20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Regional flood estimation in Australia: Application of gene expression programming and artificial neural network techniques

K. Aziz[a], A. Rahman[a], A. Shamseldin[b], M. Shoaib[b]

[a]School of Computing, Engineering and Mathematics, University of Western Sydney, NSW, Australia
[b]Department of Civil and Environmental Engineering, University of Auckland, New Zealand
Email: k.aziz@uws.edu.au

Abstract: Flood damage can be minimised by ensuring optimum capacity to drainage structures. An underdesign of these structures increases flood damage cost whereas an overdesign incurs unnecessary expenses. The optimum design of water infrastructures depends largely on reliable estimation of design floods which is a flood discharge associated with a given annual exceedance probability. For design flood estimation, the most direct method is flood frequency analysis which requires long period of recorded streamflow data at the site of interest. This is not a feasible option at many locations due to absence or limitation of streamflow records; hence regional flood estimation methods are preferred. Regional flood frequency analysis (RFFA) involves transfer of flood characteristics from gauged to ungauged catchments. The RFFA methods are widely used in practice.

In the past, different RFFA methods have been proposed for Australia, which are based on linear models such as Probabilistic Rational Method (PRM) and index flood method. More recently, regression-based methods have been investigated for Australia, which are also log-linear models. There have been successful application of non-linear models like Artificial Neural Networks (ANN), Gene Expression Programming (GEP) and Fuzzy based methods in hydrology in some other parts of the world. However, there has not been any notable application of these methods in RFFA study in Australia. This paper focuses on the application of the ANN and GEP to regional flood estimation problems in Australia. The GEP approach used in this study provides an integrated mechanism for the identification of the optimum hydrological regions for RFFA study in eastern Australia. In the preliminary study, optimum regions were obtained based on geographic and state boundaries, climatic conditions and catchment attributes. The proposed approaches were applied to 452 stations in the eastern Australia. Results depict that the GEP and ANN approach have a much better generalization capability of RFFA problems. An independent test has shown that the ANN based model provides more accurate flood quantile estimates than the GEP. Overall, the best ANN-based RFFA model is achieved when all the data set of 452 catchments are combined together to form one region, which gives an ANN-based RFFA model with median relative error of 35% to 44% and median ratios (of predicted and observed values) of 0.99 to 1.14.

Keywords: *Flood estimation, artificial neural networks, gene expression modelling, flood frequency, ungauged catchment, quantile regression technique*

1. INTRODUCTION

Estimation of large to extreme floods is a necessity in the design of major water infrastructures such as dam spillways, large weirs and major bridges. Lack of long records for reliable estimation of design flood peaks has always been a concern for hydrologist and has prompted the development of regional flood frequency analysis (RFFA) methods. RFFA techniques enable the estimation of design floods at ungauged catchments. RFFA enhances the flow statistics at sites where streamflow records are short (Shabri and Jemain, 2013).

Different regional flood estimation methods were proposed for different parts of Australia in its national guide called Australian Rainfall and Runoff (ARR). Among these, various forms of the rational method and the index flood method are the most common. However, these methods have not been updated in Australia since 1987. Because of changing climatic conditions, availability of additional streamflow data and improvements in regional flood estimation methods in the past decade, there is a need to look for new regional flood estimation techniques for Australia. Some of the recent developments in regional flood estimation methods in Australia include *L* moments based index flood method, various forms of regression techniques (Bates et al., 1998; Rahman et al., 1999; Rahman, 2005; Haddad et al., 2012). All of these techniques are based on the assumption of linear models (either in original data space or in log-log space), an assumption that may not be satisfied in many cases.

Increased computing power has created new opportunities for hydrologists for the solution of complex problems. For example, there have been applications of artificial intelligence based methods such as Artificial Neural Networks (ANN) and Gene Expression Programming (GEP) in hydrology, particularly in streamflow forecasting problems. ANN and GEP are computational models that help in input output mapping. Muttiah et al. (1997), Hall and Minns (1998), and Dawson et al. (2006) are among others who have successfully applied ANN in hydrology. There has been limited application of ANN and GEP in RFFA problems in Australia. Application of ANN and GEP may help developing improved regional flood estimation techniques for Australia. Unlike regression based approach, the ANN and GEP do not impose any fixed model structure on the data rather the data itself identifies the model form through use of artificial intelligence.

