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

Comparison of two data-driven approaches for daily river flow forecasting

Achela K. Fernando[a], **Asaad Y. Shamseldin**[b] **Robert J. Abrahart** [c]

[a] *Department of Civil Engineering, Unitec Institute of Technology, Auckland, New Zealand*
Email: afernando@unitec.ac.nz
[b] *Department of Civil and Environmental Engineering, University of Auckland, New Zealand*
[c] *School of Geography, University of Nottingham, Nottingham, UK*

Abstract: Ongoing research on the use of data-driven techniques for rainfall-runoff modelling and forecasting has stimulated our desire to compare the effectiveness of transparent and black-box type models. Previous studies have shown that models based on Artificial Neural Networks (ANN) provide accurate black-box type forecasters: whilst Gene Expression Programming (GEP: Ferreira, 2001; 2006) provides transparent models in which the relationship between the independent and the dependant variables is explicitly determined. The study presented in this paper aims to advance our understanding of both approaches and their relative merits as applied to river flow forecasting. The study has been carried out to test the effectiveness of two forecasting models: a GEP evolved equation and a model that uses a combination of ANN and Genetic Algorithms (GA). The two approaches are applied to daily rainfall and river flow in the Blue Nile catchment over a five year period.

GeneXproTools 4.0, a powerful soft computing software package, is utilised to perform symbolic regression operations by means of GEP and in so doing develop a rainfall-runoff forecasting model based on antecedent rainfall and river flow inputs. A transparent model with independent variables of antecedent rainfall and flow to forecast river discharge could be achieved.

The ANN model is developed with the assistance of a GA: the latter being used in the selection of the ANN inputs from a pre-determined set of external inputs. The rainfall and flow data for the first four years was used to develop the model and the final year of data was used for testing.

The paper describes the methods used for the selection of inputs, model development and then compares and contrasts the two approaches and their suitability for river flow forecasting. The results of the study show that the GEP model is a useful transparent model that is superior to the ANN-GA model in its performance for riverflow forecasting.

Keywords: *Gene Expression Programming, Artificial Neural Network, Genetic Algorithms, Rainfall-runoff Model, River Flow Forecasting*

1. INTRODUCTION

Rainfall-runoff modelling is a much researched area in hydrological engineering. Development of such models is intended to serve two purposes: to advance our understanding of the transformation of rainfall to runoff; and to provide practical solutions to water resources management problems. The fact that understanding the complex hydrological process involved in the rainfall to runoff transformation phenomenon is not important for the latter purpose has given rise to numerous data-driven techniques that make use of only the antecedent rainfall and flow data for forecasting purposes. These methods include Artificial Neural Networks (ANN) (Abrahart and See 2000, Fernando and Jayawardena 1998, Shamseldin 1997, Shamseldin, et al. 2007), Genetic Algorithm (GA) assisted ANN or GA-assisted parameter optimisation (Abrahart, et al. 1999, Fernando and Jayawardena 2007), Genetic Programming (Babovic and Keijzer 2002), and Gene Expression Programming (GEP) (Fernando, et al. 2009). Of these techniques GEP is the most-recent mathematical modelling technique (Ferreira 2001, 2006) but it is also somewhat different from some of the others in that the model is not completely a “black-box” and the relationship between the input (antecedent rainfall and runoff) and the output (forecast runoff) can be expressed in a mathematical representation. This paper focuses on comparing a GA-assisted ANN model with a model that is derived using GEP for forecasting daily flow in the Nile catchment.

GeneXproTools 4.0 is a powerful soft computing software package for performing GEP. In this study it is utilised to perform symbolic regression operations and develop a GEP river flow model using the most recent observed daily rainfall and flow data for the Blue Nile catchment located in East Africa.

The NeuroSolutions 5.0 is used to develop ANN-GA river flow forecasting model for the Blue Nile catchment. The structure of the ANN model is based on the Multi-Layer Perceptron with one hidden layer. In this study the GA is used to determine the optimum number of hidden neurons as well as in the selection of the ANN inputs from the independent variables of antecedent rainfall and flow.

