Application of Generalized Additive Models in Regional Flood Frequency Analysis: A Case Study for Victoria, Australia

Farhana Noor[1] and Ataur Rahman[2]

[1]Master of Philosophy Student, Western Sydney University, Sydney, Australia
[2]Associate Professor, Western Sydney University, Sydney, Australia Corresponding Author's E-mail: fnoor1989@gmail.com

Abstract

Design flow estimation for ungauged catchments is considered to be a challenging task in hydrology. Regional Flood Frequency Analysis (RFFA) can be used to estimate design flow for ungauged catchments. Commonly adopted RFFA methods include the index flood method, the rational method and the quantile regression technique. This paper examines the applicability of the generalized additive model (GAM) in RFFA. GAM establishes a well parameterized modelling framework which describes the multivariate and nonlinear characteristics of a complex dataset. This approach allows flexible specification of regression splines to represent the functional relationships between a response variable (i.e. flood quantile) and a suite of temporal and spatial covariates that can be continuous or discrete. This is done using a link function and smooth functions of the covariates such as catchment characteristics. In this study, both the GAM and log-linear model (LLM) are applied in RFFA to a data set of 114 catchments from Victoria State in Australia. Based on independent testing, it was found that GAM generally provides more accurate flood quantile estimates than the LLM. Further study is needed to confirm this finding.

Keywords: Regional Flood Frequency Analysis, GAM, Ungauged Catchment, log-linear model.

1. INTRODUCTION

Floods are one of the most common natural hazards and are considered to be one of the costliest disasters causing billions of dollars of damage globally. Floods cause loss of lives, economic damage and undermine societal wellbeing. The detrimental impacts of floods generally depend on the hydrologic, geomorphological and meteorological characteristics of the catchments and precipitation events that cause floods. The average annual flood damage is worth over $377 million and infrastructure requiring design flood estimate is over $1 billion per annum in Australia (BITRE, 2001). Therefore, reliable design flood estimation has been an important aspect of resource management and infrastructure design in Australia similar to other countries.

Design flood estimation for ungauged catchments has been a difficult task and is generally associated with a large degree of uncertainty (Haddad and Rahman, 2012). For ungauged catchments, design floods can be estimated by Regional Flood Frequency Analysis (RFFA), which consists of two principal steps, formation of homogeneous regions, and development of flood estimation equations. Various forms of RFFA techniques have been proposed in the literature (e.g. index flood method by Hosking and Wallis, 1993; Generalised Least Squares based Quantile Regression Technique by Griffis and Stedinger, 2007; and parameter regression technique by Haddad et al., 2012). In many of the previous RFFA studies, regions were formed based on geographic or administrative boundaries, often lacking hydrological similarities (Bates et al., 1998; Rahman et al., 2017). Most of the previous RFFA models are based on linearity assumptions although rainfall-runoff process is generally non-linear. Generalized Additive Model (GAM) allows complex variable transformation for the predictor variables in regression modelling to account for the non-linearity of the flood generation process. GAM has not been tested widely in RFFA, particularly in Australia. Hence, the objective of this present study

is to test the applicability of GAM in RFFA using a dataset from the State of Victoria, Australia.

2. MATERIALS AND METHODS

2.1. Log-linear Model

A multiple linear regression is often used to develop relationship between a dependent variable, *Y* and *p* predictor variables, $X_1, X_2, \ldots, X_p$. This can be expressed for i^{th} observation as below:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \quad (1)$$

where β_0 and β_j (j = 1, 2, …, p) are unknown parameters (regression coefficients) and ε_i is the error term associated with i-th observation (i = 1, 2, …, n), where n = number of observations. The error term in Eq. 1 is assumed to be normally distributed $N(0, \sigma^2)$ and the model parameters are generally estimated by the method of least squares. In RFFA, a log-linear model (LLM) is widely adopted where both the dependent and predictor variables are log-transformed before building the regression model under the assumption that it will achieve normality of the predictors and linearity between the dependent variable and predictor variables. A LLM can be presented as below:

