EA vary between regional offices; for example, the River Flow Forecasting System (RFFS) developed by the Institute of Hydrology and Logica UK Ltd (Moore *et al.,* 1994) is currently used in Yorkshire and Northumbria, while the Wessex Radar Information Project (WRIP) is the

system employed in the South Western, Anglian and North West regions (Han *et al.,* 1997). These large-scale systems generally require a substantial investment of both time and money for their development and continued maintenance yet a recent assessment of the flood forecasting models used by the EA revealed that these systems are not yet satisfactory for use in an automated operational environment (Marshall, 1996).

The complexity of natural systems and the number of processes that continuously interact to influence river levels render traditional modelling based on mirroring natural processes with process-based equations very difficult. Where processes can be modelled numerically, there are constraints imposed by the spatial availability of data and the cost of collecting this data in real time. An alternative approach, which has been slowly gaining recognition within hydrology, is the application of artificial neural networks (ANNs) to a variety of hydrological forecasting problems. ANNs offer the potential for a more flexible, less assumption-dependent approach to modelling flood processes, and they have already been demonstrated to work successfully as substitutes for rainfall-runoff models (Minns and Hall, 1996; Smith and Eli, 1995; Abrahart and Kneale, 1997; Khondker *et al.,* 1998). Most of the studies in the literature make use of feedforward multi-layer perceptrons trained with backpropagation (Rumelhart *et al.*, 1986), hereafter referred to as BPNNs. However, there are many other ANN algorithms and architectures that have yet to be investigated (Shepherd, 1997). Time Delay Neural Networks (TDNNs), which are used in speech recognition (Waibel, 1989), provide one type of network that could especially benefit flow forecasting because these networks can potentially find the temporal relationship between inputs. This has

relevance for the inclusion of upstream stations, which must be lagged in the input data by an average travel time when used with BPNNs. Multi-net approaches (Sharkey, 1999) offer another promising area for research in developing better ANN flood forecasting models. However, the ultimate test of their operational feasibility lies in their performance in real-time flood forecasting. To achieve this comparison, a set of operational evaluation measures for existing systems must be established along with a repository for benchmark data sets. In his review of EA operational forecasting models, Marshall (1996) highlighted the lack of a consistent evaluation system between offices, while much of the research in the literature reports performance using global goodness-of-fit statistics, which are rarely useful for determining how well models can predict flood hydrographs.

This paper presents a set of comprehensive evaluation measures that can be used to compare results from forecasting models. A set of BPNNs and TDNNs were developed to forecast stage on the River Tyne, Northumbria. The networks were trained on a continuous data set for one winter period and validated on winter periods for three other years. Models were developed to forecast lead times ranging from 2 to 6 hours. The majority of the evaluation measures are based on the ability to predict operational alarm levels and the rising limb of the hydrograph. The data set has been kept very small to see what can be done with relatively few storms, so absolute performance results here are of less interest than the relative results. It is hoped that this set of measures will provide a useful tool for comparing different models within a particular study catchment and initiate a dialogue for the establishment of a consistent set of operational evaluation measures and benchmark data sets in the future.

2. STUDY AREA

The study area is the upper, non-tidal section of the Tyne River, North East England as shown in Figure 1. The Tyne basin has an area of approximately 2,920 km^2. The lowest non-tidal stage gauge on the Tyne is located at Bywell, upstream of Newcastle. Historic data of stage and rainfall values are available, measured at a handful of stations throughout the North Tyne catchment. Telemetered data at 15 minute intervals for the years 1994 to 1998 were provided by the Environment Agency for Bywell and three upstream stations on the River Tyne: Reaverhill, Rede Bridge and Ugly Dub. The forecasting task is to determine if Bywell can be predicted using only information from the north portion of the Tyne in the hypothetical event that the telemetred network on the south Tyne should fail. The 15 minute data were converted to an hourly resolution. Figure 2 shows the relationship between the levels at upstream stations and Bywell, which serves to illustrate that a simple linear relationship does not exist and therefore some form of nonlinear modelling is required.