2. MODEL BASED ON GEP

The GEP model developed in this study used daily rainfall and flow data for the Blue Nile River from 1992-1997 inclusive. The catchment area upstream of the flow monitoring location is approximately 25,4230km². The flow has a very seasonal variation with a clear rainy season and a dry season. The daily average rainfall, and discharge for this river catchment are ~3.8 mm/day and 1578 m³/s respectively for the 5 year study period. The daily river flow records used in the study begin on the January 1st 1992; the first 4 years of data provided a training data set of 1433 input/output vectors; the final year of data provided a testing set of 348 vectors.

2.1. Input data

To identify which were the most appropriate input data to forecast one-day-ahead daily flow, cross correlation analysis between rainfall and flow at different temporal lags was performed.

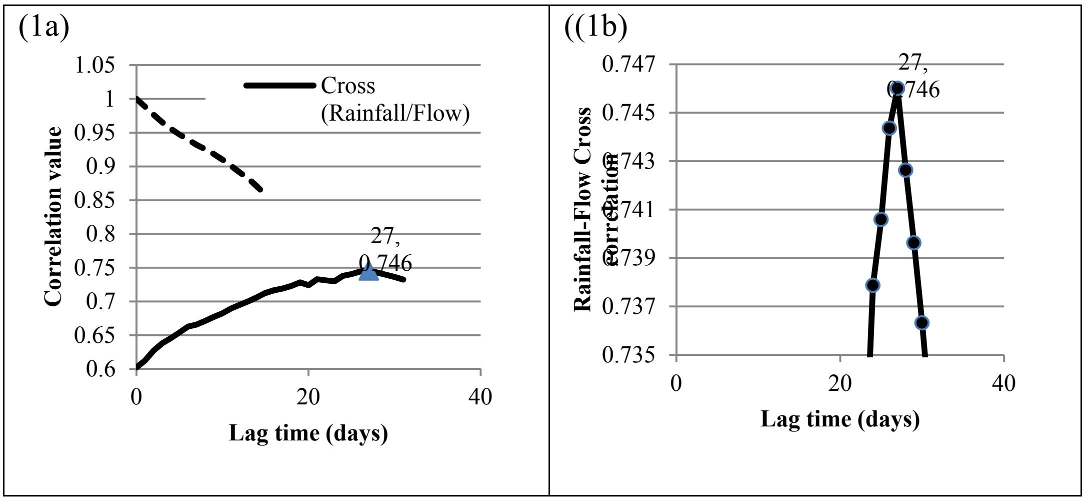


Figure 2. (a) Cross- and serial-correlation for rainfall and flow data

Figure 1a&b above show that the highest cross correlation between the antecedent rainfall values and the flow data occurs when the lag time is 27 days which implies the current flow in the river is most dependent on the rainfall that occurred 27 days prior to today.

Hence the antecedent rainfall data R(t-26), R(t-25), R(t-24) were used as input. As shown in Figure 1A, the serial correlation for flow decreases with increasing lag time and hence the significant flow data Q(t-2), Q(t-1), Q(t) were chosen as inputs to forecast the flow value Q(t+1).

2.2. GEP Model development settings

A total of 15 functions shown in Table 1 were allowed to be used in the model development within GenXPro Tools® software. The other general settings are as specified below:

- Independent variables:
  - Input d[0] : Q(t) (m^3/s)
  - Input d[1] : Q(t-1) (m^3/s)
  - Input d[2] : Q(t-2) (m^3/s)
  - Input d[3] : R(t-24) (mm/day)
  - Input d[4] : R(t-25) (mm/day)
  - Input d[5] : R(t-26) (mm/day)
- Dependent variable: Forecast flow Q(t+1) (m^3/s)
- Output : GEP model estimate of forecast flow (m^3/s)
- Number of training samples: 1433
- Number of testing samples: 348
- Number of chromosomes: 30
- Head size: 8
- Number of genes : 3. (three trees will be needed to form the final mapping function)
- Linking function : Addition (functions on the trees will be added to form the final mapping function)
- Constants: Two constants per gene with bounds of ±10
- Fitness function: Based on RMSE
- Dynamic evolution operators: Default values of mutation, inversion, transportation, recombination and transposition
- Symbolic functions: Fifteen default functions (Table 1)
- Stopping criterion: 20,000 generations

