Application of ANN in Regional Flood Estimation: A Case Study for New South Wales, Australia

Sasan Kordrostami[1], Zaved Khan[1], Ataur Rahman[1,2]

[1]School of Computing, Engineering and Mathematics, Western Sydney University, Sydney
Email: sasan.kordrostami@gmail.com
[2] Centre for Infrastructure Engineering, Western Sydney University, Sydney, Australia

Abstract

Flood estimation in ungauged catchments is often needed in hydrology. Regional flood frequency estimation (RFFE) methods can be used for this purpose. The RFFE models in Australia are mainly based on linear models, such as Index Flood Method, Quantile Regression Technique, Parameter Regression Technique and Probabilistic Rational Method. The application of non-linear RFFE techniques such as Artificial Neural Network (ANN) is quite limited in Australia. In this paper, an ANN based RFFE model is presented for New South Wales (NSW) State in Australia. It uses data from 88 gauged catchments in NSW. A total of eight predictor variables are considered and five different model forms are tested. It has been found that when all the eight predictor variables are used, the ANN based RFFE model performs the best generally; however, the gain in model accuracy from using only three predictor variables is marginal. Furthermore, the performances of ANN based RFFE models vary across the six AEPs, and there no model that is the best with respect to all the evaluation statistics adopted here. The result shows that increasing the number of predictor variables does not necessarily enhance the performance of the ANN based RFFE models. The results demonstrate the potentials of ANN based RFFE models; however, further testing is needed using a larger data set before ANN based RFFE model can be recommended for practice in NSW.

Keywords: *Statistical Hydrology, RFFE, ANNs.*

1. INTRODUCTION

Flood is one of the worst natural disasters that cause significant damage to our society. To estimate design floods for ungauged catchments, regional flood frequency estimation (RFFE) methods are widely used in practice (Haddad et al., 2012). A RFFE technique attempts to transfer flood characteristics information from a homogeneous region to ungauged catchment location(s) of interest. Most of the RFFE techniques are based on linear modelling (e.g. Haddad et al., 2015; Micevski et al., 2015; Haddad and Rahman, 2012). The application of non-linear methods in RFFE such as artificial neural network (ANN) is quite few in numbers such as Aziz et al. (2014; 2015). To fill this current research gap, this study investigates the applicability of ANN based RFFE model for New South Wales (NSW) State in Australia.

The ANN is a data-based modeling technique, which was inspired by the working mechanism of human's natural neurons. ANN has been used frequently in the past few decades to solve complex mathematical problems. This was introduced by McCulloch and Pitts (1943). This technique has widely been adopted in medical and biomedical sciences (e.g. Baxt, 1990; Agatonovic-Kustrin and Beresford, 2000), and in different fields of engineering such as pattern recognition, forecasting, and data compression.

Lapedes and Farber (1987) first used ANN in modelling non-linear time series data. It has been found that ANN possesses better generalization capabilities than regression analysis and does not require the prescription of a mathematical functional form before the model building (Lek et al., 1996). ANN is capable of discovering complex non-linear connection between observed and predicted data sets in many complex problems within science and engineering fields (Hsu et al., 1995).

The aim of this paper is to explore the applicability of ANN in RFFE in New South Wales (NSW) State of Australia. This, in particular, examines the selection of an appropriate set of predictor variables in ANN based RFFE modelling in NSW, Australia.

2. METHODOLOGY

This paper uses ANN to develop RFFE models. Unlike most of the other data processing methods, ANN finds the patterns and correlations in a sample data automatically without the use of a prescribed model form. Like human neural system, an ANN is made of plenty of single units which are called artificial neurons. Figure 1 shows a biological neuron which consists of three main parts: cell body (sama), dendrite and axon. Dendrites receive the information and transfer it to the cell body, then the signal moves through the sama and outputs through the axon, which are connected to other dendrites via synapses, if the signal received from the previous neuron reaches a certain threshold, the next neuron is activated and the signal is transferred between the neurons. Similarly, artificial neurons process the inputs (X) and compare them with a given threshold (θ) to estimate the proper output.

