A Comparative Study of Artificial Neural Network Techniques for River Stage Forecasting

C.W. Dawson[1], L. See[2], R.J. Abrahart[3], R.L. Wilby[4], A.Y. Shamseldin[5], F. Anctil[6], A.N. Belbachir[7], G. Bowden[8], G. Dandy[8], N. Lauzon[6], H. Maier[8]
[1]Department of Computer Science, Loughborough University, UK
[2]School of Geography, University of Leeds, UK
[3]School of Geography, University of Nottingham, UK
[4]Environment Agency, Trentside Office, Nottingham, UK
[5] Department of Civil and Environmental Engineering, University of Auckland, Private Bag 92019, Auckland, NZ
[6]Department of Civil Engineering, Université Laval, Canada
[7]PRIP, Vienna University of Technology, Austria
[8]School of Civil & Environmental Engineering, University of Adelaide, Australia

***Abstract*--- Although artificial neural networks have been applied to problems within hydrology for over ten years, there is little consensus on the 'best' type of neural network model to use and the most effective means of training the chosen model. In order to explore the different approaches neural network modellers use to forecasting river stage, an international comparison study was undertaken during 2004.**

This research was based on a set of rainfall and river stage data covering three winter periods for an unidentified river basin in England (with a catchment of 331,500 Ha in the north of the country), sampled at 15 minute intervals. Several neural network enthusiasts took part in the study from a number of different countries. The preferred methodologies and forecasting outputs from a number of 'blind' models of river stage developed by the participants have been collated and are presented in this paper.

I. INTRODUCTION

Artificial neural networks (ANNs) have been applied to the problem of hydrological modelling since the early 1990s and there are now over 200 papers published in this field. Despite this extensive literature base, a common set of operational methodologies has still to emerge, although some attempts have been made to define one (Dawson and Wilby, 2001). In addition, the extensive range of different types of network, training algorithms and software tools available, means that a standard implementation of this kind of model has not emerged and the application of these models in real time is still awaited.

In order to explore and evaluate the approaches that different *neurohydrologists* employ, an inter-comparison exercise was established. This exercise involved the dissemination of a benchmark catchment data set to a number of participants. Each was given the freedom to develop two ANN models for three forecast horizons: t+8, t+16 and t+24 hours ahead. In this paper the results of the t+24 hour ahead forecasts are presented and discussed as this is the most challenging of the three lead times.

Participants in this exercise were provided with background information about the unknown catchment and a standard rainfall-runoff data set (split into two calibration sets and one test data set). Each participant was asked to submit two modelling solutions for each lead time, one that used upstream data as predictors and one that did not. The participants were free to develop any neural network model using any appropriate training algorithm, and to use the two calibration data sets provided in whatever manner was considered prudent (for example, using one for training and one for validation).

This paper discusses the motivation and background to this investigation, it provides further particulars on the design of the experiment, and concludes with a discussion of the results and some suggestions for further research. This paper is also an extension of work from a similar study undertaken during 2003, which was based on an experimental catchment.

II. CATCHMENT DESCRIPTION

Table I provides important hydrological statistics of the catchment in northern England that formed the basis of this exercise. These were the only data provided to the participants to ensure that the modelling was undertaken 'blindly'. In other words, none of the participants were disadvantaged through lack of first hand knowledge of the catchment.

III. BENCHMARK DATA SETS

Three data sets consisting of 15 minute data were made available in this study. Two data sets were provided for calibrating the network models, covering the periods 1 October 1993 to 31 March 1994 (containing 17,472 data points) and 1 October 1995 to 31 March 1996 (containing

17,568 data points). A test set was provided covering the period 1 October 1994 - 31 March 1995 (containing 17,472 data points). Each of the data sets contained river stage data (m) at three upstream sites, rainfall data at five catchment rain gauges (mm) and stage data (m) at the target site. The peak stage in the two training sets was 5.04m and 4.13m, while the peak stage in the test set was 5.78m (thus providing a reasonable test of modelling skill).