Table **1.** Function Set

Function	Symbol
Addition	+
Subtraction	-
Multiplication	*
Division	/
Floating-point remainder	Mod
Power	Pow
Square root	Sqrt
Exponential	Exp
10^x	POw10
Absolute Value	Abs
x to the power of 2	x2
x to the power of 3	x3
x to the power of 4	X4
Sine	Sin
Cosine	Cos

2.3. GEP Model output

The expression trees obtained for the model are as shown below in Fig. 3. As the expression trees are summative, the equation derived for the three expression trees and the final forecast model respectively are:

Tree 1 =[Mod((Sin(Sin(R(t-26)))+3.95),3.84)]**3
Tree 2 = Q(t)-Mod(exp(R(t-24)),Sqrt(Q(t)-4.17)))
Tree 3 = Sqrt((Abs(Q(t-1))**Cos(Abs(Mod((R(t-26)-R(t-24)),4.52)))))
Q(t+1) = Tree 1 + Tree 2 + Tree 3

A significant detection is that in estimating Q(t+1), Q(t) contributes as a direct summation while Q(t-2) has not played any part. It must also be noted that R(t-24) and R(t-26) have been included in the model while R(t-25) has not.

It is plausible for this method to produce other similar models with different combinations of functions that may also be equally accurate. It is noteworthy that GEP can produce a suite of equifinal models. Also, the GEP model is transparent in that it directly provides the equation involved in the input-output transformation. Equations can indeed be written for ANN models too, however, not with the same ease. At this stage of the research neither the comparison of the equifinal models nor the interpretability of the hydrology in physically meaningful way using the functions involved is considered.