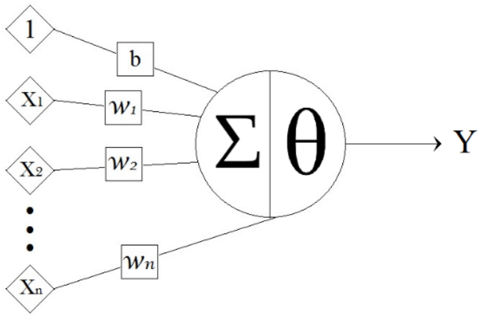

Figure 1 Schematic diagram of an artificial neuron

In Figure 1, X_i are the inputs and W_i are the specific weight matrix for each input, which determine the output by considering a given vector. It can be observed that there is a constant value of 1 among the inputs, which is introduced to the neuron by its unique weight b, known as bias. Bias allows the ANN to change the activation function. Consideration of bias for all networks is not essential, but it can improve the performance of a network significantly.

The mathematical relation of an artificial neuron can be shown as as below (Araghinejad, 2013):

$$I = W \times X + b \tag{1}$$

$$Y = \begin{cases} 1 \; if \; I \geq \theta \\ 0 \; if \; I \leq \theta \end{cases}$$

where X = inputs, W = weight matrix, b = bias, I = sum of all the weighted inputs, θ = threshold, and Y = output.

In this study, ANN is applied to develop prediction equation to estimate flood quantiles for six different annual exceedance probabilities (AEPs) i.e. 1 in 2, 1 in 5, 1 in 10, 1 in 20, 1 in 50 and 1 in 100 years. Model for each flood quantile is developed separately. Initially, for each of the selected gauged catchments, at-site flood quantiles of the 6 AEPs are estimated by using ARR FLIKE software

(Kuczera, 1999) by adopting LP3 distribution (with Bayesian method). The ANN based RFFE model prediction is compared with the at-site FLIKE estimates.

3. STUDY AREA AND THE DESCRIPTION OF DATA

A total of 88 small to medium sized gauging stations in NSW are selected for this study. The data preparation method of these catchments can be found in Rahman (2005), Haddad et al. (2010) and Rahman et al. (2009). The selected catchments are natural and free from major regulations. The catchments are smaller than 1000 km^2 with the median size of 260 km^2. The periods of annual maximum flood records range from 25 to 82 years. The following eight predictor variables are used in this study: (1) area of catchment (*AREA)* in km^2; (2) design rainfall intensity for 6-hour duration and 2-year return period (*I62*) in mm/h; (3) mean annual rainfall (MAR) (mm); (4) shape factor (*SF*); (5) mean annual areal evapo-transpirationation (MAE) in mm; (6) stream density (*SDEN*) in km/km^2; (7) main stream slope of the central 75% of the main stream (S1085) and (8) fraction forest (FOREST). The data were obtained from Australian Rainfall and Runoff (ARR) Project 5 (Rahman et al., 2016).

The selected 88 stations were sub-divided into 70% (62 stations) for training, 15% (13 stations) for validation, and 15% (13 stations) for testing. All catchments in each of these groups were chosen randomly. Also, 5 different models based on different combinations of the 8 catchment characteristics were selected. A three-layer feed forward neural network was considered for the prediction with 2 hidden layers. The Levenberg-Marquardt algorithm was chosen for training purposes, and Hyperbolic Tangent sigmoid function was used as the activation function. The modelling was done using MATLAB software.

The following statistical measures were used to evaluate the performance of the ANN based RFFE models using the set of validation catchments (consisting of 13 stations, selected randomly):

Root mean squared normalised error:

$$\text{RMSNE} = \sqrt{\frac{1}{n}\sum\left(\frac{\hat{Q}_i - Q_i}{Q_i}\right)^2} \tag{2}$$

Relative root mean squared error:

$$\text{RRMSE} = \frac{\sqrt{\frac{1}{n}\sum\left(\frac{\hat{Q}_i - Q_i}{Q_i}\right)^2}}{\bar{Q}} \tag{3}$$

Coefficient of determination:

$$R^2 = 1 - \frac{\sum(\hat{Q}_i - Q_i)^2}{\sum(\hat{Q}_i - \bar{Q}_i)^2} \tag{4}$$