The average travel times between stations were calculated by plotting hydrographs of historical storm events. They represent the typical time difference for a peak flow passing between two stage gauges. Data from the upstream stations were offset by the average travel time before training with the BPNNs. However, the TDNNs did not see the lagged data in order to determine whether these networks could model the travel time as part of the training process.
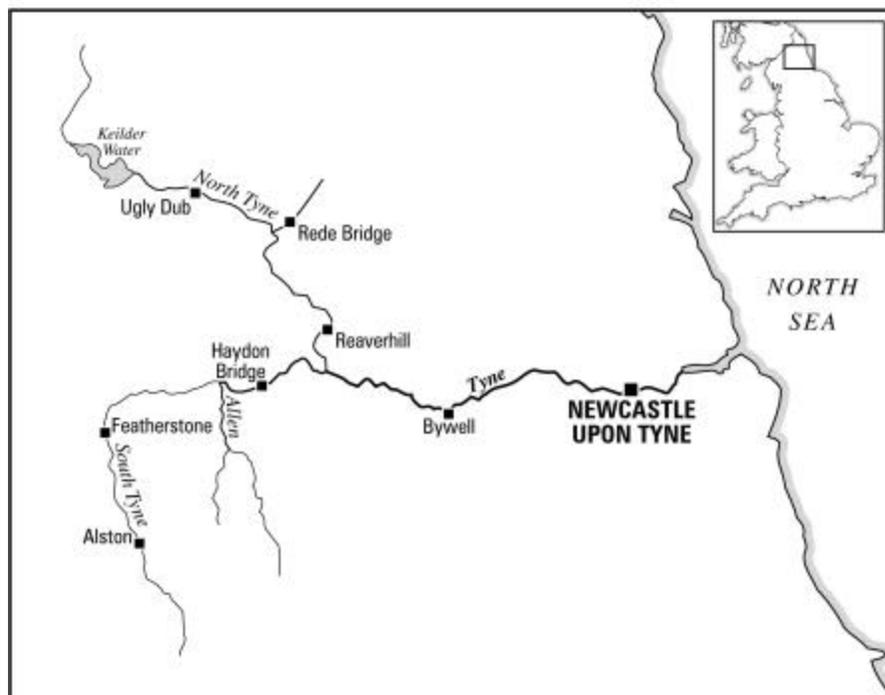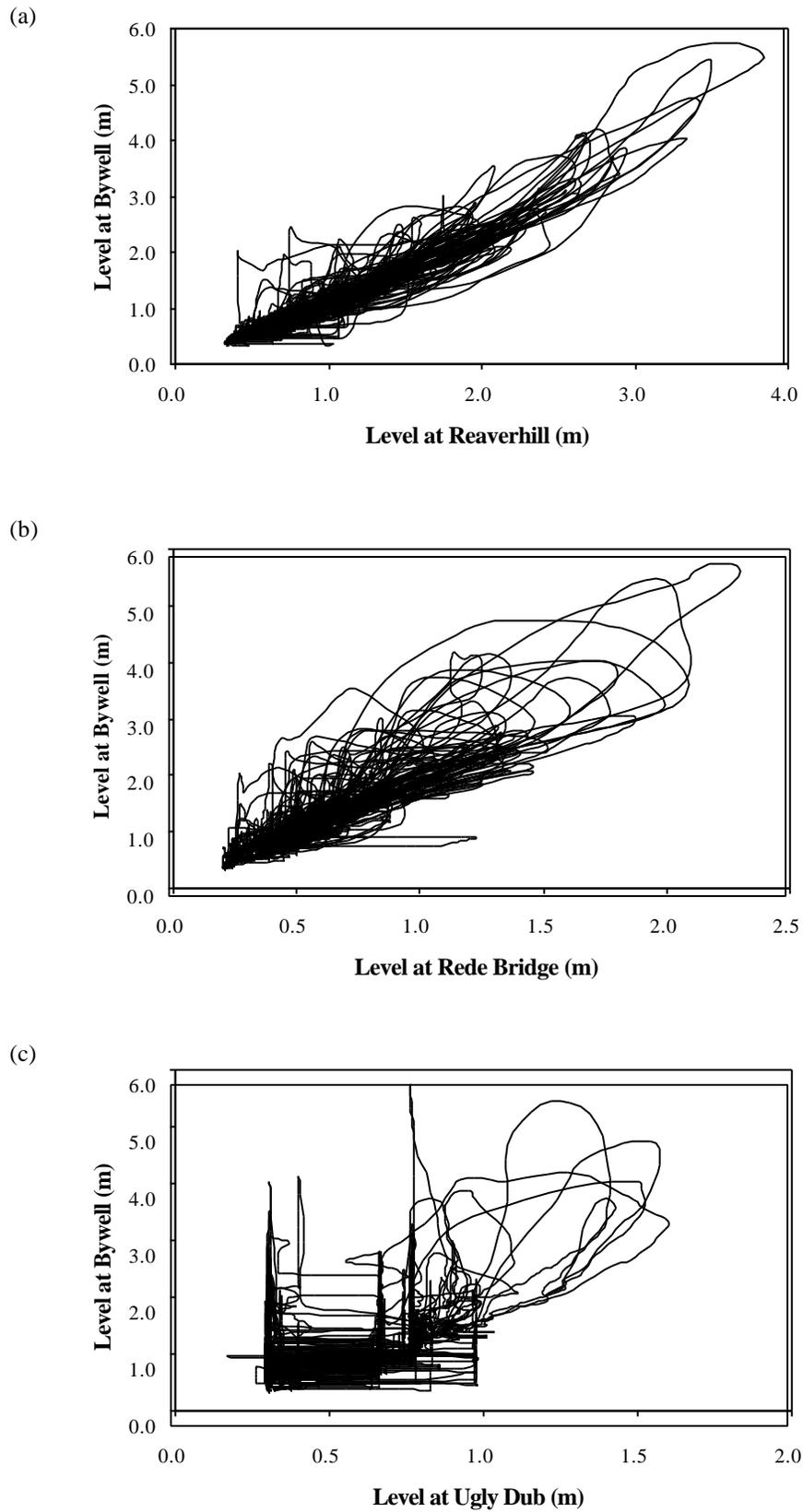

Figure 1: Location of the Tyne River in the UK

(a)

(b)

(c)

Figure 2: Plots of river level at Bywell against levels at (a) Reaverhill (b) Rede Bridge and (c) Ugly Dub

Every major river under the jurisdiction of the Environment Agency has between one and four flood risk warning levels defined at a range of points along the river. Each warning level represents a particular level of risk of flood occurrence when the river reaches that particular level. Risk is measured in terms of the human and economic consequences of the flood. The alarm levels for Bywell gauging station in the 1990-1999 period and the interpretation of the alarm levels are shown in Table 1; the status and interpretations were revised in 2000. These alarm levels were used in devising the set of operational evaluation measures.

Table 1: Alarm levels for the River Tyne at the Bywell stage gauge (Source: Environment Agency)

Status	Interpretation	Level
Alarm	First operational warning level	3.5m
Yellow	Flooding of low lying farmland and roads near rivers and the sea	4.4m
Amber	Flooding may affect isolated properties, roads and large areas of farmland near rivers and the sea	5.3m
Red	Flooding will affect many properties, roads and large areas of farmland	5.8m

3. ERROR MEASURES FOR EVALUATION

Evaluation measures are frequently referred to as "goodness-of-fit" measures, since they measure the degree to which the predicted hydrograph fits the actual hydrograph. Goodness-of-fit statistics can be grouped into two types: relative and absolute.