TABLE I
CASE STUDY CATCHMENT DESCRIPTORS

Catchment area (Ha)	331500
Elevation (metres)	10-710
Geology	Mixed: Carboniferous Limestone and Millstone Grit to the west [headwaters]; Permo-Triassic rocks to the east [basin out].
Soils	Peats and stagnogley soils in the uplands; stagnogleys, sandy gley soils and brown earths in the lowlands.
Land–use	Land use reflects both topographic and precipitation influences: moorland (24%) and grassland (33%) predominate to the west; tilled land (31%) to the east; 4% of the catchment is woodland; 5% could be classed as urban or suburban.
Annual rainfall (mm)	906 mm [1969 - 1990]
Annual runoff (mm)	464 mm [1969 - 1990]
Runoff (%)	51
Drainage	Mixed: three major sub-catchments are involved. Minor baseflow component: baseflow index is 0.43.

It was left to the participants to decide how to use the two calibration data sets; for example, they could use both sets for training their models or use one set for training and one set for validation (i.e. selecting the 'best' model). The participants were also free to decide how to pre-process the data into appropriate predictors for the ANN models – for example, identifying strong antecedent correlations, using rainfall averages and/or moving averages where appropriate.

The participants were asked to produce two models for the t+24 hour ahead forecasts – one that could use upstream data (pre-processed as required) with at least a 24 hour lead time and one that could not (i.e. it could only use antecedent rainfall pre-processed as required). Both models could use antecedent river stage at the target site as an additional predictor providing at least a 24 hour lead time was used.

The data sets also 'included' a number of missing values for the rain gauges and the upstream sites. In addition, some of the rainfall data appeared to be somewhat inaccurate – for example, one gauge recorded a rainfall of 62.4mm in 15 minutes (the UK's record for 15 minutes is 50mm). The participants were left to deal with these missing and inaccurate data as they felt appropriate.

IV. EXPERIMENTS

Of the seventeen participants originally contacted to take part in this study, five produced models using the benchmark data sets. Table II summarizes the different approaches used by the participants in this study. The table shows the variation in software employed - from off-the-shelf packages, such as the Neural Network Toolbox for MATLAB, to software written by the participant themselves in Pascal. Networks used included the common Multi Layer Perceptron (MLP), self organising maps and an ARMAX model (auto regressive moving average with exogenous inputs) for comparison. Data were either normalised or standardised to [0.1, 0.9].

The decision as to when to terminate training was based on cross validation with the second training set or after a certain number of epochs had been performed.

A number of approaches were used to pre-process the data and identify suitable predictors for the models. The simplest predictors were unadjusted upstream flow, unadjusted rainfall at each site, and antecedent flow at the target site. More sophisticated predictors included moving averages of rainfall, averaging rainfall across all rain gauges and calculating appropriate lags for each of the predictors identified. In order to speed up the training process one participant reduced the first training set from 17,472 data points to 1,000 data points by selecting every sixteenth data point from the data set. This did not lead to any reduction in model performance and speeded up training time significantly. Finally, a cross-section of training algorithms was employed including Backpropogation, Bayesian regularization and SOM-batch training.

It is noted that no comparisons have been made with physical or conceptual rainfall-runoff models. The purpose of this study was *not* to compare results with other approaches but to compare alternative neural modelling approaches with one another and a standard statistical approach.

TABLE II
SUMMARY OF DIFFERENT APPROACHES USED IN THE STUDY

Software Used	Own software on Matlab, Matlab with Neural Network Toolbox, own software (Pascal), SOM Toolbox 2.
Network Types	Self organizing map – multiple linear regression, MLP.
Activation Functions	Sigmoid.
Normalisation / Standardisation	Normalisation, [0.1, 0.9].
Stopping Criteria	Number of epochs, validation set, minimise SSE.
Predictors	Mean of all rain gauges, moving average of each rain gauge, Q, upstream flow, sum of rainfall over time, various lagged rainfall and upstream flow.
Training Algorithms	Bayesian regularization, BP, SOM-batch training algorithm.