$$ln(y_i) = \beta_0 + \sum \beta_j ln(X_j) + \varepsilon_i \quad (2)$$

2.2. Generalized Additive Model

Generalized Additive Model (GAM) (Hastie and Tibshirani, 1986; Wood, 2006) allows non-linear functions of each of the predictor variables, while maintaining the additivity of the model, which is achieved by replacing each linear component in Eq. 1 $\beta_j X_{ij}$ by a smooth non-linear function $fj(X_{ij})$. A GAM can then be written as:

$$y_i = \beta_o + \sum_{j=1}^{p} fj(x_{ij}) + \varepsilon_i$$
$$= \beta_o + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + \ldots\ldots + f_p(x_{ip}) + \varepsilon_i \quad (3)$$

GAM allows fitting a non-linear function f_j to each X_j i.e. manually trialing process of numerous transformations on each of the predictor variables can be avoided. Additionally, GAM allows checking the impact of each X_j on Y individually. In this model, the smoothness of function f_j for the variable X_j is summarized via degrees of freedom. In GAM, the linear predictor predicts a known smooth monotonic function of the expected value of the response, and the response may follow any distribution from exponential family or may have a known mean variance relationship, allowing a quasi-likelihood approach (Wood, 2006).

In GAM, to estimate the smooth function f_j a spline may be adopted. A number of spline types are available (e.g. *P*-splines, cubic splines and *B*-splines). In this study, thin plate regression splines are adopted as they provide fast computation, and do not require selection of knot locations and have optimality in approximating smoothness (Wood 2003, 2006).

2.3 Data Selection

This study uses flood and catchment data from 114 catchments of Victoria State, Australia (Fig1). These catchments are subset of Australian Rainfall Runoff Project 5 database (Rahman et al., 2016). These catchments have not undergone any major land use change and are not affected by any major regulation during the period of streamflow data availability. The catchment area of the selected 114 catchments range from 3 to 997 km^2 (mean: 317.5 km^2 and median: 270.5 km^2). The annual maximum

streamflow record length of selected stations varies from 26 to 62 years, with a mean of 38-years, median of 39 years and standard deviation of 5years. The dependent variables are flood quantiles with annual exceedance probabilities (AEPs) of 1 in 2, 1 in 5, 1 in 10, 1 in 20, 1 in 50 and 1 in 100, represented respectively by Q_2, Q_5, Q_{10}, Q_{20}, Q_{50}, Q_{100}. The physiographical and meteorological variables (predictor), available for each catchment, are summarized in Table 1.

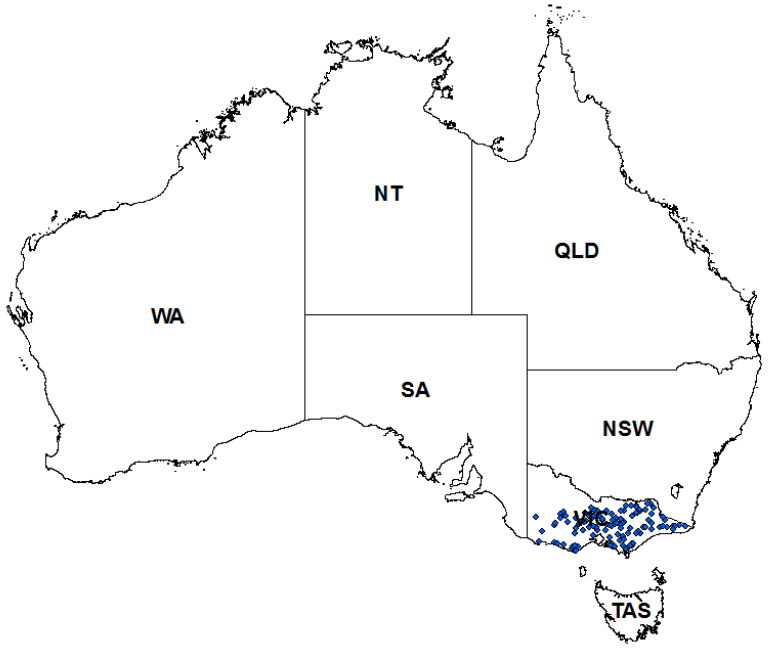

Figure 1. Location of the selected 114 catchments in Victoria, Australia.

