19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Long term flow forecasting for water resources planning in a river basin

C. Sivapragasam [a], N. Muttil [b] and V.M. Arun [a]

[a] *Department of Civil Engineering, Kalasalingam University, Krishnankoil, Tamil Nadu, India*
[b] *School of Engineering and Science, Victoria University, PO Box 14428, Melbourne, VIC 8001, Australia*
Email: nitin.muttil@vu.edu.au

Abstract: Although lot of work have been reported in forecast of inflows to reservoir, yet each basin needs unique treatment to improve the forecast accuracy, and as such it is very difficult to generalize the inflow modeling, particularly for monthly flow models. Further, for planning of water resources in a river basin, long term flow forecasts are more important. In this work, an approach is suggested to improve the forecasting of monthly inflows to the Manimuthar reservoir in the Tamarabarani river basin in Tamil Nadu, India. The releases from Manimuthar reservoir act as a support to the Tamarabarani river system which plays a vital role in agricultural production of the region. For the agricultural planning, it is desired to have monthly prediction, and hence an attempt has been made herewith towards this.

The first step towards this involved a careful selection of the input variables. Antecedent inflows from previous years for the same month form a part of the input vector. Further, rainfall information is also included in the input vector, wherein such information is first estimated in those crucial locations along the river reach or the surrounding region which plays an important role in deciding the inflow to reservoir. This rainfall information is estimated using a Kriging based methodology (based on the Kriging standard deviations), which was presented in a previous study (Sivapragasam et al., 2011).

Genetic Programming (GP), an evolutionary algorithm based data-driven modelling technique is chosen as the modeling tool for this study. As compared to the traditional data-driven techniques like artificial neural networks (ANN), GP has unique advantages in that it does not assume any functional form of the solution. For instance, in ANNs, the network needs to be defined initially and then the coefficients (weights of the ANN) will found by the learning algorithm. In contrast, in GP, the learning method finds both the form of the model and the coefficients that fit the problem well.

The results indicate significant improvement in inflow forecast accuracy, especially when the rainfall values from additional stations around the Manimuthar reservoir region are included in the input vector. Based on the results demonstrated for the forecasting of inflows for the Manimuthar reservoir, the proposed approach appears to be quite innovative and has potential for improving monthly flow forecasts.

Keywords: *Flow forecasting, water resources planning, Genetic Programming (GP)*

1. INTRODUCTION

Due to the importance of hydrologic forecasting, a considerable number of forecasting models and methodologies have been developed and applied in inflow forecasting which can be categorized as process-driven methods and data-driven methods. The process-based modeling approach is a knowledge-driven modeling process that explains the underlying process. Various forms of rainfall-runoff models such as lumped, semi distributed and distributed, and snowmelt-runoff models are in this category. Data driven models are based on a limited knowledge of the physics of the watershed system and they depend on data describing input and output characteristics. They are essentially black-box models that characterize the relationships between inputs and outputs without a consideration of the details or explicit simulation of the underlying physical process.

In one of the earliest studies, Karunanithi et al. (1994) used a cascade correlation algorithm to predict the flow at a location in a river by using the flow data at different locations along the river and along its tributaries, as input. A five previous day window of each input station is used to account for the time dependence of the phenomenon. Their model performed better than the commonly used power model. Many more such studies are reported by various researchers (Coulibaly et al, 2000; Thirumalaiah and Deo 2000; Kisi, 2008a). Although monthly forecasts studies have been relatively less, in the recent past, few studies have been undertaken specifically for monthly river flow forecasting (Kisi, 2008b, Firat and Turan, 2010, Singh et al., 2011).

It has to be noted that though lot of works have been reported to forecast inflows to reservoir, yet each basin/catchment needs unique treatment to improve the forecast accuracy, and as such it is very difficult to generalize the inflow modeling. No single forecasting model is powerful and general enough to outperform others for all types of catchments and under all circumstances or even one catchment with different behavioral phases (Shamseldin, 2004). Selection of appropriate input parameters in the model and also the design of appropriate methodology are very crucial.

In order to facilitate decision making process, an attempt is made in this study to forecast monthly inflow to Manimuthar reservoir in the Tamarabarani basin (India) using Genetic Programming (GP) as a data driven process. This is primarily because of GP's ability to select appropriate input variables for the model (parameters which most appropriately describes the process) in the process of mapping a complex non-linear input-output relationship between variables.

