Comparison between Quantile Regression Technique and Generalised Additive Model for Regional Flood Frequency Analysis

FARHANA NOOR

STUDENT ID: 18815782

A thesis submitted for the fulfilment of Master of Philosophy Degree in Western Sydney University, Australia

Principal Supervisor: Associate Professor Dr Ataur Rahman

Co-supervisor: Professor Dr Arumugam Sathasivan

Centre for Infrastructure Engineering

School of Computing, Engineering and Mathematics

Western Sydney University

June 2018

ABSTRACT

Design flood estimates are needed for the planning and design of hydraulic structures, and in many other water and environmental management tasks. Design flood estimation is a challenging task, in particular for poorly gauged and ungauged catchments. In Australia, there are numerous ungauged catchments; for these catchments Regional Flood Frequency Analysis (RFFA) techniques are generally adopted to estimate design floods.

Most of the RFFA techniques previously adopted in Australia are based on rational method and/or linear modelling approaches. However, with the recent advancements in statistical computation methods, there are several other techniques becoming popular gradually in hydrological applications which can account for non-linearity in the rainfall-runoff processes. Generalized additive model (GAM) is one of the recently developed techniques which can deal with the non-linearity, which has not been widely explored in hydrological research, in particular for the RFFA problems. Therefore, this research is devoted to examining the applicability of GAM in RFFA and compare its performances with one of the most widely used linear RFFA technique (log-log linear model).

This study is carried out using data from 114 small to medium sized gauged catchments of Victoria, Australia. This data has primarily been sourced from Australia Rainfall Runoff (ARR), Project 5 Regional Flood Methods. This study is based on a number of alternative groups, e.g. a combined group consisting of all the 114 catchments and sub-groups formed based on cluster analysis. Four regions are formed using hierarchical and k-means clustering techniques. All the five groups are used for developing log-log linear models and GAM based models. The predictor variables for each of these models are selected based on the statistical significance of the predictor variables, i.e. *p*-statistics. For validation of the developed prediction models, a 10-fold cross validation method is adopted.

The performances of the prediction models for the alternative models are assessed using a number of statistical measures including coefficient of determination (R^2), median relative error (RE) and median Q_{pred}/Q_{obs} ratio values. It is found that, none of the models from the combined group and clustering groups perform equally well for the six average recurrence intervals (ARIs) (2, 5, 10, 2, 50 and 100 years) with respect to the selected statistical measures. Overall, log-log linear model from clustering group A1 is found to be the best

performing model. GAM based RFFA models perform better for smaller ARIs (i.e., 2, 5 and 10 years); which is as expected since the hydrological behaviour of catchments for smaller ARIs is generally more non-linear, e.g. higher loss and hence rainfall produces lower runoff for more frequent events.

Some predictor variables (e.g., *evap*), which were not adopted in the previous RFFA models, in Australia are found to be significant in the GAM based RFFA models. Overall, it is found that consideration of non-linearity via GAM can add new dimensions in RFFA modelling for selecting appropriate predictor variables and to deal with non-linearity.

Overall, the results of this study demonstrate that GAM has a strong potential to enhance the accuracy of RFFA models in Australia; however, additional predictor variables are needed (than what are included in this study) to capture the non-linearity more explicitly between runoff and flood producing variables.

STATEMENT OF AUTHENTICATION

I, Farhana Noor, declare that all the materials presented in this Master of Philosophy Thesis entitled 'Comparison between Quantile Regression Technique and Generalised Additive Model for Regional Flood Frequency Analysis' are my own work and that any work adopted from other sources is duly cited and referenced as such.

This thesis contains no material that has been submitted previously, in whole or in part, for any award or degree in other university or institution.