2.4 Model Development

Log-linear Model

In development of the log-linear model (LLM), both the response variables (Q_2, Q_5, Q_{10}, Q_{20}, Q_{50}, Q_{100}) and predictor variables are log-transformed. Predictor variables in the regression model are selected using a backward stepwise selection procedure. For the LLM, only four predictor variables are found to be statistically significant: AREA, $I_{6,2}$, SF and SDEN.

The developed LLM can be written by:

$$\ln(Q) = bo + b_1 \ln(AREA) + b_2 \ln(I_{6,2}) + b_3 \ln(SF) + b_4 \ln(SDEN) \quad (4)$$

Generalized Additive Model

In the development of GAM model, predictor variables are selected based on the backward stepwise procedure for each of the quantiles. Five predictor variables are found to be statistically significant; AREA, $I_{6,2}$, RAIN, SDEN and EVAP. During building of the prediction equation in GAM, the "Gaussian family" is adopted with 'identity' link function as this is the most common approach.

The general form of the developed prediction equation in GAM is given by:

$$\ln(Q) = \alpha + s(AREA) + s(I_{6,2}) + s(RAIN) + s(EVAP) + s(SDEN) \quad (5)$$

Table 1. Descriptive statistics of the selected predictor variables for the 114 catchments in Victoria, Australia

Variable	Unit	Notation	Min	Mean	Max	SD
Catchment area	km^2	AREA	3	317.54	997	244.65
Catchment shape factor	-	SF	0.281	0.79	1.4341	0.22
Main stream slope	m/km	S1085	0.8	13.38	69.9	12.30
Stream density	km/km^2	SDEN	0.52	1.53	4.25	0.53
Fraction of catchment covered by forest	-	FOREST	0.01	0.59	1	0.35
Rainfall intensity (6-h duration and 2-year return period)	mm/h	$I_{6,2}$	24.6	34.29	46.7	5.27
Mean annual rainfall	mm	RAIN	484.39	931.64	1760.81	319.01
Mean annual potential evapotranspiration	mm	EVAP	925.9	1035.47	1155.3	42.80

2.5 Model validation

The model performance is evaluated by a 10-fold cross validation method. In this approach, the dataset is randomly divided into modelling and test sub-sets. The model is calibrated on the modelling sub-set and the model is tested on the test sub-set (Haddad et al., 2013). The following statistical measures are adopted in this study to assess the model accuracy:

Coefficient of determination,
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \quad (6)$$

Root mean square error,
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2} \quad (7)$$

Relative root mean square error,
$$rRMSE = 100\sqrt{\frac{1}{n}\sum_{i=1}^{n}[(z_i - \hat{z}_i)/z_i]^2} \quad (8)$$

Mean bias,
$$BIAS = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i) \quad (9)$$

Relative mean bias,
$$rBIAS = 100\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)/z_i \quad (10)$$

where z_i and $\hat{z}_i$ are respectively the local (at site) and regional flood quantile estimates at catchment *i*, $\bar{z}$ is the local mean of flood quantile (for a given return period) and *n* is the number of catchments in the test data set.

3. RESULTS AND DISCUSSION

The predicted and observed flood quantiles for 2, 10 and 50 year return periods are presented in Figure 2 for both the LLM and GAM. From these plots, it can be seen that GAM generally provides a better match between the observed and predicted flood quantiles. However, there is a remarkable degree of scatter for large flood quantiles.