GP has found many applications in hydraulics and water resources field, including forecasting. Harris et al (2003) used GP to develop mathematic models to study the effect of vegetation on velocity distribution across a channel through experimental studies in the laboratory flume. They emphasized using data together with its dimensions rather than using traditional method of dimensionless data. Giustolisi (2004) applied GP used to determine Chezy's resistance coefficient for full circular corrugated channels and stressed upon the scientific discovery capability of GP. Muttil and Lee (2005) presented a "real-time" modeling of algal blooms at a monitoring station in a costal area located in Hong Kong, China using GP. Several studies based on use of GP as a rainfall-runoff model builder have also been carried out (Babovic and Keijzer, 2000; Khu et al., 2001; Liong et al., 2001).

This paper first presents details of the study area and data used for the analysis in Section 2, which is followed by a brief description and advantages of GP in Section 3. The details of the modelling that has been carried out is presented in Section 4 and Section 5 presents the results and discussion. Finally, conclusions are drawn in Section 6.

2. STUDY AREA AND DATA USED

Tamarabarani River basin is one of the important basins in Southern Tamil Nadu (India) with rich surface water resource. Tamarabarani River has its origin in Western Ghats near Senkottai district of Tamil Nadu State with a catchment area of about 6000 km^2. In the development of Tamarabarani River basin, Manimuthar River plays a crucial role as it is one of the major tributaries, particularly so for catering to the agricultural need. As such, for agricultural planning, it is desired to have monthly forecast models.

The monthly inflow data of Manimuthar reservoir (for a period from 1970 to 1995) was obtained from Public Works Department, Thirunelveli District, Tamil Nadu. The monthly averages and standard deviation for the period of 1970 - 1995 are plotted in Figure 1. It can be inferred that

(a) Inflow peaks are obtained almost during fixed time period every year (during north – east monsoon)

(b) The standard deviation of average monthly rainfall is quite high implying difficulty in accurate prediction of inflow.

Further, monthly precipitation data is available for the above mentioned period for the rain gauge located near Manimuthar dam. There are 15 more rain gauge stations spread throughout the Tamarabarani basin where similar rainfall information are available. A detailed work has been carried out using this rainfall information to suggest a new network design for the basin (Sivapragasam et al., 2011).

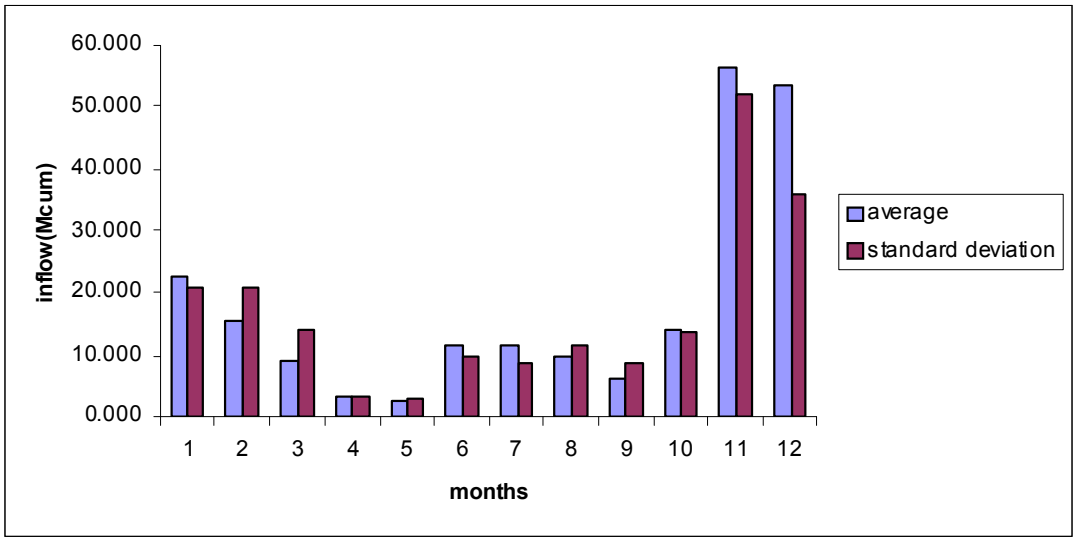

Figure 1. Monthly average and standard deviation for Manimuthar inflow.