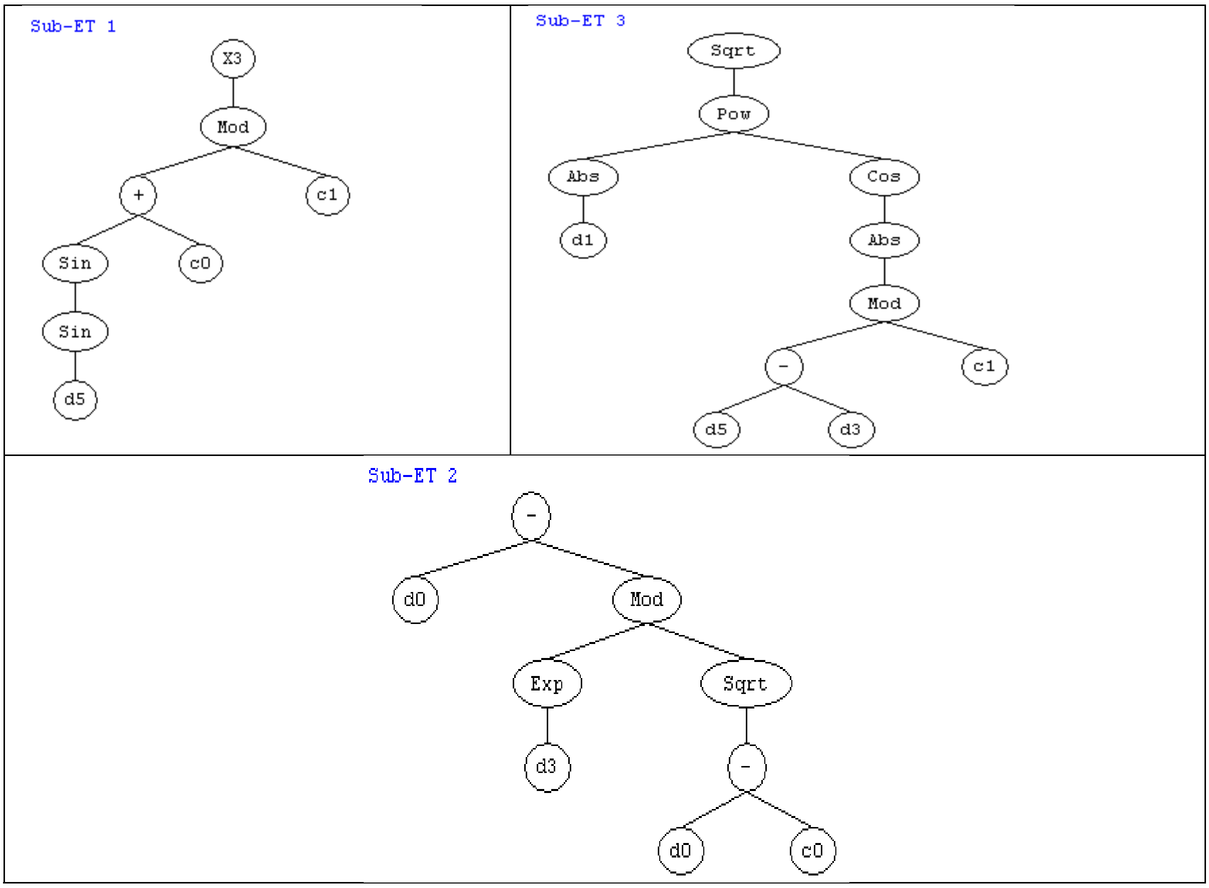


Figure 3. The Expression trees (summative) obtained for the GEP Model

3. MODEL BASED ON GA ASSISTED ANN

3.1. Model build

The structure of Artificial Neural Network (ANN) used in the study is based on the Multi-Layer Perceptron (MLP). The MLP used in this study consists of a network of interconnected computational elements (neurons), linked together by connection pathways, which are arranged in a series of layers. The neuron layers of the MLP are the input layer, the output layer, and the hidden layer located between the input and output layers. The input layer receives the ANN external input array with each input element being assigned to only one neuron. The input neuron relays its external input without transformation to each of the neurons in the hidden layer. Thus, each neuron in this hidden layer has an input array consisting of the outputs of the input layer neurons. Each hidden layer neuron produces only a single output which becomes an element of the input array to each neuron in the subsequent (output) layer.

In the case of hidden and output layer neurons, the process of the input-output transformation is achieved by a transfer function which is a non-linear transformation of the total sum of the products of each of its input array elements with its corresponding synaptic weight plus a constant. The weights and the threshold values are the parameters of the network, which are determined during training/calibration. The calibration is achieved by minimising the least squares objective function using non-linear optimisation algorithms. A non-linear transfer function with upper and lower limits of ±1 is generally used for neurons in the hidden and the output layers. As the actual input values are usually outside this range a rescaling of inputs and outputs is done in order to compare the actual value and the output of the network.

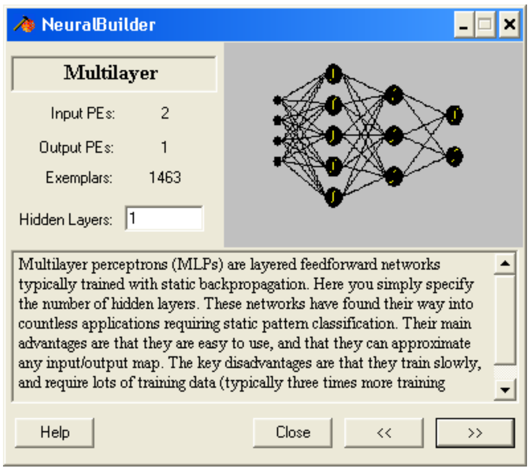
In this study, a GA was used to determine the optimum input(s) to be selected from the same set of inputs used in developing the GEP model, i.e. R(t-26), R(t-25), R(t-24), Q(t),Q(t-1) and Q(t-2). No cross-validation set was used during training.