Mean bias:

$$BIAS = \frac{1}{n}\sum\left(Q_i - \hat{Q}_i\right)^2 \tag{5}$$

Relative mean bias:

$$rBIAS = 100\,\frac{1}{n}\,\frac{\sum(Q_i - \hat{Q}_i)}{Q_i} \tag{6}$$

Absolute median relationship:

$$absRE = \left|\frac{\hat{Q}_i - Q_i}{Q_i}\right| \times 100 \tag{7}$$

Ratio between predicted and observed quantiles:

$$r = \frac{\hat{Q}_i}{Q_i} \tag{8}$$

Here $\hat{Q}_i$ and Q_i are respectively the predicted (from ANN) and observed quantiles (from FLIKE), $\bar{Q}_i$ is the average of the observed quantiles for any given period, and n is the number of the considered catchments. These evaluation statistics are adopted from Blöschl et al. (2013).

4. RESULTS

Five different ANN-based RFFE models were compared: Model 1 (all the 8 predictors); Model 2 (AREA, I62); Model 3 (AREA, I62, SF); Model 4 (AREA, I62, SF, MAR, MAE, S1085, FOREST) and Model 5 (AREA, I62, SF, SDEN).

The current ARR RFFE model for NSW region (Rahman et al., 2016) consists of three predictors, AREA, I62 and SF (i.e. similar to Model 3 here). One important question here is whether addition of a greater numbers of predictors increases the model accuracy or not. The results from independent testing (using the validation data set) of the ANN based RFFE models using 7 evaluation statistics are summarised in Tables 1 to 5.

Close examination of results in Tables 1 to 5 show that Model 1 (based on all the 8 predictors) performs very well with respect to the median relative error (RE) for all the AEPs except 1 in 2. Overall, Model 3 performs the best with respect to the RE. The performances of models are different across the six AEPs, i.e., for AEPs of 1 in 2 and 1 in 20, Model 3 is the best, but for AEPs of 1 in 5, 1 in 10, 1 in 50 and 1 in 100, Model 1 is the best performer. In terms of BIAS, Model 4 is the poorest model. In terms of coefficient of determination (R^2), Model 1 is the best model, except for 1 in 2 AEP which shows a relatively smaller R^2.

The best two models (Model 1 and Model 3) are compared in Figure 2 with respect to RMSE, which shows that these two models perform quite similarly i.e. there is little differences in the RMSE values for the two models across the six AEPs.

Table 1. Summary of ANN-based RFFE model performance based on independent testing (Model 1)

Quantile	Median (ABS) RE(%)	Median Ratio	R^2	RMSE	RMSNE	RRMSE	rRMSE	BIAS	rBIAS
Q2	61.31	0.71	0.42	76.50	5.28	0.76	151.47	-20.08	-125.13
Q5	23.39	0.96	0.71	133.79	1.51	0.50	527.68	-5.89	-2.67
Q10	10.25	1.04	0.75	189.37	1.17	0.46	270.89	13.66	18.83
Q20	34.06	0.79	0.64	304.97	2.25	0.52	420.77	-102.98	-26.97
Q50	33.99	1.03	0.72	376.37	1.65	0.45	137.24	5.10	29.04
Q100	33.09	1.04	0.52	636.20	2.05	0.61	505.32	3.32	29.22

Table 2. Summary of ANN-based RFFE model performance based on independent testing (Model 2)

Quantile	Median (ABS) RE(%)	Median Ratio	R^2	RMSE	RMSNE	RRMSE	rRMSE	BIAS	rBIAS
Q2	36.13	1.02	0.71	53.65	0.89	0.54	89.44	2.26	-22.09
Q5	32.93	1.09	0.73	127.81	1.23	0.48	123.06	10.63	-40.19
Q10	34.19	1.06	0.68	212.53	1.79	0.51	319.50	3.44	-59.29
Q20	39.50	1.20	0.57	337.10	1.30	0.58	168.51	-16.63	-30.76
Q50	38.09	1.12	0.59	461.41	1.87	0.55	348.80	-2.96	-56.57
Q100	47.72	0.89	0.05	894.86	2.47	0.85	609.85	311.68	-54.28

Table 3. Summary of ANN-based RFFE model performance based on independent testing (Model 3)