3.1 Relative goodness-of-fit statistics

Relative goodness-of-fit measures are non-dimensional indices, which provide a relative comparison of the performance of one model against another. There are three commonly used relative goodness-of-fit statistics:

(a) The Coefficient of Determination

$$R^2 = \left[\frac{\sum_{i=1}^{N}(O_i - \overline{O})(P_i - \overline{P})}{\sqrt{\sum_{i=1}^{N}(O_i - \overline{O})^2}\sqrt{\sum_{i=1}^{N}(P_i - \overline{P})^2}} \right]^2 \tag{1}$$

where O_i is the observed value at time i, P_i is the predicted value at time i, N is the total number of observations, $\overline{O}$ is the mean of O over N and $\overline{P}$ is the mean of P over N. It measures the degree of collinearity between two sets of values. The coefficient has a range of 0.0 to 1.0 with a higher value indicating a higher degree of collinearity.

(b) The Coefficient of Efficiency (Nash and Sutcliffe, 1970)

$$E = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \tag{2}$$

where a value of 1.0 represents a 'perfect' prediction while a model with an index of 0.0 is no more accurate than predicting the mean observed value for all i. This measure was introduced by Nash and Sutcliffe (1970) and is widely used in the hydrological literature.

(c) The Index of Agreement

$$d = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}\left[|P_i - \overline{O}| + |O_i - \overline{O}| \right]^2} \tag{3}$$

which was proposed by Willmott (1981) as an adaptation of the Nash and Sutcliffe efficiency index. The alteration to the denominator seeks to penalise differences in the mean predicted and mean observed values. Inspection shows the degree of similarity between the coefficients of determination and efficiency. Due to the squaring of the error terms in all three measures, they are considered to be overly sensitive to outliers in the data set.

More generic forms of both E and d can be developed, by varying the power applied to the error terms to an arbitrary, positive integer power j, denoted as E_j and d_j respectively:

$$E_j = 1 - \frac{\sum_{i=1}^{N}|O_i - P_i|^j}{\sum_{i=1}^{N}|O_i - \overline{O}|^j} \tag{4}$$

$$d_j = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}\left[|P_i - \overline{O}| + |O_i - \overline{O}| \right]^2} \tag{5}$$

By this convention the original coefficient of efficiency E becomes E_2 and the original index of agreement d becomes d_2. Particularly useful are E_1 and d_1 since difference terms are not inflated by their distance from the mean.

3.2 Absolute goodness-of-fit statistics

In contrast, absolute goodness-of-fit statistics are measured in the units of the flow/stage measurement. The following are commonly used absolute goodness-of-fit statistics:

(a) Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - P_i)^2}{N}} \qquad (6)$$

As with the various relative statistics, the squared error term places undue importance on the outliers in the dataset.

(b) Mean Absolute Error (MAE)

$$MAE = N^{-1}\sum_{i=1}^{N}|O_i - P_i| \qquad (7)$$

Taking the absolute value of the error term rather than its square removes the bias towards outlying points in the data set.

(c) Comparisons of Means and Standard Deviations

$$? m = \overline{O} - \overline{P}$$
$$? s = s_O - s_P \qquad (8)$$

Whilst not in themselves fully descriptive goodness-of-fit statistics, comparisons of the mean of the observed and predicted values and their standard deviations do provide a useful measure of the performance of the model.

3.3 Flood Specific Evaluation Measures

The following are a list of additional evaluation measures that are specifically aimed at measuring flood forecasting performance:

(a) Proportion of time in the correct state of alarm

$$E = N^{-1}\sum_{i=1}^{N} f(O_i, P_i)$$
$$f(O_i, P_i) = \begin{cases} 1 & O_i, P_i \text{ are in the same state} \\ 0 & O_i, P_i \text{ are in different states} \end{cases} \qquad (9)$$

A useful model will be able to accurately predict when a river enters a higher flood risk alarm status, since this will allow precautionary measures to be taken earlier. Hence the lower the proportion of missed, late or false alarms to correctly predicted alarms, the better the performance of the model. The score is reduced by both a failure to predict an alarm event that subsequently occurs, and by false alarms for events that do not occur.

The next two measures were originally suggested by Smith (2000) in his investigation of neural network forecasting models for the River Tyne. These were found to be useful indicators in combination with the other measures outlined above.

(b) Root Mean of the Flow Weighted Error (RM_FWE)

$$RM\_FWE = \sqrt{\frac{\sum_{i=1}^{N} O_i|O_i - P_i|}{N}} \qquad (10)$$

For each data point, the absolute error between the observed and predicted values is calculated, which is then multiplied by the observed level. The mean of these flow weighted errors is then calculated in the usual manner. This way the sensitivity of the error calculation is increased for high flow and storm events and reduced for low and normal flows. However, the multiplication by observed level implies that the importance of the accuracy of prediction is linearly proportional to the observed level. Whilst there is no reason to suppose that a linear scale of importance is the most appropriate, it is plausibly more appropriate than rating importance independent of flow.

(c) Root Mean Gradient Weighted Error

$$RM\_GWE = \sqrt{\frac{\sum_{i=2}^{N}|O_i - O_{i-1}||O_i - P_i|}{N-1}} \qquad (11)$$

Whilst the RM_FWE has increased sensitivity to high flow data, it is still relatively insensitive to the errors in the lower part of the rising limb of a hydrograph. For this reason a gradient weighted error measurement is also proposed. For each data point, the absolute error between the observed and predicted values is calculated. This is multiplied by the absolute difference between the current observed value and the previous observed value, which is approximately proportional to the current gradient of the hydrograph. The mean of these flow weighted errors is then calculated in the usual manner. In this way the sensitivity of the error calculation is increased for periods of rapid change in levels and reduced for periods of stable flow.