3.2. ANN-GA Model development settings

The inputs to the ANN-GA model in NeuroSolutions® were the antecedent rainfall and flow. In the model building process the settings summarised in Table 2 were used. Figure 4 below shows a screen dump of a typical MLP for the training set.

Table 2. Parameter setting for ANN-GA model build in NeuroSolutions®

Parameter	**Setting**
Input data range	First 4 years of rainfall and flow data
Input optimisation	Genetic Algorithm
Number of hidden layers/nodes	1/2
Processing elements	Genetic Algorithm
Transfer function at hidden layer	Logistic function
Learning rule	Conjugate gradient
Transfer function at output layer	Sigmoid function
Maximum training epochs	1000
Termination criterion	MSE, Threshold of 0.01
Weight update	Batch



Multilayer

Input PEs: 2

Output PEs: 1

Exemplars: 1463

Hidden Layers: 1

Multilayer perceptrons (MLPs) are layered feedforward networks typically trained with static backpropagation. Here you simply specify the number of hidden layers. These networks have found their way into countless applications requiring static pattern classification. Their main advantages are that they are easy to use, and that they can approximate any input/output map. The key disadvantages are that they train slowly, and require lots of training data (typically three times more training

Figure 4. Typical MLP structure used by NeuroSolutions® for training data set

3.3. ANN-GA Model output

The ANN-GA model created by NeuroSolutions® consisted of one element Q(t). These were chosen by the GA optimization technique in the software. The model had 2 neurons in the hidden layer. There was no attempt to prevent over-fitting by using a cross validation set.

4. MODEL OUTCOMES

Figure 5 below shows the comparison of the forecasts of the GEP and the GA-assisted ANN models with respect to the actual flow values. Both model predictions seem to closely follow the actual flow hydrograph except at some peak values.

Figures 6 shows that GEP model predictions are better correlated with the actual flow for both training and testing sets; ANN-GA model underpredicts the high flows.

In Table 3 below is summarised some of the basic statistics of the model outputs for comparison with the relatively superior value bold and underlined. The observations made on the graphs are confirmed; that the GEP model performs better overall than the GA-ANN model. The additional benefit of the GEP model is that it is transparent.

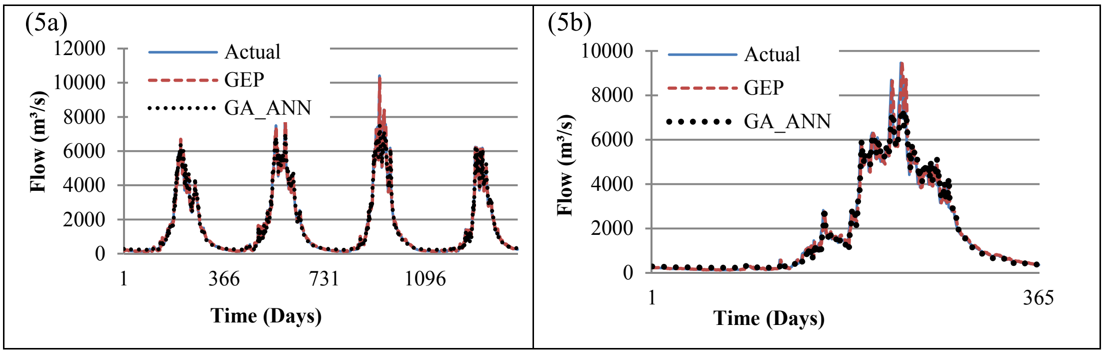


Figure 5. Actual and model predicted flow for (a) training data set (b) Testing data set