3. GENETIC PROGRAMMING

The basic search strategy behind GP (Koza, 1992) is a genetic algorithm. It differs from this traditional genetic algorithm in that it typically operates on parse trees instead of bit strings. A parse tree is built up from a "terminal set" (the variables in the problem) and a "function set". Suppose the terminal set consists of a single variable x and some constants, and the function set consist of the operators for multiplication, division, addition and subtraction, the space of available parse trees constitute all polynomials of any form over x and the constants. An example of such a parse tree can be found in Figure 2, with a "tree size" of 3. Tree size is the maximum "node depth" of a tree, where "node depth" is the minimal number of nodes that must be traversed to get from the "root node" of the tree (see Figure 2) to the selected node.

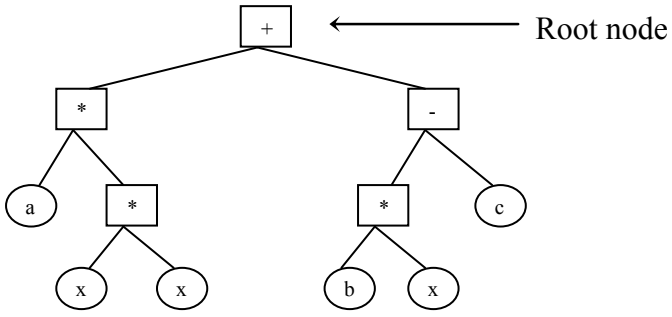

Figure 2. A parse tree representing a*x*x + b*x - c

As a genetic algorithm, GP proceeds by initially generating a population of random parse trees, calculate their fitness – a measure of how well they solve the given problem – and subsequently selects the better parse trees for reproduction and recombination to form a new population. This process of selection and reproduction iterates until some stopping criterion is satisfied. The recombination takes place by crossover

and mutation. For a detailed description of genetic programming from a water resources perspective, the interested reader is referred to Muttil and Lee (2005), Khu et al. (2001) and Babovic and Keijzer (2000).

What makes GP unique compared to the traditional data-driven methods is that it does not assume any functional form of the solution. In for instance regression, the model to use is fixed at the onset, and the regression method will subsequently find the coefficients. For artificial neural networks, the network needs to be defined and then the coefficients (weights) will found by the learning algorithm. In GP in contrast, the initial building blocks (terminals and functions) are defined, and the learning method will subsequently find both the form of the model and the coefficients that fit the problem well (Muttil and Lee, 2005).

In this study, GPKernel, developed by DHI Water and Environment (Babovic and Keijzer, 2000) is used for implementing GP. GPKernel is a command line based tool for finding functions on data. For a detailed explanation of various features of GPKernel, the reader is referred to Babovic and Keijzer (2000).

4. MODEL DEVELOPMENT

The validation was done on last one year of data i.e. for 1995. Remaining years are used for training and testing.

The forecast performance is evaluated using the root mean square error (RMSE) goodness-of-fit measure, as presented in Eqn (1) below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\left(X_m\right)_i - \left(X_s\right)_i\right]^2} \tag{1}$$

where X is any variable that is being forecasted; the subscripts *m* and *s* represent the measured and forecasted values and *n* is the total number of training records.

4.1. One Month Lead Time Forecast

For one lead time monthly forecast, it is most appropriate to use the direct forecast approach or single model approach i.e. the output to the GP model is the one lead forecast directly. Most researchers adopt this method in their studies. Typically, the model can be functionally represented as:

$$Q_{t+n} \equiv f\left(Q_t, Q_{t-1}, R_t\right) \tag{2}$$

where Q_{t+n} is n^{th} month ahead inflow (and n=1 for one month ahead), Q_t is the inflow during current period and R_t is the rainfall during current period.