Quantile	Median (ABS) RE(%)	Median Ratio	R^2	RMSE	RMSNE	RRMSE	rRMSE	BIAS	rBIAS
Q2	29.48	1.02	0.77	48.32	1.18	0.48	117.71	-0.91	-40.08
Q5	52.24	0.64	0.22	218.10	0.98	0.82	98.07	110.80	12.49
Q10	33.23	1.00	0.72	197.63	0.75	0.48	56.86	26.56	-7.02
Q20	33.43	0.98	0.68	287.39	1.30	0.49	168.22	21.09	-24.73
Q50	38.90	1.12	0.36	575.39	2.25	0.69	505.42	44.02	-68.93
Q100	38.32	0.93	0.37	726.09	3.50	0.69	1223.31	187.66	46.73

Table 4. Summary of ANN-based RFFE model performance based on independent testing (Model 4)

Quantile	Median (ABS) RE(%)	Median Ratio	R^2	RMSE	RMSNE	RRMSE	rRMSE	BIAS	rBIAS
Q2	38.28	1.23	0.72	52.62	3.29	0.53	329.24	-97.66	329.24
Q5	79.96	1.36	0.64	316.98	5.90	1.19	590.09	-12.04	590.09
Q10	29.84	1.09	0.62	231.52	3.25	0.56	1057.57	-59.87	1057.57
Q20	46.99	1.11	0.29	432.04	5.27	0.74	2777.88	-122.62	2777.88
Q50	30.69	1.07	0.55	483.04	2.07	0.58	429.41	-66.57	429.41
Q100	35.42	1.24	0.38	719.13	4.86	0.68	2366.70	-62.77	2366.70

Table 5. Summary of ANN-based RFFE model performance based on independent testing (Model 5)

Quantile	Median (ABS) RE(%)	Median Ratio	R^2	RMSE	RMSNE	RRMSE	rRMSE	BIAS	rBIAS
Q2	319.61	2.29	-9.77	329.04	27.88	3.29	2788.22	-174.27	53.26
Q5	84.89	0.99	0.04	242.55	7.67	0.91	767.27	-8.30	73.48
Q10	44.94	0.94	0.45	278.34	1.49	0.67	221.19	48.93	-15.29
Q20	38.44	0.98	-0.03	520.12	1.89	0.89	356.23	109.56	-20.44
Q50	46.69	0.76	0.39	559.18	4.84	0.67	2344.07	213.44	112.79
Q100	67.21	1.01	-0.21	1007.54	7.60	0.96	5769.06	164.90	-170.32