2. ANN AND GEP-BASED MODELS IN HYDROLOGY

Most hydrologic processes are highly nonlinear, with a higher degree of spatial and temporal variability. The uncertainty in parameter estimates makes them more complicated. ANN is a parallel system and is capable of resolving paradigms that linear computing cannot solve. The ANN has been widely adopted for a range of hydrological problems such as rainfall-runoff modeling, streamflow forecasting and water quality modeling (e.g. Govindaraju, 2000; Chokmani et al., 2008; Turan and Yurdusev, 2009). The Task Committee on Application of ANN in Hydrology by ASCE (2000) stated that ANN should be classified as empirical models, which treat hydrologic systems (such as a watershed) as a black-box and attempt to find a relationship between historical inputs (e.g. rainfall and streamflow) and outputs (e.g. catchment runoff measured at a stream gauge). There have been relatively few applications of ANN to RFFA to estimate flood quantiles in ungauged catchments. For example, Dawson et al. (2006) applied ANN to develop a model for index flood using data from 850 UK catchments and found that ANN provided more accurate flood quantile estimates than the QRT. They pointed out that ANN are heavily data dependent and cannot explicitly account for physical processes, reducing confidence in model predictions. Other ANN based RFFA studies include Muttiah et al. (1997) who used a large data set from the US to predict 2 year peak flood. In Australia, Daniell (1991) adopted ANN to 14 catchments in Australian Capital Territory (ACT) to develop a RFFA model; however, due to limited data set the method could not produce any meaningful prediction. Recently, in Australia Aziz et al. (2010, 2011, 2012 and 2013) have applied ANN-based RFFA methods to eastern Australia and found that ANN-based RFFA methods can provide quite accurate regional flood estimates. The main focus of this paper is to compare the ANN-based RFFA methods with Gene expression programming (GEP).

The GEP is a computing method that is capable of generating a 'transparent' and structured representation of the system. This has been applied with success in water resource engineering (Rabunal et al. 2007; Guven et al. 2008). This has drawn the hydrologists in investigating the use of GP in estimating the river flow data and runoff estimate model (Seckin and Guven, 2012; Guven and Talu, 2010). The most relevant study to RFFA has been conducted by Seckin and Guven (2012). They applied GEP for the estimation of peak flood discharges at ungauged sites across Turkey. The study covered 543 ungauged sites across Turkey. Drainage area, elevation, latitude, longitude, and return period were used as the inputs while the peak flood discharge was the output.

3. WORKING STRUCTURE OF ANN AND GEP

The ANN method adopted in this study is based on the structure of the multi-layer perception (MLP), which has been widely used in hydrological modelling (Shamseldin, 1997). A network of interconnected neurons linked by connection pathways form the MLP structure as shown in Figure 1. In this study, the adopted ANN model has three layers of neurons or nodes: an input layer, a hidden layer and an output layer. The layers of neuron communicate via a weighted connection network. There are four types of weighted connections: feedforward, feedback, lateral, and time-delayed connections. Each neuron has a number of inputs and a number of outputs (leading to the subsequent layer or out of the network). In Figure 1 neurons are shown by circles and lines representing the connections. The computation required at each neuron is simple where each input is multiplied by a connection parameter known as weight, and combined (usually with certain bias) to produce a single value. A transfer function is used to operate this value. This functional form helps to determine the response of a node to the total input signal it receives. The functional form used in this study is sigmoid which is a bounded, monotonic, non-decreasing function that provides a graded and nonlinear response. This function enables a network to map any nonlinear process. Typically the hyperbolic tangent sigmoid function used in this study is:

$$f(x)=\frac{1-e^{-x}}{1+e^{-x}} \tag{1}$$

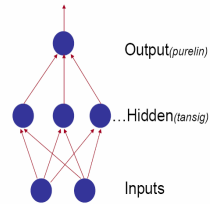

Figure 1. Configuration of a feedforward three-layered ANN.

GEP automatically generates algorithms and expressions for the solution of a problem, which are coded as a tree structure with its leaves (terminals) and nodes (functions). The generated candidates (programs) are evaluated against a "fitness function" and the candidates with higher performance are then modified and re-evaluated. This modification evaluation cycle is repeated until an optimum solution is achieved. Chromosomes and the expression trees (ETs) are two important components of the GEP. The ETs are the expression of the genetic information encoded in the chromosomes. To predict a flood quantile, a set of independent variables (predictor variables) is to be used in the GEP. A set of functions (e.g., e^x and sin(x)) and arithmetic operations (+, -, /, *) are used. The terminals and the functions form the junctions in the tree of a program. The GEP gene contains head and a tail. The head contains the functions and the terminals are represented by the symbols while tail only contains the terminals. For each problem the length of the head of the gene *h* is selected whereas tail's length is a function of length of the gene's head.