In this study, the following 3 different models are studied:

- Model 1: Monthly forecast with only antecedent inflow. This model will have the current inflow as the only antecedent inflow input.
- Model 2: Monthly forecast with antecedent inflow and rainfall. This model will contain both current inflow and current rainfall (at Manimuthar reservoir).
- Model 3: Monthly forecast with antecedent inflow and rainfalls from additional stations. This model will contain current inflow, and rainfall information from ten additional imaginary stations around the Manimuthar region. The rainfall from 'imaginary stations' refer to rainfall information estimated based on Kriging analysis (on existing rain gauges) in locations where physically no rain gauges are there as reported in a previous study (Sivapragasam et al., 2011). These locations are selected based on site visit and discussion with the site engineers. The estimated rainfall for the current time period is included in the model to improve the inflow forecast. As rainfall events are highly unlikely to be uniform, and since the reservoir catchment is characterized by only one rain gauge, it is desired to use estimated rainfall information in regions adjoining the catchment which might provide additional information for improving the forecast. The model can be functionally represented as:

$$Q_{t+1} = f\left(Q_t, R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8, R_9, R_{10}\right) \tag{3}$$

  In the training of GP, only the basic arithmetic functions are used in the GP function set.

4.2. Higher Lead Time Forecast

For higher lead times, it was decided to go for individual monthly models as it is highly unlikely to improve the results by a single model. In fact, the preliminary results did indicate poor prediction when such models were adopted. After a discussion with the dam site engineer, it is proposed to construct model based on historical information of inflow during the same month in the previous years. This is because, at least as far as Manimuthar reservoir is concerned, the inflows during the same months in previous years can be assumed to be closely related as the catchment didn't undergo any major changes in the past.

In this study, the inflows during previous three years during the same month were found to be most appropriate. The model can be functionally represented as:

$$Q_{y+1} = fn(Q_y, Q_{y-1}, Q_{y-2}) \quad (4)$$

where Q_y is the inflow for a given month in y^{th} year, Q_{y-1} is the inflow for the same month in $(y-1)^{th}$ year and so on. Q_{y+1} is the inflow for the same month during next year.

When rainfall information is included, it includes rainfall at Manimuthar station during a given month in the current year and also rainfall information from the selected imaginary locations during the same period. The model can then be functionally represented as shown in Eqn (5) below:

$$Q_{y+1} = fn(Q_y, Q_{y-1}, Q_{y-2}, R_y, R_1, \ldots, R_{10}) \quad (5)$$

In the training of GP, besides the basic arithmetic functions (+, -, x, /), some trigonometric functions were also introduced in the function set to capture the effect of possible non-linearity.

5. RESULTS AND DISCUSSION

5.1. One Month Lead Time Forecast

The results in terms of goodness-of-fit measures are presented in Table 1.

Table 1. Forecast results for one month lead time

Name	Training	Validation	GP equations
	RMSE	RMSE	
Model 1	19.759	39.318	$Q_{t+\Delta t} = Q_t(1 - 0.62Q_t)$
Model 2	18.605	34.736	$Q_{t+\Delta t} = 0.4Q_t + 0.174R$
Model 3a	15.109	21.772	$Q_{t+\Delta t} = \frac{R_m}{12} + 5 + \frac{Q_t}{2 + R_6}$
Model 3b	15.990	9.956	$Q_{t+1} = (13 + R_m + 6Q_t + 2R_6)(1 + 0.25R_6)^{-1}$

From the results presented in Table 1, the following observations can be made regarding the three models:

- Model 1: Performance of Model 1 is quite poor with a validation RMSE of 39.31 Mm^3.
- Model 2: With the inclusion of current rainfall, Model 2 marginally improves the forecast with a RMSE of 34.73 Mm^3, which is an improvement of about 11% as compared to Model 1.
- Model 3: With rainfall information available from 10 additional imaginary rain gauge stations, the RMSE improved to 21.77 Mm^3. The improvement in RMSE for Model 3 was about 37% as compared to the RMSE for Model 2. A closer look at the GP derived model indicates that out of 12 input variables, only rainfall information from two locations are found to affect the modeling (viz., Manimuthar and location no. 6). Model 3b was obtained using only the most significant rainfall information as obtained in Model 3(a) above for GP training. The result indicates a significant improvement in the forecast.