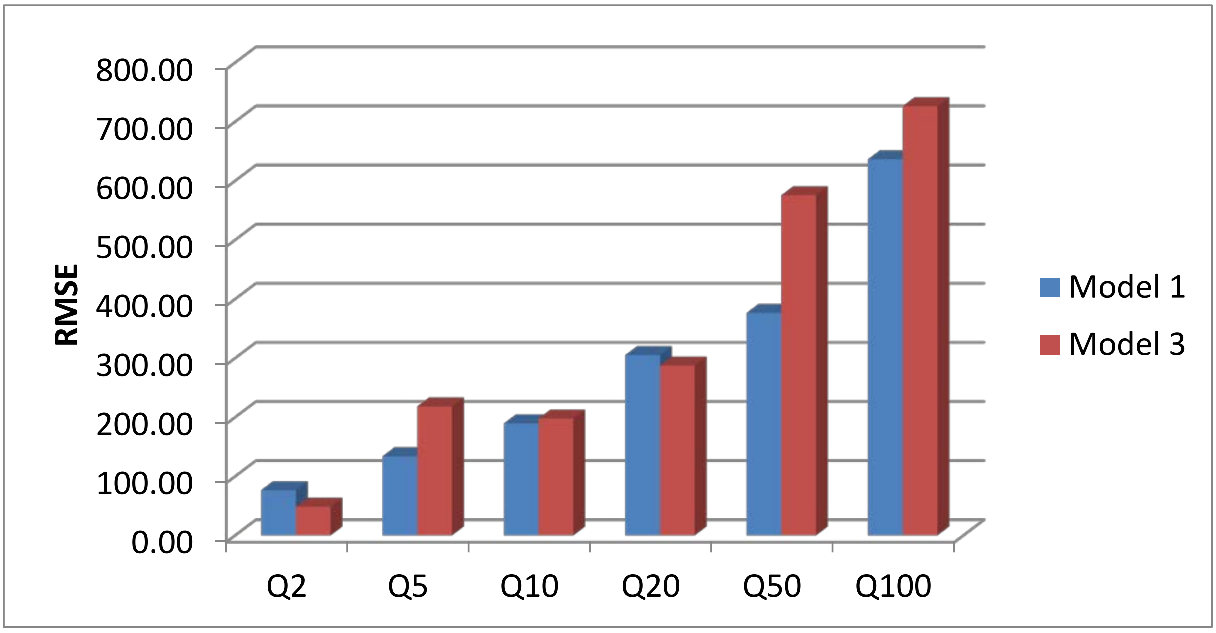

Figure 2 Comparison of RMSE values for Models 1 and 3

5. CONCLUSION

ANN-based RFFE model is developed using data from 88 NSW catchments. Five different combinations of eight predictor variables are compared. A split-sample validation technique is adopted to compare the model performance based on 9 different statistics. It has been found that when all the eight predictor variables are used, the ANN based RFFE model performs the best generally; however, the gain in model accuracy from using only three predictor variables is marginal. Furthermore, the performances of ANN based RFFE models vary across the six AEPs, and there no model that is the best with respect to all the evaluation statistics adopted here. The result shows that increasing the number of predictor variables do not necessarily enhance the performance of the ANN based RFFE models. The results demonstrate the potentials of ANN based RFFE models; however, further testing is needed using a larger data set before ANN based RFFE model can be recommended for practice in NSW.

ACKNOWLEDGMENTS

Authors acknowledge ARR Project 5 team, Engineers Australia, Australian Bureau of Meteorology and Department of Water, NSW for sharing the data used in this research.

REFERENCES

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717-727.

Araghinejad, S. (2013). *Data-driven modeling: using MATLAB® in water resources and environmental engineering* (Vol. 67). Springer Science & Business Media.

Aziz, K., Rahman, A., Fang, G., & Shrestha, S. (2014). Application of artificial neural networks in regional flood frequency analysis: a case study for Australia. *Stochastic Environmental Research and Risk Assessment*, *28*(3), 541-554.

Aziz, K., Rai, S., & Rahman, A. (2015). Design flood estimation in ungauged catchments using genetic algorithm-based artificial neural network (GAANN) technique for Australia. *Natural Hazards*, *77*(2), 805-821.

Baxt, W. G. (1990). Use of an artificial neural network for data analysis in clinical decision-making: the diagnosis of acute coronary occlusion. *Neural Computation*, *2*(4), 480-489.

Blöschl, G. (Ed.). (2013). *Runoff prediction in ungauged basins: synthesis across processes, places and scales*. Cambridge University Press.

Haddad, K., Rahman, A., Ling F. (2015). Regional flood frequency analysis method for Tasmania, Australia: a case study on the comparison of fixed region and region-of-influence approaches, Hydrological Sciences Journal, 60, 12, 2086-2101.

Haddad, K. & 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, 430-431 (2012), 142-161.

Haddad, K., Rahman, A., & 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. & 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.

Hsu, K. L., Gupta, H. V., & Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research*, *31*(10), 2517-2530.

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

Lapedes, A., & Farber, R. (1987). *Nonlinear signal processing using neural networks: Prediction and system modelling* (No. LA-UR-87-2662; CONF-8706130-4).

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, *90*(1), 39-52.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115-133.

Micevski, T., Hackelbusch, A., Haddad, K., Kuczera, G., & Rahman, A. (2015). Regionalisation of the parameters of the log-Pearson 3 distribution: a case study for New South Wales, Australia, Hydrological Processes, 29, 2, 250-260.

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. & 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, 181 pp.

Rahman, A., Haddad, K., Kuczera, G., & Weinmann, P.E. (2016). Regional flood methods. In: Australian Rainfall & Runoff, Chapter 3, Book 3, edited by Ball, Engineers Australia, 78-114 (In press).