4. STUDY AREA AND DATA

This study focuses on the eastern states of Australia. This includes the states of New South Wales (NSW), Victoria (VIC), Queensland (QLD), and Tasmania (TAS). This part of Australia is selected because of a rich spatial and temporal data of gauged catchments in this region. This data is more comprehensive than other parts of Australia. Filling of gaps, checking for trends, outliers and rating curve error in streamflow as detailed in Haddad et al., (2010), Rahman et al. (2009) and Rahman et al. (2012) were adopted to prepare the streamflow data. Here, annual maximum flood series data are used. A total of 452 stations were finally selected for this study (Figure 2), which include data from four states: NSW (96) VIC (131), QLD (172) and TAS (53). The catchment sizes of the selected 452 stations range from 1.3 km^2 to 1900 km^2 with the median value of 256 km^2. The annual maximum flood record lengths of the selected stations range from 25 to 75 years (mean: 33 years).

5. METHOD

The available data set of 452 stations was divided into 80% (362 stations) for training, 20% (90 stations) for testing. The training and testing data sets were selected randomly out of the total 452 stations. Both ANN and GEP based RFFA models were developed to predict the 2, 5, 10, 20, 50 and 100 years ARI floods. Two best performing predictor variables (Catchment area (*A)* and design rainfall intensity ($I_{tc,ARI}$) were selected for

prediction equation (Aziz et al., 2010). All the gauging stations (452) from eastern states of Australia are combined in this study to form one region (Aziz et al., 2011, 2013).

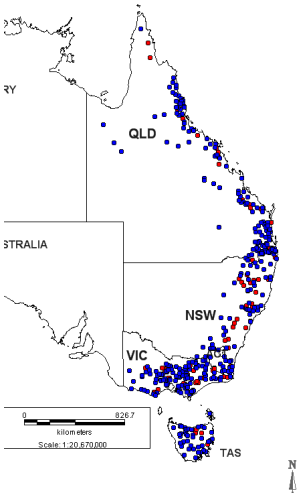

Figure 2. Location of study catchments (Red colour shows the test catchments).

In case of ANN, a feedforward ANN consisting of three layers (input, hidden and output layers) was used with the training algorithm known as 'backpropagation of error'. Three hidden-layered neural networks were selected with 7, 3 and 1 neurons to each of these three layers. Two inputs (A, I_{tc_ARI}) were used in one input layer and one output layer with one output (Q_{pred}). The transfer function used for the hidden layers and the output layer was hyperbolic tangent sigmoid function. Transfer functions calculate a layer's output from its net input. Each predictor and predictand was standardized to the range of [0.05, 0.95], such that extreme flood events which exceeded the range of the training data set could be modelled between the boundaries [0, 1] during testing. A learning rate of 0.05 was used together with a momentum constant of 0.95.

In order to obtain the best GEP model, the MSE values between the observed and predicted flood quantiles were calculated and the training was undertaken to minimise this error. Mean squared error (MSE) was taken as fitness function. Lavenberg-Marquardt method was used as the training algorithm to minimise the mean squared error. MATLAB and GenXProTools were used for the ANN and GEP analysis respectively. In order to develop the combined model in GenXProTools®, the parameter settings as shown in Table 1 were used.

Three statistical measures were used for model evaluation as discussed below.

- Coefficient of efficiency

$$\text{CE} = 1 - \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - \overline{Q}_i)^2} \tag{2}$$

- Ratio between predicted and observed flood events

$$\text{Ratio} = r = \frac{Q_{predicted}}{Q_{observed}} \tag{3}$$

- Relative error (RE):

$$\text{RE}\ (\%) = \text{Abs}\left[\frac{(Q_{pred} - Q_{obs})}{Q_{obs}} \times 100\right] \tag{4}$$

Where Q_{pred} is the flood quantile estimate from the GEP, ANN-based or QRT model and Q_{obs} is the at-site flood frequency estimate obtained from LP3 distribution using a Bayesian parameter fitting procedure (Kuczera, 1999).