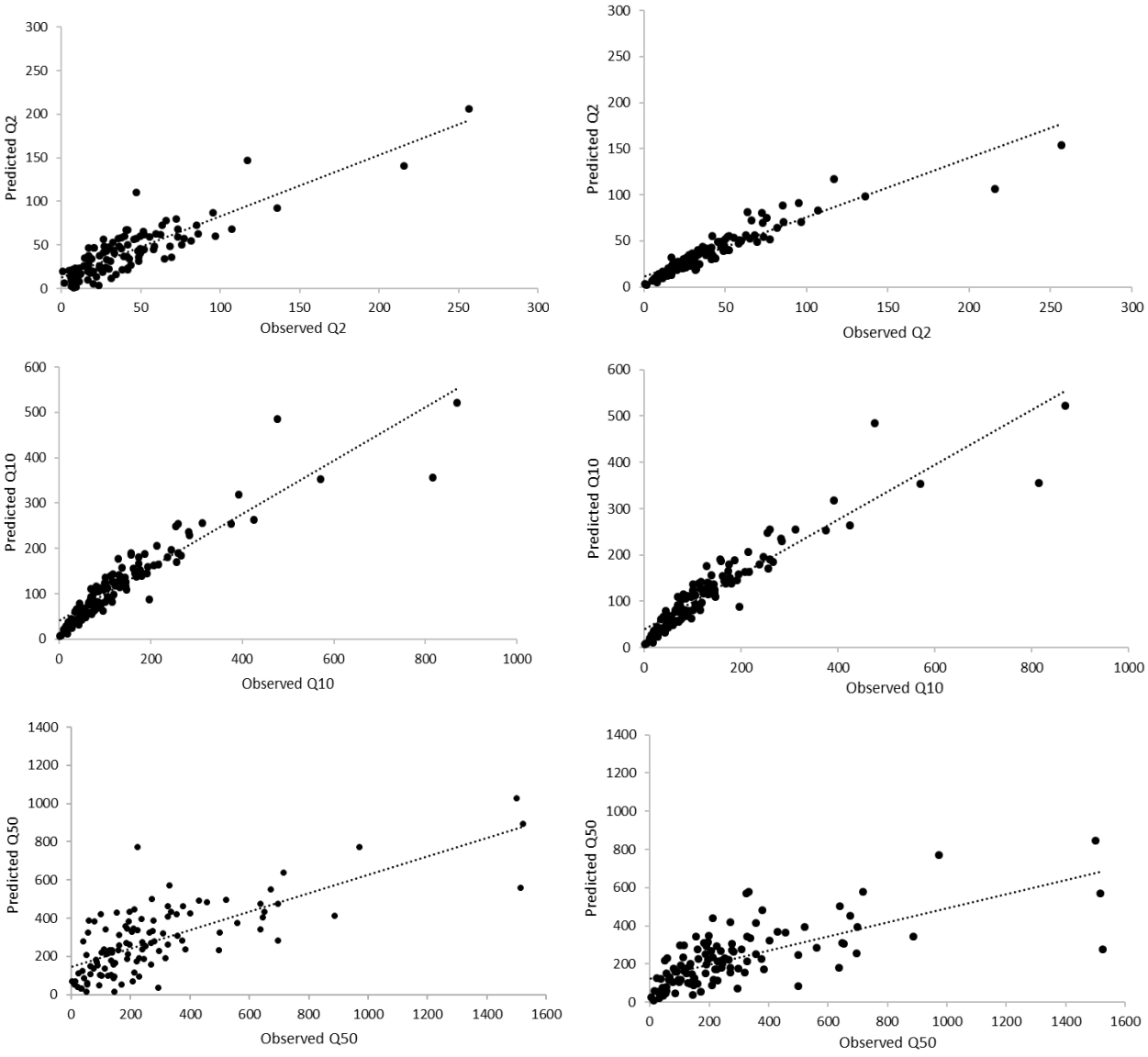

Figure 2. Observed versus predicted flood quantiles (Q_2, Q_{10}, Q_{50}) for GAM (left side) & LLM (right side)

Table 2 summarizes the results of the 10-fold cross validation for both the GAM and LLM. It can be seen from this table that the GAM has much higher R^2 values than the LLM. For both the LLM and GAM, R^2 values reduce with increasing return period as expected. The BIAS is relatively smaller for the GAM than the LLM. The RMSE values are also smaller for the GAM than the LLM. For rBIAS, the LLM has smaller values than the GAM for all the return periods except Q_{10}. For rRMSE, the LLM has smaller values than the GAM for all the return periods. Overall, the GAM outperforms the LLM in predicting the flood quantiles.