There are two disadvantages to this measure. Unlike RM_FWE, sensitivity is reduced for any stable period of high flow. In addition to increasing sensitivity to the rising limb of the hydrograph, sensitivity is also increased to the falling limb, which is of less importance for flood risk warning. It is not possible to use the true (signed) gradient since this would improve scores for poorly fitting falling limbs and reduce those where the fit was good. Despite these problems, it may prove that the RM_GWE is a useful measure in addition to the RM_FWE.

3.4 Suggested Evaluation Measures for Operational Flood Forecasting

Many of the principal measurements that are used in the hydrological literature have been reviewed by Legates and McCabe (1999). They strongly discourage the use of R^2 as a goodness-of-fit statistic

as well as E_2 and d_2, but recommend the use of E_1 and d_1 instead to reduce the influence of outliers and because E_1 has the advantage that the value 0.0 has a meaning that can be directly interpreted. In addition they recommend that a complete assessment of model performance should include at least one absolute error measure (e.g. RMSE or MAE) with any supplementary information such as comparison of means and standard deviations. In their conclusion they stress that any individual goodness-of-fit statistic is only one tool in assessing the performance of a model and that a range of statistics are required. For the development of a set of evaluation measures, the recommendations of Legates and McCabe (1999) are followed for both the entire data set and for data above the alarm level as well as including the more flood specific evaluation measures outlined in section 3.3.

4. ANN EXPERIMENTS ON THE UPPER TYNE RIVER

This section provides an overview of the artificial neural network models used in this study and briefly describes the ANN modelling experiments for the Upper Tyne River.

4.1 Artificial Neural Network Models

ANNs are information processors, trained to represent the implicit relationships and processes that are inherent within a data set. The original inspiration for ANNs was biological so much of the terminology of ANNs reflects this biological heritage. Good descriptions can be found in Kasabov (1996), Bishop (1995) and Lin and Lee (1996). The basic structure of an ANN consists of a number of simple processing units, also known as neurons or nodes. The basic role of each node is to take the weighted sum of the inputs and process this through an activation function such as the sigmoid. A connection or link joins the output of one node to the input of another. Each link has a 'weight', which represents the strength of the connection. Collectively, the values of all the weights in a network represent the current state of learning of the network, in a distributed manner. These weights are altered during the training process to ensure that the inputs produce an output that is close to the desired value. The arrangement of nodes and interconnecting links is called the architecture. In many architectures, nodes are arranged in layers, whereby there are no links between nodes in the same layers. Data enters the network through the input units of the input layer, are fed forward through successive hidden layers and emerge from the output units in the output layer of the network. This is called a feedforward network or multi-layer perceptron (MLP) (Rumelhart *et al.*, 1986) because the flow of information is in one direction only from input to output units.

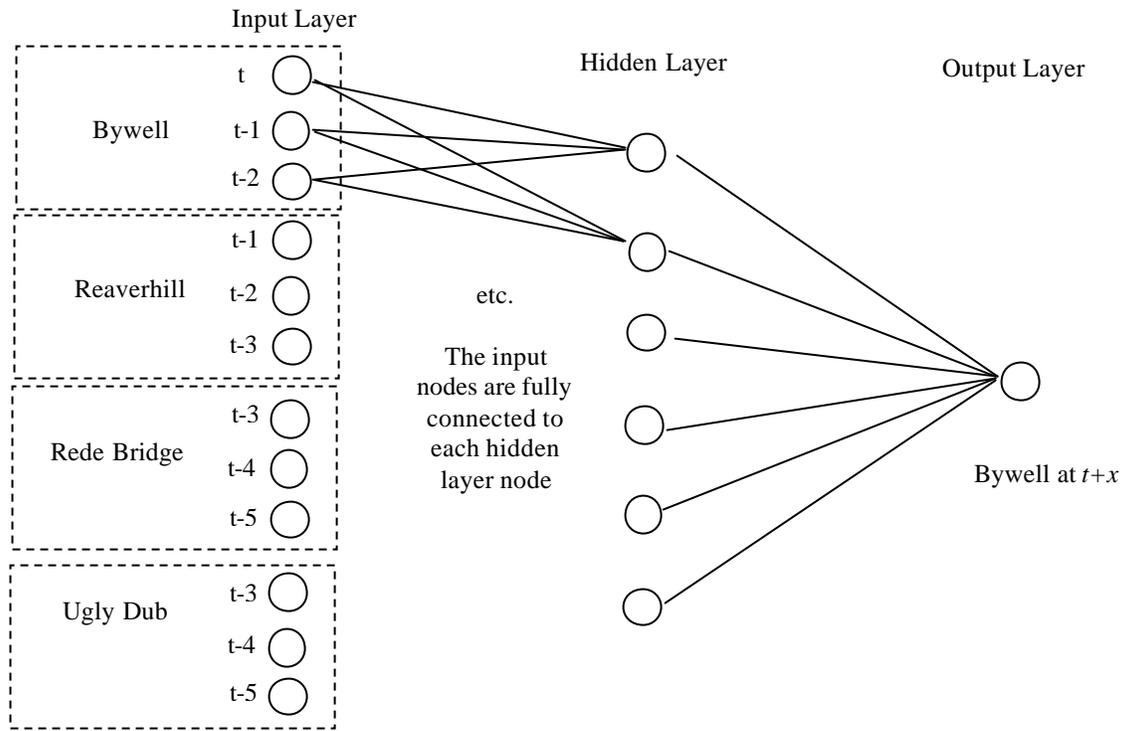A learning function or algorithm is used to adjust the weights of the network during the training phase. There are many learning algorithms for training a MLP but backpropagation is one of the most common (Rumelhart *et al.*, 1986). Backpropagation is based on the concept of steepest gradient descent of the error surface. During the learning period, both the input vector and the target output vector are supplied to the network. The network then generates an error signal based on the difference between the actual output of the network and the target vector. This error signal is then used to adjust the weights of the network appropriately. The error for a hidden processing unit is derived from the error that has been passed back from each processing unit in the next forward layer. This error is weighted using the same connection weights that modified the forward output activation value, and the total error for a hidden unit is thus the weighted sum of the error contributions from each individual unit in the next forward layer. Following training, input data are then passed through the trained network in its non-training mode, where the presented data are transformed within the hidden layers to provide the modelling output values.