Table 1. Parameters used in GEP model

Parameters	Description	Amount
P1	Chromosomes	20
P2	No of genes	5
P3	Head size	6
P4	Tail size	7
P5	Fitness function error type	MSE
P6	Linking function	Subtraction
P7	Mutation rate	0.044
P8	Function set	+,-,*,/,x^2, x^3, sqrt, Exp, Ln, Sin, Cos, 3Rt, Atan, Pow, Log, Log2
P9	Inversion rate	0.1
P10	Gene recombination rate	0.1
P10	One point recombination rate	0.3
P10	Two point recombination rate	0.1
P10	Gene Transposition rate	0.1
P10	Data type	Floating-Type

6. RESULTS

Table 2 shows the median ratio (*r)* values for the ANN, GEP and QRT based RFFA models. The results based on ANN and GEP are comparable and closer to 1. For higher ARIs GEP provides better results than ANN. ANN based model performs better for smaller ARIs than a GEP based RFFA model whereas, GEP shows much better values of *r* for higher ARIs e.g. Q_{50} and Q_{100}. When median *r* values of QRT are compared with ANN and GEP based models, two non linear models outperform except for Q_5. Overall, in terms of *r* values GEP based RFFA model performs well as compared to other two models.

Table 2. Median Q_{pred}/Q_{obs} ratio values for ANN, GEP and QRT

Quantiles	Q_{pred}/Q_{obs} ratio (median)		
	ANN	GEP	QRT
Q_2	1.04	1.07	1.15
Q_5	0.99	1.10	1.06
Q_{10}	1.02	1.04	1.35
Q_{20}	1.04	1.02	1.13
Q_{50}	1.14	1.05	1.19
Q_{100}	1.10	1.02	1.28
Overall	**1.06**	**1.05**	**1.19**

The CE values for these models (ANN and GEP) are also compared with each other and also with QRT based RFFA technique as shown in Table 3. ANN based method provides CE value in the range of 0.52 (Q_{100}) to 0.73 (Q_2) which is quite reasonable. In the case of GEP the values of CE ranges from 0.51(Q_2) to 0.68 (Q_{20}). In fact, the CE values for GEP are better for higher ARIs; however, the ANN based models provide better results for smaller ARIs. The CE values for the GEP and ANN are close enough to each other, but a significant difference can be seen for CE values. Figure 3 shows the plot of observed and predicted flood quantiles for 20 years ARI from the ANN based model, which shows quite a good fit for the training and validation data sets. Similar results were found for other ARIs for ANN and GEP based models. This was observed that for majority of the cases the model prediction matches very well with the observed quantiles, but in few cases there were notable differences, which are expected in RFFA for Australia (e.g. Haddad and Rahman, 2012).

Table 3. Coefficient of efficiency (CE) values for ANN, ANFIS, GEP and QRT

Quantiles	Q_{pred}/Q_{obs} CE (median)		
	ANN	GEP	QRT
Q_2	0.73	0.51	0.35
Q_5	0.61	0.67	0.37
Q_{10}	0.63	0.56	0.30
Q_{20}	0.71	0.6	0.37
Q_{50}	0.68	0.63	-8.42
Q_{100}	0.52	0.67	0.38
Overall	**0.65**	**0.62**	**-1.11**

Table 4 shows the median RE values for the RFFA models based on GEP, ANN and QRT. When comparing the results of GEP and ANN, ANN based RFFA technique provides median RE value in the range of 35% to 44%. On the other hand, the median RE values for GEP based RFFA model are in the range of 37% to 45%, which are quite close to ANN based models. But, in this case of median RE values, the GEP outperforms the ANN for higher ARIs e.g. Q_{50} (37%) and Q_{100} (44%). On the contrary, the ANN performs well in the case of smaller ARIs e.g. median RE values of 37% for Q_2 and 35% for Q_{20}. The QRT provides the results in the range of 42% (Q_{20}) to 65% (Q_2). For all the ARIs, non-linear techniques outperform the linear one. But overall when RE value is considered, the ANN based model performs better than other two models as shown in Figure 4.