Table 2. Comparison of LLM and GAM in RFFA based on a 10-fold cross validation

		Q_2	Q_5	Q_{10}	Q_{20}	Q_{50}	Q_{100}
R^2	LLM	0.502	0.466	0.484	0.389	0.417	0.385
	GAM	0.737	0.7	0.676	0.627	0.537	0.495
BIAS	LLM	-4.275	-10.83	-17.467	-30.718	-48.48	-68.96
	GAM	6.023	1.23	-8.93	1.442	3.44	-4.186
RMSE	LLM	26.102	64.8	67.007	149.97	214.462	284.616
	GAM	18.997	48.5	38.271	117.23	191.246	257.991
rBIAS	LLM	13.9	16.1	52.734	23.26	26.259	30.114
	GAM	39.952	49.7	5.029	57.27	60.748	68.787
rRMSE	LLM	69.913	83.4	163.346	103.613	107.965	117.574
	GAM	215.258	266.	241.98	256.124	218.276	228.192

4. CONCLUSION

This study compares GAM and LLM in RFFA using data from 114 catchments in Victoria. The main advantage of GAM is that it allows non-linear variable transformation in regression modelling relatively easily. From this preliminary analysis, it has been found that GAM can be applied successfully in RFFA. Based on independent testing, it was found that the predicted and observed flood quantiles have a closer match for the GAM than the LLM. The GAM also shows higher R^2, smaller BIAS and RMSE as compared with the LLM. However, for rBIAS and rRMSE, the LLM generally provides better model validation results. Further study is needed to compare these two models in RFFA such as examination of outlier catchments and boxplots of relative errors to confirm that the GAM fits the observed flood data better.

ACKNOWLEDGMENTS

The authors would like to thank ARR Revision Project 5 team for providing the data for this study.

REFERENCES

Bureau of Infrastructure, Transport and Regional Economics (BITRE) (2001). Economic costs of natural disasters in Australia, Commonwealth of Australia, Canberra.

Bates BC, Rahman A, Mein RG, Weinmann PE (1998). Climatic and physical factors that influence the homogeneity of regional floods in southeastern Australia. Water Resources Research, 34(12), 3369-3381.

Griffis VW, Stedinger JR (2007). The use of GLS regression in regional hydrologic analyses. Journal of Hydrology, 204, 82-95.

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 JR (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, Zaman M, Shrestha S (2013). Applicability of Monte Carlo cross validation technique for model development and validation using generalized least squares regression, Journal of Hydrology, 482, 119-128.

Hastie T, Tibshirani R (1986). Generalized additive models. Statistical Science, 1, 297–310.

Hosking JRM, Wallis JR (1993). Some statistics useful in regional frequency analysis. Water Resources Research, 29(2), 271-281.

Rahman A, Haddad K, Kuczera G, Weinmann PE (2016). Regional flood methods. In: Australian Rainfall & Runoff, Chapter 3, Book 3, edited by Ball et al., Commonwealth of Australia.

Rahman A, Charron C, Ouarda TBMJ, Chebana F (2017). Development of regional flood frequency analysis techniques using generalized additive models for Australia. Stochastic Environmental Research and Risk Assessment, 1-17. doi:10.1007/s00477-017-1384-1.

Wood SN (2003). Thin plate regression splines. Journal of the Royal Statistical Society, 65, 95–114.

Wood SN (2006). Generalized additive models: An introduction with R. CRC Press, Florida.