Signed by:

Farhana Noor

Date: 27/06/2018

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and thankfulness to my Principal Supervisor, Associate Professor Dr. Ataur Rahman, for the continuous support, guidance and invaluable assistance. He has been a great inspirer for me throughout my MPhil study. This work would not have been possible without his support, encouragement and most importantly the patience during this research. I am also grateful to my Co-supervisor Professor Dr. Arumugam Sathaa Sathasivan and HDR Director, Associate Professor Dr. Dongmo Zhang for their valuable support towards the completion of my research degree. I could not be prouder of my academic roots, and hope that I live up to the research values and the dreams that my supervisors have passed to me.

I would like to thank Australia Rainfall Runoff Project 5 Revision Team to provide me useful data to conduct this study. I would also like to thank to all the staff and fellow researchers at Western Sydney University for their kind support and assistance throughout my research study.

I would not have completed this road if not for my parents, Engr. Mr. Nurul Alam and Advocate Mrs. Zakri Jahan, who instilled within me a quest for knowledge, which has driven me to continue with the research.

Last but not the least, I heart-fully thank and appreciate the constant support and patience of my loving husband Engr. Maruf Adnan during this study. His endless patience and care during the journey will never be forgotten. It would not have been travelled the road to MPhil degree without his constant inspiration.

LIST OF PUBLICATIONS

1. Noor, F., Rahman, A. (2017). Application of Generalized Additive Models in Regional Flood Frequency Analysis: A Case Study for Victoria, Australia, Proceedings of the 1st International Conference on Water and Environmental Engineering, pp. 74-80, 20-22 Nov 2017, Sydney, Australia.

2. Noor, F., Rahman, A. (2018). Comparison of log-log linear model with Generalised Additive model for Regional Flood Frequency Analysis for Victorian Catchments. (In preparation for Natural Hazards)

3. Noor, F., Rahman, A. (2018). Validation of Generalised Additive model for Regional Flood Frequency Analysis in Victoria, Australia (In preparation for 2nd International Conference on Water and Environmental Engineering, 19-23 Jan, 2019, Dhaka, Bangladesh.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. General

The thesis focuses on the applicability of Generalized Additive Model (GAM) for regional flood estimation. The performance of GAM is compared with the widely used log-log linear model for design flood estimation in ungauged catchments. This chapter begins by presenting a background to this research, need for this research, research questions to be investigated and research tasks undertaken and an outline of this thesis.

1.2. Background of the proposed research

Flood is considered as one of the costliest and disturbing natural disasters. Floods cause loss of lives, economic damage and undermine societal wellbeing (Rahman, 2017). The detrimental impacts from floods can be even worse due to the negative geomorphological impacts of floods, e.g. erosion, sedimentation and destruction of vegetation and wild life.

Flooding aftereffects can be substantial on both spatial and temporal scale. In the period 1852 to 2011, 951 people were killed and another 1326 injured by floods in Australia (Carbone and Hanson, 2013). The average annual flood damage is worth over $377 million and infrastructure requiring design flood estimate is over $1 billion per annum in Australia (Gentle et al., 2001). The state of New South Wales (NSW) alone has an average annual cost of flood damage of over $172 million, which is almost 46% of the average annual flood damage cost for Australia. The state of Queensland is second largest in terms of flood damage, with an average annual cost of $125 million. Importantly, the 2010-11 devastating flood in Queensland caused flood damage over $5 billion (Queensland Reconstruction Authority, 2011).

Floods in Australia are triggered by several causes which include excessive precipitation, infrastructure failures and cyclonic effects. Other associating factors that act as drivers to determination of flood magnitudes include catchment and land use characteristics. Rapid

urbanisation, infiltration of waterbodies and land encroachment increase the risk of flooding in a given catchment. Flooding often emerges as a serious threat to livelihoods and infrastructure systems in urban areas due to rapid increase in runoff volume due to larger impervious area and shorter response time. Moreover, climate change has a tremendous impact including more frequent extreme rainfall events resulting in increased flood risk (Ishak et al., 2013).