Table 4. Median relative error values (%) for ANN, GEP and QRT

Quantiles	Q_{pred}/Q_{obs} RE (median)		
	ANN	**GEP**	**QRT**
Q_2	37.56	45.87	65.38
Q_5	40.39	44.95	45.35
Q_{10}	44.63	42.08	57.62
Q_{20}	35.62	41.53	42.64
Q_{50}	39.09	37.87	48.71
Q_{100}	44.53	44.47	51.72
Overall	**40.3**	**42.8**	**51.9**

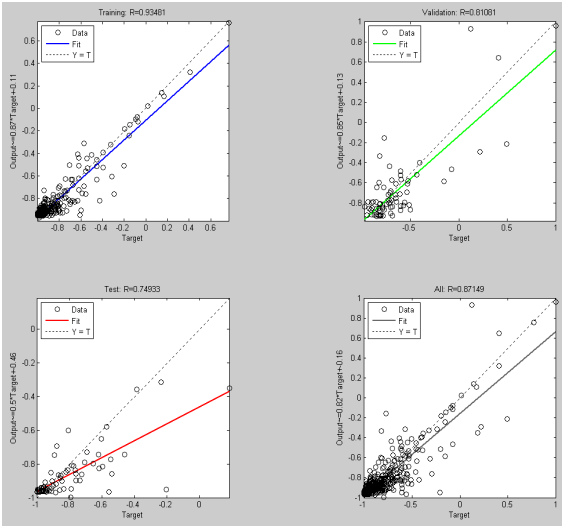

Figure 3. Plot of observed (target) and predicted (output) quantiles for Q_{20} (ANN based model)

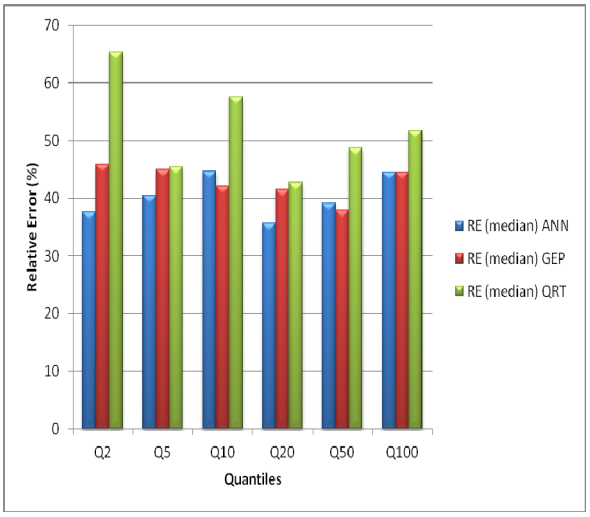

Figure 4. Comparison of RE values between ANN, GEP and QRT based RFFA models

7. CONCLUSION

This paper compares three RFFA methods, two non-linear (ANN and GEP) and one linear (QRT). It has been found that the ANN and GEP provide quite accurate flood estimation in eastern Australia. It has been found that that a backpropagation feedforward ANN consisting of three layers is the best performing model when two predictor variables (catchment area and design rainfall intensity) are used. This model shows a median relative error value in the range of 35% to 44%, median ratio of predicted and observed flood quantiles in the range of 0.99 to 1.14 and coefficient of efficiency values in the range of 0.52 to 0.73. The ANN and GEP methods can be extended to other Australian states.

REFERENCES

ASCE. (2000). Task Committee, 2000. Artificial neural networks in hydrology-I: Preliminary concepts. *Journal of Hydrologic Engineering*, 5, 2, 115–123.

Aziz, K., Rahman, A., Fang, G., Haddad, K., Shrestha, S. (2010). Design flood estimation for ungauged catchments: Application of Artificial Neural Networks for eastern Australia. World Environment and Water Resources Congress, ASCE, Providence, Rhodes Island, USA.

Aziz, K., Rahman, A., Fang, G., Shrestha, S. (2011). Artificial Neural Networks Based Regional Flood Estimation Methods for Eastern Australia: Identification of Optimum Regions. 33rd Hydrology and Water Resources Symposium, 26 June-1 July 2011, Brisbane, Australia.

Aziz, K., Rahman, A., Fang, G. Shrestha, S. (2012). Comparison of Artificial Neural Networks and Adaptive Neuro-fuzzy Inference System for Regional Flood Estimation in Australia, Hydrology and Water Resources Symposium, Engineers Australia, 19-22 Nov 2012, Sydney, Australia.

Aziz, K., Rahman, A., Fang, G., Shreshtha, S. (2013). Application of Artificial Neural Networks in Regional Flood Frequency Analysis: A Case Study for Australia, *Stochastic Environment Research & Risk Assessment*. DOI 10.1007/s00477-013-0771-5.

Bates, B.C., Rahman, A., Mein, R.G. and Weinmann, P.E. (1998). Climatic and physical factors that influence the homogeneity of regional floods in south-eastern Australia. *Water Resources Research*, 34, 12, 3369 – 3381.