V. ERROR MEASURES

There is a general lack of consistency in the way that rainfall-runoff models are assessed or compared (Legates and McCabe, 1999) and one should not rely on individual error measures when assessing ANN model performance (Dawson and Wilby, 2001). Because of these considerations a number of *complementary* error measures have been used in this study including:

- RMSE (Root Mean Squared Error), which is used in many studies and provides a good measure of fit at high flows (Karunanithi et al., 1994).
- CE (Coefficient of Efficiency) and r^2 (r-squared), which are independent of the scale of data used. According to Shamseldin (1997) a CE value above 0.9 is 'very satisfactory', a value between 0.8 and 0.9 is 'fairly good' while a value below 0.8 is 'unsatisfactory'.
- MAE (Mean Absolute Error) which is not weighted towards high flow events and provides an indication of overall accuracy.
- SE (Standard Error) which provides a measure of the spread of the errors produced by the model (calculated as the standard deviation of the errors).

VI. RESULTS

The results of the experiments are presented in Tables 3 and 4. The participants are represented in these tables as A, B, C, D and E. The two models each participant produced (one using upstream data and one that did not) are represented by Xa and Xb respectively. Table 3 summarizes the model structures produced by the participants which were evaluated against the test data set. The data points listed in Table 3 represent the amount of data available for testing each participant's model once lags, moving averages etc. had been calculated. The x in the structure indicates that this parameter is unknown.

Table 4 provides the summary statistics for all the models when assessed against the test data set. As an example of the results produced, Fig. 1 shows the observed test data, in this case plotted with the results of the ANN model produced by participant D (using upstream data in the model).

TABLE III
SUMMARY OF MODELS

Model	Structure	Data points
Aa	4-4-1	17313
Ab	3-10-1	17313
Ba	12-x-1	17007
Bb	7-x-1	17007
Ca	x-x-1	17376
Cb	x-x-1	17376
Da	8-10-1	17357
Db	7-20-1	17357
Ea	2-4-1	17280
Eb	6-6-1	17279

TABLE IV
RESULTS OF MODELS DURING TESTING

Model	RMSE (m)	CE	MAE (m)	r^2	SE (m)
Aa	0.4681	81.98	0.3129	0.9063	0.4673
Ba	0.5329	76.78	0.3702	0.8770	0.5329
Ca	0.3111	92.06	0.2347	0.9768	0.2859
Da	0.5164	78.11	0.3265	0.8938	0.5065
Ea	0.5324	76.66	0.4132	0.7922	0.5071
Ab	0.4958	79.79	0.3246	0.8952	0.4952
Bb	0.5074	78.95	0.3433	0.8947	0.5037
Cb	0.9461	25.59	0.7686	0.9152	0.5516
Db	0.5763	72.74	0.3733	0.8649	0.5656
Eb	0.5399	76.00	0.3460	0.7729	0.5303
Summary					
Min:	0.3111	25.59	0.2347	0.7729	0.2859
Max:	0.9461	92.06	0.7686	0.9768	0.5656

We begin first by looking at those models that included upstream data as predictors (Xa). According to all statistics the model produced by participant C was the most accurate when assessed using the test data set. This model was the statistical ARMAX model that was used for comparative purposes. The models produced by participants B and E were the least accurate depending upon which test statistic is considered. Model Ba contained the most predictors of all the models produced – using lagged unadjusted data from all five rain gauges, and unadjusted flow data from the upstream

sites (at different lag periods). This shows that increased complexity does not necessarily provide the most accurate model and parsimonious models such as Ca can often prove to be more accurate and efficient.

For those models that did not include upstream data as predictors (Xb), the results are less conclusive. For example, the model produced by participant C in this case is the least accurate according to the CE statistic (25.59%) although according to the r-squared statistic, it is the most accurate (0.9152). Model Cb has the largest RMSE and MAE (0.9461m and 0.7686m respectively) while model Db has the largest SE statistic (0.5656m).