Time Delay Neural Networks (TDNNs) are another in the class of feedforward neural networks. They are in many ways similar to MLPs and train using an algorithm similar to backpropagation. TDNNs are designed to detect temporal relationships in a succession of inputs, independent of the absolute time, i.e. they seeks relationships between inputs in their position relative to each other, rather than their position in the data set. In backpropagation, networks with the temporal sequence are converted to a spatial sequence.
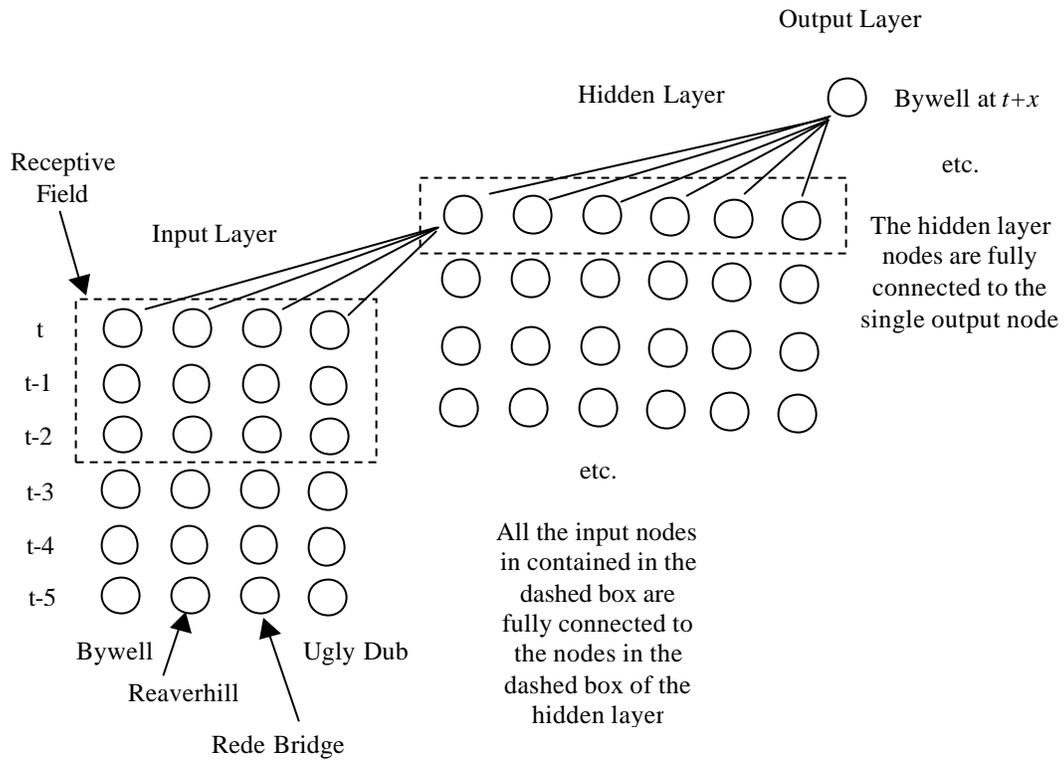
Figure 3a represents a MLP that would be used for predicting the river level at Bywell using inputs from Bywell and 3 upstream stations. The twelve input nodes, which correspond to three different input times for each of the four stations, comprise the input layer to the network. There are six nodes in the hidden layer and one output node in the output layer, which is the prediction at Bywell for $t+x$, where x is the lead time of the forecast. Figure 3b shows a three dimensional representation of a three layer time delay neural network using the same inputs. Compared to the MLP the input nodes are organised in a matrix, where one column represents the input from one station and each row relates to the age of the reading, such that the top row of the input layer is the current value and the sixth row is the sixth previous value.

The top row of the hidden layer of the TDNN behaves in an identical manner to the hidden layer of the MLP. That is, each node in the top row of the hidden layer of the TDNN receives inputs from all the nodes in the 'receptive field' for that row and no other nodes. The 'receptive field' for a particular row of a non-input layer is a region of the previous layer: in this case a 4x3 area of the 4x6 input layer. The receptive field is shown in Figure 3b. Hence there are four possible receptive field positions of the input layer, each of which correspond to a row of the hidden layer. This relationship is made explicit in Table 2.

(a) Conventional feedforward network

Input Layer

Hidden Layer

Output Layer

Bywell
t
t-1
t-2

Reaverhill
t-1
t-2
t-3

Rede Bridge
t-3
t-4
t-5

Ugly Dub
t-3
t-4
t-5

etc.

The input nodes are fully connected to each hidden layer node

Bywell at $t+x$

(b) Time Delay Neural Network

Output Layer

Hidden Layer

Bywell at $t+x$

Receptive Field

Input Layer

etc.

The hidden layer nodes are fully connected to the single output node

t
t-1
t-2
t-3
t-4
t-5

Bywell

Reaverhill

Rede Bridge

Ugly Dub

etc.