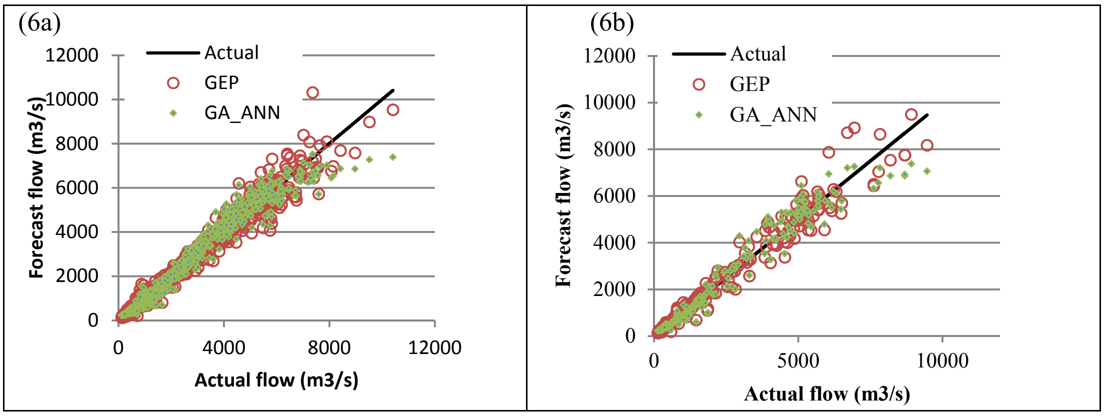


Figure 6. Scattergrpah for predicted flows for (a) training data set (b) Testing data set

Table 3: Summary of model performance statistics

Performance indicator	**GEP**	**ANN_GA**	**Actual**
RMS Error (Training) (m^3/s)	**270**	295	
RMS Error (Testing) (m^3/s)	**366**	392	
R^2 Error (Training)	**0.980**	0.976	
R^2 Error (Testing)	0.971	0.966	
Mean Flow (Training) (m^3/s)	**1536**	1549	1529
Mean Flow (Testing) (m^3/s)	1779	**1774**	1771
Peak Flow (Training) (m^3/s)	**10316**	7524	10412
Peak Flow (Testing) (m^3/s)	**9492**	7377	9467

5. CONCLUSIONS AND DICUSSION

In this study a comparative river flow forecasting study has been conducted using two data driven models, namely GEP and ANN-GA. The daily data of the Blue Nile catchment located in East Africa is used in the comparison. The results of the comparison reveal that the GEP model performs better than a calibrated ANN-GA model. An added advantage of the GEP model is that the flow forecast can be easily expressed in explicit terms of the independent variables and therefore is transparent.

6. REFERENCES

Abrahart, R. J., and See, L. (2000). "Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecast in two contrasting catchments." *Hydrological processes*, 14, 2157-2172.

Abrahart, R. J., See, L. M., and Kneale, P. E. (1999). "Using pruning algorithms and genetic algorithms to optimise network architectures and forecasting inputs in a neural rainfall-runoff model." *Journal of Hydroinformatics*, 1(2), 103-114.

Babovic, V., and Keijzer, M. (2002). "Rainfall Runoff Modelling Based on Genetic Programming." *Nordic Hydrology*, 33(5), 1-346.

Fernando, A. K., and Jayawardena, A. W. (1998). "Runoff forecasting using RBF networks with OLS algorithm." *Journal of Hydrologic Engineering*, 3(3), 203-209.

Fernando, A. K., and Jayawardena, A. W. (2007). "Use of a supercomputer to advance parameter optimisation using genetic algorithms." *Journal of Hydroinformatics*, 9(4), 319-329.

Fernando, A. K., Shamseldin, A. Y., and Abrahart, R. J. "Using gene expression programming to develop a combined runoff estimate model from conventional rainfall-runoff model outputs." *Proc., 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation*, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, 2377-2383.

Ferreira, C. (2001). "Gene Expression Programming: A New Adaptive Algorithm for Solving Problems." *Complex Systems*, 13(2), 87-129

Ferreira, C. (2006). *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*, 2nd Edition, Springer-Verlag, Germany.

Shamseldin, A. Y. (1997). " Application of neural network technique to rainfall-runoff modelling." *Journal of Hydrology*, 199(3), 272-294.

Shamseldin, A. Y., O'Connor, K. M., and Nasr, A. E. (2007). "A comparative study of three neural network forecast combination methods for simulated river flows of different rainfall-runoff models." *Hydrological Sciences*, 52(5), 896-916.