The second set of results highlight the problems of using single error measures for assessing the accuracy of models. For example, using the CE statistic one might conclude that model Cb is particularly inaccurate, whereas when using the r-squared statistic one might conclude that this is the strongest model. Fig. 2 highlights why this model produced these seemingly contradictory results. Although the model follows the general shape of the hydrograph quite well (hence the r-squared statistic is reasonably 'high') it does so at a much lower level (and thus statistics such as MAE and RMSE are 'poor'). If one were to implement this model for real time flood forecasting one would need to be aware that while the model might be modelling general changes in flow accurately at t+24 hours ahead, it would be generally underestimate the flow magnitude.

VII. CONCLUSIONS

This study has enhanced collaboration between scientists in this promising field of research and the results, like many other studies before, show the potential benefits of neural network rainfall-runoff models.

Although only a limited number of participants took part in the study another project of this nature is currently been undertaken with a simplified benchmark data set (missing data and anomalies have been excluded). Those wishing to take part in a follow-up study, or with benchmark data that could be used, should contact Dr C.W. Dawson via email at C.W.Dawson1@lboro.ac.uk.

REFERENCES

[1] C.W. Dawson, and R.L. Wilby, "Hydrological modelling using artificial neural networks", *Progress in Physical Geography*, vol. 25 (1), pp. 80–108, 2001.

[2] N. Karunanithi, W.J. Grenney, D. Whitley, and K. Bovee, "Neural networks for river flow prediction", *Journal of Computing in Civil Engineering*, vol. 8, pp. 201–220, 1994.

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

[4] A.Y. Shamseldin, "Application of a neural network technique to rainfall-runoff modelling", *Journal of Hydrology*, vol. 199, pp. 272–294, 1997.