All the input nodes in contained in the dashed box are fully connected to the nodes in the dashed box of the hidden layer

Figure 3: (a) Conventional feedforward and (b) time delay network architectures for forecasting river level at Bywell using historical levels from Bywell and 3 upstream stations

Table 2: The relationship of rows of the hidden layer and the corresponding receptive field for the TDNN shown in Figure 3b

Row of Hidden Layer	Row of input layer in corresponding receptive field
1	1 to 3
2	2 to 4
3	3 to 5
4	4 to 6

The choice of the number of rows in the hidden layer and the size of the receptive fields must be made so that there is exact correspondence between the number of rows in the hidden layer and the number of possible positions for receptive fields in the previous layer. The number of rows in a non-input layer is the time delay of the layer. The number of rows in the input layer is the total delay length. Continuing the comparison with the MLP, each row in the hidden layer of the TDNN is equivalent to the status of the hidden layer of the MLP at a recent point in time. (i.e. the top represents the current time and each subsequent row represents the status of the MLP hidden layer one time step earlier). The integration function of the output node can recognise relationships between the different rows of the hidden layer and hence the temporal patterns over the entire delay length time period are considered. The update function for TDNNs is similar to backpropagation but the convergence tends to be slower and hence longer training times are needed.

4.2 Outline of Experimental Runs

A series of feedforward BPNNs and TDNNs were trained on historical data for the winter period of one year (Oct to Apr 1994/95) to forecast levels at Bywell for lead times of 2, 4 and 6 hours ahead. This winter period was chosen because it contains the highest event for the entire 4 year time period. This will ensure that the network has seen the full range of events in the training data set and not be forced to extrapolate to events never seen before. The networks were then validated using independent data from three further winter periods from 1995 to 1998. The neural network model inputs for all the experiments were the levels at Bywell, Reaverhill, Rede Bridge and Ugly Dub. No rainfall information was available. The 15-minute data were averaged to produce hourly values and normalised between 0.1 and 0.9 prior to training. Neural networks were trained using backpropagation with momentum (BPNNs) and backpropagation for a time delay neural network (TDNNs). Training was stopped when the errors in both the training and validation data sets were at a minimum to avoid problems with overfitting of the data.

5. ANN EXPERIMENTS ON THE UPPER TYNE RIVER

Once the networks were trained, the validation data sets were used in predictive mode. The following goodness-of-fit statistics were calculated:

- RMSE (m)
- MAE (m)
- E_1
- d_1
- Difference of averages
- Difference of standard deviations
- False alarms
- RM_FWE (m)
- RM_GWE (m)

These measures were calculated for the entire winter period and for a subset of the data, which included only alarm levels. The exception is the false alarms, which are calculated only for the subset. Table 3 lists the number of observations where alarm levels equalled or exceeded 3.5 m. The goodness-of-fit statistics are provided in Tables 4 to 6 corresponding to lead times of 2, 4 and 6 hours.

Table 3: Number of observations exceeding alarm levels for each winter period

Winter Period	Number of observations
94/95	49
95/96	5
96/97	25
97/98	6

The following observations can be summarised from Tables 4 to 6:

- the results generally show an increase in RMSE, MAE, RM_FWE and FM_GWE as the lead time increases, which is to be expected. The TDNNs generally had lower values for all lead times compared to the BPNNs indicating they are handling the data better. The RM_FWE and RM_GWE are more difficult measures to interpret but they clearly indicate the years in which higher floods occurred because these measures are higher for the winter period of 1995/96.

- E_1 and d_1 also follow the same expected pattern except they decrease as the lead time increases. Similarly, the TDNNs generally had higher values for all lead times compared to the BPNNs.

Table 4: Error measures for a 2 hour lead time. The top value in each row is the statistic for all data in the winter period while the bottom value is the statistic for a subset of the data containing only values exceeding alarm levels.

Measure	BackPropagation				TDNN			
	94/95	95/96	96/97	97/98	94/95	95/96	96/97	97/98
RMSE (m)	0.0631	0.1080	0.1693	0.1248	0.0799	0.1049	0.1228	0.1128
	0.2044	1.2108	1.1878	0.7103	0.2939	0.3487	0.5808	0.4954
MAE (m)	0.0252	0.0497	0.0637	0.0572	0.0494	0.0664	0.0680	0.0675
	0.1354	1.0484	0.8583	0.6603	0.2335	0.2982	0.4418	0.4093
E_l	0.9351	0.7937	0.8004	0.8271	0.8730	0.7248	0.7867	0.7958
d_1	0.9676	0.8980	0.9015	0.9141	0.9355	0.8578	0.8899	0.8935
	0.8947	0.1084	0.2924	0.1157	0.7847	0.3212	0.4856	0.2338
Average Difference	-0.0028	0.0312	0.0262	0.0324	-0.0328	-0.0507	-0.0438	-0.0394
	0.0057	1.0484	0.3645	0.6604	0.1034	0.2093	-0.0078	0.1178
Std Dev Difference	0.0007	0.0240	0.0026	0.0164	0.0158	0.0284	-0.0069	0.0079
	-0.0544	-0.4555	-0.7590	-0.1719	0.1924	-0.1678	-0.3384	-0.4135
False Alarms	0.1020	0.8000	0.6800	1.0000	0.2245	0.4000	0.4800	0.5000
RM_FWE (m)	0.2123	0.2457	0.3256	0.2807	0.2759	0.2497	0.2904	0.2698
	0.7594	2.0157	1.8603	1.5455	1.0614	1.0752	1.3235	1.2132
RM_GWE (m)	0.0578	0.0537	0.0902	0.0658	0.0593	0.0440	0.0678	0.0549
	0.2300	0.6289	0.5952	0.3548	0.2428	0.3363	0.4284	0.3076

Table 5: Error measures for a 4 hour lead time. The top value in each row is the statistic for all data in the winter period while the bottom value is the statistic for a subset of the data containing only values exceeding alarm levels.