Table 2. Forecast results for monthly models

Name	RMSE		GP equations
	Training	Validation	
Feb	2.971	0.727	$Q_{t+1} = \frac{1}{Q_1}\left[\frac{Q_1^{\,2} + QR_mQ_1 + 3R_m}{Q_3 - 1.2}\right] - (Q_3 + R_m)$
March	3.325	0.860	$Q_{t+\Delta t} = Q_3\left\{\left[Q_1 + 8R_m + 2Q_3 + 2Q_1 + Q_2\right]Q_3^{-1} + R_m + Q_2\right\}\left\{R_m - Q_2\right\}^{-1}$
April	0.668	0.655	$Q_{t+1} = \left[3\left(\frac{Q_2^2 R_3}{R_2}\right) + 6.04\left(\frac{R_1}{Q_2 - R_3 + R_1}\right) + R_3 + \left(\frac{R_3 - R_2}{R_2 - R_1}\right)\right]R_3^{-1} - \left(\frac{Q_1 - R_1}{2R_2 - R_1}\right)$
May	1.501	0.744	$Q_{t+1} = \frac{R_mQ_1Q_2^2}{2.2Q_3^2}\left[Q_1Q_2 + Q_2 + 2.2\frac{Q_1}{Q_3}\right] + 1.7$
June	2.183	0.897	$Q_{t+1} = \left[\left[\frac{R_m}{Q_2} - R_m - Q_1\right] - \left[\left[4Q_1 + 2Q_3 + \frac{R_m}{R_m - (Q_1 + Q_2 + Q_3)}\right]\left[\frac{Q_3}{R_m - Q_1 - Q_3}\right]^{-1}\right]Q_1 + Q_3 + R_m\right]^{-1}$
July	2.116	0.94	$Q_{t+1} = 12 + \left[\left[\frac{17 + Q_3 + Q_2}{8.4 - Q_3 - R_m}\right]\left[Q_1 + Q_3 - \frac{R_m(Q_1 + Q_3)}{2Q_2Q_3 + 4Q_2^2}\right]\right]\left[\frac{R_m}{Q_3Q_2}\right]^{-1}$
August	1.628	0.779	$Q_{t+1} = \frac{1}{1 + Q_3}\left[\frac{R_m}{Q_3}(2Q_1 - Q_2) - \frac{Q_3}{R_m(Q_3 - Q_1)} + 3Q_3 + Q_2 + Q_1\right]$
Sep	2.277	0.882	$Q_{t+1} = 2.3\left[\left[12Q_1 + (R_m - 2Q_1 - Q_2)\frac{Q_3}{Q_2}\right]\left[Q_3 + R_m\right]^{-1}\right]$
Oct 1	2.927	0.907	$Q_{t+1} = \left[Q_1Q_2 - Q_2^2 + \frac{Q_2}{0.009 * Q_3} + 3Q_1Q_3 + \frac{Q_2}{0.006 * Q_3} + R_m\right]\left[\frac{Q_2}{Q_3^2} + R_m\right]^{-1}$
Oct 2	1.245	0.822	$Q_{t+1} = \left[3R_m + R_2 - \frac{R_3}{Q_3} + \left(\frac{Q_3R_2 - R_3}{Q_3 - R_m}\right) + \frac{2R_m}{Q_3^2} - 2\frac{R_1}{Q_3} - \frac{Q_2R_2}{Q_3}\right]R_m^{-1}$
Nov	2.815	0.902	$Q_{t+1} = \left[Q_3Q_1 + Q_1^3 + 2Q_3Q_1^2 - Q_3Q_2Q_1 - Q_1^2Q_2 - 0.25 * Q_1Q_2^2\right]\left[0.5 * R^2Q_3^3 - R_mQ_3Q_2 + R_m{}^2Q_3Q_2Q_1\right]^{-1}$
Dec	3.614	0.732	$Q_{t+1} = 19.217 + 2R_m + 2Q_3 + Q_1 + \left[Q_3(R_m + 4Q_3 - Q_1 - 26.35)\right]\left[Q_2Q_1(R_m - Q_3)(Q_1 + Q_2 - 2Q_3)\right]^{-1}$

5.2. Higher Lead Time Forecast

The results are presented in Table 2, from which the following observations can be made:

- Comparison of model performance with only antecedent inflows and that with a combination of antecedent inflow and rainfall clearly indicate a drastic improvement in the forecast accuracy when rainfall information from surrounding regions are included in the model.
- Except for the months of April and October, for all other months, only rainfall information from Manimuthar dam site is found to be sufficient to model the inflow process.
- For the months of April and October rainfall information from location nos 1, 2 and 3 are found to significantly affect the process. These locations are far removed from the reservoir site. A discussion with

the site engineer revealed that there are small streams in that region and as such the appearance of these variables in the model is quite reasonable.