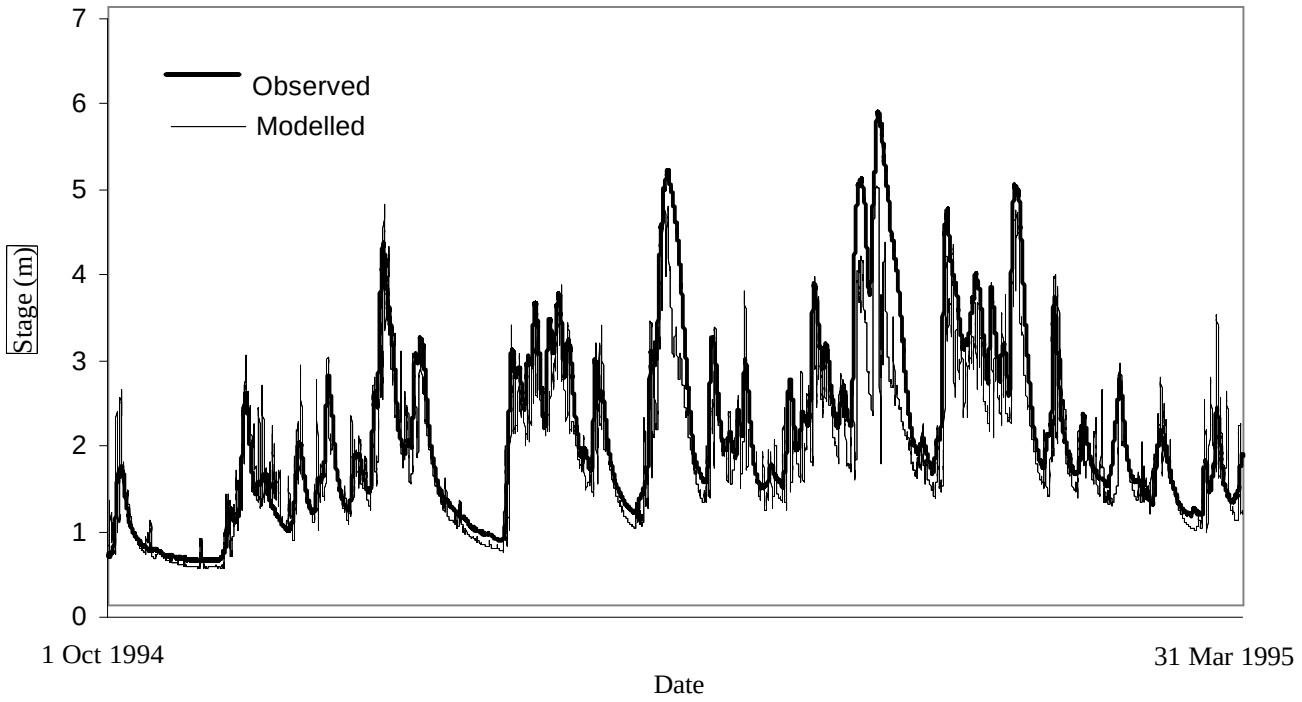

Fig. 1. Model Da during testing

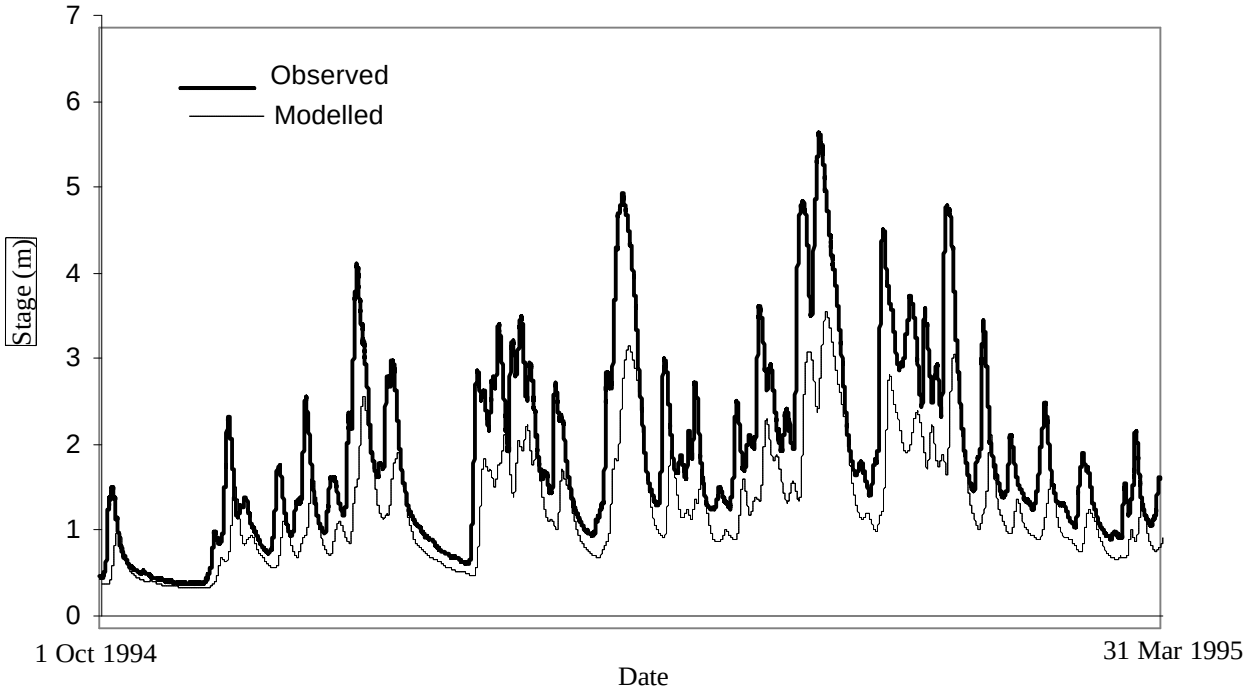

Fig. 2. Model Cb during testing