Chokmani, K., Ouarda, B.M.J.T., Hamilton, S., Ghedira, M. H., and Gingras, H. (2008). Comparison of ice-affected streamflow estimates computed using artificial neural networks and multiple regression techniques. *Journal of Hydrology*, 349, 83–396

Daniell, T.M. (1991). Neural networks – applications in hydrology and water resources engineering. International Hydrology & Water Resources Symposium. Perth, Australia, 2-4 October. 1991.

Dawson, C.W., Abrahart, R.J., Shamseldin, A.Y. and Wilby, R.L. (2006). Flood estimation at ungauged sites using artificial neural networks, *Journal of Hydrology*, 319, 391–409.

Govindaraju, R.S., (2000). Artificial neural networks in hydrology II. Hydrological applications. *Journal of Hydrologic Engineering*, 5 (2), 124–137.

Guven, A., Avtek, A., Yuce MI, Aksoy H (2008). Genetic programming based empirical model for daily reference evapotranspiration estimation. CLEAN soil Air Water 36(10-11):905-912.

Guven, A., Talu, N. E., (2010). Gene-expression programming for estimating suspended sediment in Middle Euphrates Basin, Turkey. *Clean Soil Air and Water*, 38(12), 1159–1168.

Haddad, K., and Rahman, A. (2012). Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework: Quantile Regression vs. Parameter Regression Technique. *Journal of Hydrology*, 20, 142-161.

Haddad, K., Rahman, A., and Stedinger, J.R. (2012). Regional Flood Frequency Analysis using Bayesian Generalized Least Squares: A Comparison between Quantile and Parameter Regression Techniques, *Hydrological Processes*, 26, 1008-1021.

Haddad, K., Rahman, A., Weinmann, P.E., Kuczera, G. and Ball, J.E. (2010). Streamflow data preparation for regional flood frequency analysis: Lessons from south-east Australia. *Australian Journal of Water Resources*, 14, 1, 17-32.

Hall, M.J., and Minns, A.W. (1998). Regional flood frequency analysis using artificial neural networks Proc. 3rd Intl. Conf. on Hydroinformatics, Vol. 2. Balkema, Rotterdam, 759–763.

Kuczera, G., (1999). Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference. *Water Resources Research*, 35, 5, 1551-1557.

Muttiah, R.S., Srinivasan, R. and Allen, P.M. (1997). Prediction of two year peak stream discharges using neural networks. *Journal of the American Water Resources Association*, 33 (3), 625–630.

Rabunal, JR, Puertas J, Suarez J, Rivero D (2007) Determination of the unit hydrograph of a typical urban basin using genetic programming and artificial neural networks. *Hydrol Process* 27(4):476–485

Rahman, A., Bates, B.C., Mein, R.G. and Weinmann, P.E. (1999). Regional flood frequency analysis for ungauged basins in south-eastern Australia. *Australian Journal of Water Resources*. 3(2), 199-207, 1324-1583.

Rahman, A. (2005). A quantile regression technique to estimate design floods for ungauged catchments in South-east Australia. *Australian Journal of Water Resources*, 9(1), 81-89.

Rahman, A., Haddad, K., Kuczera, G. and Weinmann, P.E. (2009). Regional flood methods for Australia: data preparation and exploratory analysis. Australian Rainfall and Runoff Revision Projects, Project 5 Regional Flood Methods, Report No. P5/S1/003, Nov 2009, Engineers Australia, Water Engineering, pp. 181.

Rahman, A., Haddad, K., Zaman, M., Ishak, E., Kuczera, G. and Weinmann, P.E. (2012). Regional flood methods, Stage II, Project 5 report, School of Engineering, University of Western Sydney, Prepared for Engineers Australia, Report No. P5/S2/015, 319.

Shabri, A. and Jemain, A.A. (2013). Regional flood frequency analysis for Southwest Peninsular Malaysia by LQ-moments. *Journal of Flood Risk Management*, doi: 10.1111/jfr3.12023.

Sekin, N. Guven, A. (2012). Estimation of peak flood discharges at ungauged sites across Turkey, *Water Resources Management*, (2012) 26, 2569–2581.

Shamseldin, A.Y. (1997). Application of a neural network technique to rainfall-runoff modeling. *Journal of Hydrology*, 199, 272–294.

Turan, M.E. and Yurdusev, M.A. (2009). River flow estimation from upstream flow records by artificial intelligence methods. *Journal of Hydrology*, 369, 71–77.