Measure	BackPropagation				TDNN			
	94/95	95/96	96/97	97/98	94/95	95/96	96/97	97/98
RMSE (m)	0.1390	0.1424	0.2371	0.1682	0.1269	0.1298	0.1658	0.1411
	0.5973	1.8715	1.8436	1.2420	0.5414	0.6193	0.9920	0.8271
MAE (m)	0.0588	0.0563	0.0885	0.0711	0.0723	0.0701	0.0775	0.0750
	0.3274	1.7645	1.6008	1.1516	0.4634	0.4333	0.7774	0.7551
E_l	0.8487	0.7663	0.7228	0.7853	0.8139	0.7094	0.7569	0.7730
d_1	0.9249	0.8849	0.8623	0.8932	0.9050	0.8519	0.8749	0.8820
	0.7606	0.0673	0.1232	0.0698	0.5976	0.2299	0.2845	0.0776
Average Difference	0.0206	0.0031	0.0014	0.0048	-0.0316	-0.0423	-0.0316	-0.0256
	0.1749	1.764	1.2231	1.1516	0.3156	0.4271	0.3131	0.6627
Std Dev Difference	0.0077	-0.0088	0.0039	0.0382	0.0155	0.0505	-0.0023	0.0209
	-0.1862	-0.4446	-0.8550	-0.3578	0.1364	-0.3735	-0.5421	-0.3715
False Alarms	0.1837	1.0000	0.8400	1.0000	0.3877	0.4000	0.5600	0.8333
RM_FWE (m)	0.3275	0.2839	0.4032	0.3324	0.3469	0.2692	0.3354	0.3050
	1.2146	2.6160	2.5374	2.0422	1.4674	1.2851	1.7662	1.6555
RM_GWE (m)	0.0868	0.0628	0.1042	0.0755	0.0749	0.0536	0.0845	0.0640
	0.3832	0.7711	0.7002	0.4402	0.3648	0.4756	0.5616	0.3641

Table 6: Error measures for a 6 hour lead time. The top value in each row is the statistic for all data in the winter period while the bottom value is the statistic for a subset of the data containing only values exceeding alarm levels.

Measure	BackPropagation				TDNN			
	94/95	95/96	96/97	97/98	94/95	95/96	96/97	97/98
RMSE (m)	0.2121	0.1644	0.2550	0.1933	0.1816	0.1356	0.2255	0.1677
	1.0260	1.7930	1.6575	1.2391	0.8485	1.6006	1.6067	1.1278
MAE (m)	0.1297	0.0913	0.1217	0.1028	0.0916	0.0772	0.1070	0.0938
	0.6211	1.7782	1.4203	1.1930	0.7227	1.4888	1.3971	1.0014
E_l	0.6659	0.6213	0.6188	0.6894	0.7641	0.6798	0.6644	0.7163
d_1	0.8279	0.8076	0.8053	0.8389	0.8791	0.8321	0.8263	0.8517
	0.5844	0.0669	0.1754	0.0675	0.4439	0.0788	0.9738	0.0673
Average Difference	-0.0581	-0.0330	-0.0443	-0.0219	-0.0165	-0.0439	-0.0410	-0.0391
	0.5351	1.7782	0.9923	1.1930	0.6125	1.4887	-0.7727	0.9873
Std Dev Difference	0.0497	0.0665	0.0147	0.0579	0.0304	0.0136	0.0048	0.0414
	-0.2560	-0.0128	-0.8463	-0.2604	0.1193	-0.4207	-0.7727	-0.4217
False Alarms	0.4694	1.000	0.9200	1.000	0.6939	1.0000	0.8800	0.8300
RM_FWE (m)	0.4579	0.3297	0.4387	0.3755	0.4102	0.3010	0.4144	0.3518
	1.7082	2.6240	2.3977	2.0752	1.8391	2.4024	2.3697	1.9072
RM_GWE (m)	0.0995	0.0600	0.1029	0.0763	0.0938	0.0605	0.1012	0.0735
	0.4942	0.7382	0.6715	0.4523	0.4487	0.7181	0.6693	0.4138

- the average differences are very small for the entire data set and the negative sign shows that the TDNNs tended to overpredict for the entire data set. The BPNNs tended to underpredict on average for lead times of 2 and 4 hours and then overpredict as the lead time increased. For the alarm subset, both the TDNNs and the BPNNs tended to underpredict. This provides a good source of extra information above the other global measures.

- the difference in standard deviations indicates a difference in the variation of predictions relative to the observed data. The results generally indicated a slightly smaller variation overall but a higher one for the alarm data set. The TDNNs predicted values with variations closer to the observed data relative to the BPNNs.

- the false alarm rate increased with lead time and the TDNN models generally had a lower rate of false alarms than the BPNNs.

The combination of evaluation measures clearly highlights the difference between performance for the entire winter period and that for the alarm data set. Unfortunately the alarm data set is very small for Bywell so the networks were heavily biased towards low flow events. Therefore, it is not surprising that poor absolute results were achieved for both forecasting models in an operational context. Interestingly, better results have been achieved when predicting stage at Bywell using the South Tyne River stations only (Kneale *et al.*, 2000), which raises the question as to whether Bywell is mainly responsive to the South Tyne flows. If this were true, then the stations on the North Tyne could not be used to predict Bywell successfully in the hypothetical event that the South Tyne stations fail. Alternatively, should

data sets be sampled to contain only flood hydrograph events and used in the training and independent validation process? In theory these will provide more meaningful global goodness-of-fit statistics. Finally, should a better sampling scheme such as bootstrapping be used in the training process (Abrahart, 2001)? Each of these issues will be investigated in the next stage of the research. However, the TDNN models did show a better performance overall and therefore hold some promise as to another type of neural network that could be investigated for rainfall-runoff modelling.