6. CONCLUSIONS

This study presents an application of forecasting inflows of the Manimuthar reservoir in Tamil Nadu, India using a data-driven modelling technique, Genetic Programming (GP). The main aim of this study was to test an innovative way to select the variables in the input vector for the GP model. In the input vector, along with the current inflow, the rainfall information from ten additional imaginary stations around the Manimuthar region is also used. This rainfall information is first estimated in those crucial locations along the river reach or the surrounding region which plays an important role in deciding the inflow to reservoir. It is observed that when the rainfall information at the additional imaginary stations are included in the input vector, the forecasts show significant improvement indicating that incorporation of appropriate input variables play a crucial role in the model forecast accuracy. Thus, the proposed approach appears to be quite innovative and has potential for improving monthly flow forecasts.

ACKNOWLEDGMENTS

The authors would like to acknowledge and thank the Public Works Department (Thirunelveli District), Tamil Nadu, India for providing the monthly inflow data of Manimuthar reservoir. Further, the first and the last author wish to extend their sincere thanks for financial support from CSIR, New Delhi for carrying out this work.

REFERENCES

Babovic, V. and Keijzer, M. (2000). Genetic programming as a model induction engine. *Journal of Hydroinformatics*, 2 (1), 35-60.

Coulibaly, P., Anctil, F. and Bobee, B. (2000). Daily reservoir inflow forecasting using artificial neural networks with stopped training approach, *Journal of Hydrology*, 230 (3-4) 244-257.

Firat, M. and Turan, M. E. (2010). Monthly river flow forecasting by an adaptive neuro-fuzzy inference system. *Water and Environment Journal*, 24, 116–125.

Giustolisi, O. (2004). Using genetic programming to determine Chezy resistance coefficient in corrugated channels, *Journal of Hydroinformatics*, 6 (3), 157–173.

Harris, E.L., Babovic, V. and Falconer, R.A. (2003). Velocity prediction in compound channels with vegetated floodplains using genetic programming, *International Journal of River Basin Management*, 1 (2), 117–123.

Karunanithi, N., Grenney, W.J., Whitley, D. and Bovee, K. (1994). Neural networks for river flow prediction. *Journal of Computing in Civil Engineering*, 8, 201–219.

Khu, S.T., Liong S.Y., Babovic, V., Madsen, H. and Muttil, N. (2001). Genetic programming and its application in real-time runoff forecasting. *Journal of American Water Resources Association*, 37 (2), 439-451.

Kişi, O. (2008a). Stream flow forecasting using neuro-wavelet technique. *Hydrological Processes*, 22 (20), 4142–4152.

Kişi, O. (2008b). River flow forecasting and estimation using different artificial neural network techniques. *Hydrology Research*, 39 (1), 27–40.

Koza, J.R. (1992), Genetic Programming: On the Programming of Computers by Means of Natural Selection, The MIT Press, Cambridge, MA.

Liong, S.Y., Gautam, T.R., Khu, S.T., Babovic, V. and Muttil, N. (2002). Genetic Programming: A new paradigm in rainfall-runoff modeling, *Journal of American Water Resources Association*, 38 (3), 705-718.

Muttil N. and Lee J.H.W. (2005), Genetic programming for analysis and real-time prediction of coastal algal blooms. *Ecological Modelling*, 189 (3-4), 363-376.

Shamseldin, A.Y. (2004). Hybrid neural network models. Chapter 4 in: Abrahart, R.J., Kneale, P.E. and See, L.M. (eds.) Neural Networks for Hydrological Modelling. Rotterdam: A.A. Balkema Publishers.

Singh, M., Singh, R. and Shinde, V. (2011). Application of software packages for monthly stream flow forecasting of Kangsabati River in India. *International Journal of Computer Applications*, 20 (3), 7-14.

Sivapragasam, C., Arun, V.M. and Muttil, N. (2011). Re-design of rain gauge network using genetic programming based ordinary Kriging. In: 34th IAHR World Congress - Balance and Uncertainty, 26 June - 1 July (2011), Brisbane. Australia: IAHR & Engineers Australia, pp 428 - 433.

Thirumalaiah, K. and Deo, M. C. (2000). Hydrological forecasting using neural networks, *Journal of Hydrologic Engineering*, 5(2), 180-189.