6. CONCLUSIONS

This paper has presented a set of evaluation measures for the purpose of comparing results from forecasting models including a combination of global and flood specific goodness-of-fit measures. Backpropagation and Time Delay Neural Networks were trained to forecast stage on the River Tyne, Northumbria using only upstream stations on the north branch. The networks were trained on a continuous data set for one winter period and validated on winter periods for three other years. Models were developed to forecast lead times ranging from 2 to 6 hours. The combination of measures provided a good picture of the forecasting performance of the models. However, the establishment of a consistent set of operational evaluation measures and benchmark data sets should be undertaken by researchers and operational staff. If there could be some general agreement on error measures and a set of benchmark data sets supplied with these error measures, it would facilitate the comparison of the neural network and physical models in the future. This will go a long way towards addressing a major criticism of neural network models in the hydrological literature, i.e. the lack of

comparison with existing operational forecasting systems.

Clearly there are a very limited number of alarm level events and increasing the data set should improve the absolute results. Nevertheless these experiments with limited data replicate situations where new stations are established. Issues regarding the creation of appropriate training data sets and sampling procedures as well as the hydrological conditions of the River Tyne need to be investigated further. The results also showed that, in general, the TDNNs perform better than the BPNNs and may prove a profitable algorithm for hydrological investigations.

REFERENCES

Abrahart, R.J. (2001) "Forecasting comparison between split-sample validation and single model bootstrap neural network rainfall-runoff modelling". To be presented at GeoComputation 2001, Brisbane Australia, 24-26 September 2001.

Abrahart, R.J. and Kneale, P.E. (1997) "Exploring neural network rainfall-runoff modelling", *Proceedings of the Sixth National Hydrology Symposium,* University of Salford, 9.35-9.44.

Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press: Oxford.

Garrick, M., Cunnane, C. and Nash J.E. 1978. "A criterion of efficiency for rainfall-runoff models", *Journal of Hydrology*, vol. 36, pp. 375-381.

Han, D., Cluckie, I.D., Wedgwood, O. and Pearse, I. (1997) "The North West Flood Forecasting System (WRIP North West)", BHS Sixth National Hydrological Symposium, University of Salford.

Kasabov N.K. (1996) *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering.* MIT Press: Cambridge, Massachusetts.

Khonder, M.U.H., Wilson, G. and Klinting, A. (1998) "Application of neural networks in real time flash flood forecasting". In Babovic, V. and Larsen, C.L. (eds.) *Proceedings of the Third International Conference on HydroInformatics,* A.A.Balkema: Rotterdam, pp. 777-782.

Kneale, P. and See, L. (1999) *Developing a Neural Network for Flood Forecasting in the Northumbria Area of the North East Region, Environment Agency. Final Report.* University of Leeds: Leeds.

Kneale, P.E., See, L., Cameron, D., Kerr, P. and Merrix, R. (2000) "Using a prototype neural net forecasting model for flood predictions on the Rivers Tyne and Wear". British Hydrological Society, 7th National Hydrology Symposium.

Legates, D.R. and McCabe, G.J. (1999) "Evaluating the use the "goodness-of-fit" measure in hydrologic and hydroclimatic model validation". *Water Resources Research*. vol. 35, pp. 233-241.

Lin, C.T. and Lee, C.S.G. (1996) *Neural Fuzzy Systems – A Neuro-Fuzzy Synergism to Intelligent Systems.* Prentice Hall: Upper Saddle River

Marshall, C. (1996) *Evaluation of Integrated Flood Forecasting Systems*, R&D Technical Report W17, Environment Agency, Bristol.

Minns, A.W. and Hall, M.J. (1996) "Artificial neutral networks as rainfall-runoff models", *Hydrological Sciences Journal*, vol. 41, pp. 399-417.

Moore, R.J., Jones, D.A., Black, K.B., Austin, R.M., Carrington, D.S., Tinnion, M. and Akhondi, A. (1994) "RFFS and HYRAD: Integrated systems for rainfall and river flow forecasting in real-time and their application in Yorkshire", *BHS Occasional Paper* 4, University of Salford, Salford.

Nash, J.E. and Sutcliffe, J.V. (1970) "River flow forecasting through conceptual models, I, A discussion of principles", *Journal of Hydrology*, vol. 10, pp. 282-290.

Nemec, J. (1986) *Hydrological Forecasting: Design and Operation of Hydrological Forecasting Systems*, D. Reidel Publishing Company: Dordrecht.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) "Learning internal representation by error propagation". In D.E. Rumelhart, J.L. McClelland (eds.) Parallel Distributed Processing: *Explorations in the Microstructure of Cognition*, Vol.1. MIT Press: Cambridge MA, pp. 318-362.

Sharkey, A.J.C. (1999) *Combining Artificial Neural Networks: Ensemble and Modular Multi-Net Systems*. Springer-Verlag: London.

Shepherd, A.J. (1997) *Second-Order Methods for Neural Networks*. Springer-Verlag Limited: London.

Smith A. (2000) *A comparison of neural network designs with regards their capability for river level forecasting, for the Tyne River, NE England.* Unpublished MSc Dissertation. School of Geography, University of Leeds: Leeds.

Smith, J. and Eli, R.N. (1995) "Neural network models of rainfall-runoff process", *Journal of Water Resources Planning and Management*, vol.121, pp. 499-509.

Waibel, A. (1989) "Modular construction of time-delay neural networks for speech recognition", *Neural Computing*, vol.1, pp. 39-36.

Willmott, C.J. (1981) "On the validation of models", *Physical Geography*, vol.2, pp. 184-194.