Considering the aftermaths of flooding and to ensure the accuracy of a flood forecasting system, the development of a dependable flood risk assessment technique is very important in order to reduce the flood damage cost (Caballero and Rahman, 2014). To develop a reliable flood risk assessment technique, improved methods as well as adequate flood and rainfall data are needed. Flood damage can be reduced if design floods can be estimated more accurately. A well-designed flood infrastructure largely depends on the accuracy of design flood estimation.

Design floods can be defined as the flood discharge associated with a given annual exceedance probability (AEP). Design flood estimation is required in numerous engineering applications, e.g. design of bridge, culvert, weir, spill way, detention basin, flood protection levees, highways, floodplain management, flood insurance studies and flood damage assessment tasks (Aziz et al., 2014). In order to estimate design floods, the most common method used is flood frequency analysis, which requires recorded streamflow data of adequate length at the selected catchment. The accuracy of flood frequency analysis results largely depends on availability of good quality flood data in terms of data quality and quantity. From a statistical point of view, flood estimation from a small sample may give unreasonable or physically unrealistic parameter estimates, especially for probability distributions with a large number of parameters (three or more).

Flood estimation of data poor regions has become a considerable issue in recent years due to effects of some devastating floods in Australia. There are several regional flood estimation methods which have been adopted over the years to estimate the design floods for ungauged catchments. These include Index Flood Method, the Rational Method and Quantile Regression Technique. Regional flood frequency analysis (RFFA) has been considered as one of the efficient methods to ascertain the design flood estimation in data poor regions and

ungauged catchments. This research focuses on regional flood estimation in order to enhance the accuracy of design flood estimates.

Design flood estimation is widely used in practice. At-site flood frequency analysis is used if streamflow data of longer length (generally over 20 years) is available. In many instances, recorded streamflow data is absent or of limited length, and under these circumstances, regional flood estimation methods are adopted. ARR1987 recommended Probabilistic Rational Method in some Australian states. ARR2016 has recommended the RFFE model which is based on regional LP3 distribution where its parameters are estimated using GLS regression. Also, in ARR2016 regions are formed using a region-of-influence approach in the data-rich regions of Australia.

Most of the above RFFA approaches are linear methods, i.e. they cannot incorporate the nonlinearity between floods and flood producing variables. In this regard, GAM can be adopted which can account for the nonlinearity (e.g.,Asquith et al., 2013; Chebana. et al., 2014; Rahman. et al., 2018). In Australia, there has been limited application of GAM in RFFA e.g. Rahman et al. (2018) applied GAM to New South Wales (NSW) state. Hence, this thesis aims to test the applicability of GAM in RFFA to a new region of Australia, which is the state of Victoria. This also compares the performance of GAM based RFFA models with log-log linear models for Victoria.

1.3. Research questions

This thesis is devoted to answering the following research questions in relation to the development of GAM based RFFA models for Victoria.

- Whether the Generalized Additive Model can produce more accurate regional flood estimates as compared to the log-log linear model?

- What is the best set of predictor variables for the development of log-log linear model and GAM based RFFA models?

- Whether cluster analysis can result in better regions for RFFA and reduce uncertainty in RFFA?

1.4. Overview of adopted methodology

To answer the above research questions (identified in Section 1.4), the following tasks are carried out in this study:

- A critical literature review on the most commonly used RFFA and GAM based methods to identify the gaps in the current state of knowledge and further research opportunities in RFFA.

- Selection of catchments from Victoria, collation of streamflow data, selection of catchment characteristics that govern flood generation process and preparation of climatic and catchment characteristics data set.

- Selection of the best performing set of predictor variables for the log-log linear model and GAM based RFFA models.

- Comparison of different candidate regions based on catchment characteristics data using cluster analysis and identification of the best performing region(s) for log-log linear model and GAM based RFFA model.

- Comparison of the performance of the log-log linear model and GAM using a set of independent test catchments.

Figure 1.1 below presents a flow chart illustrating the major tasks involved in this study.

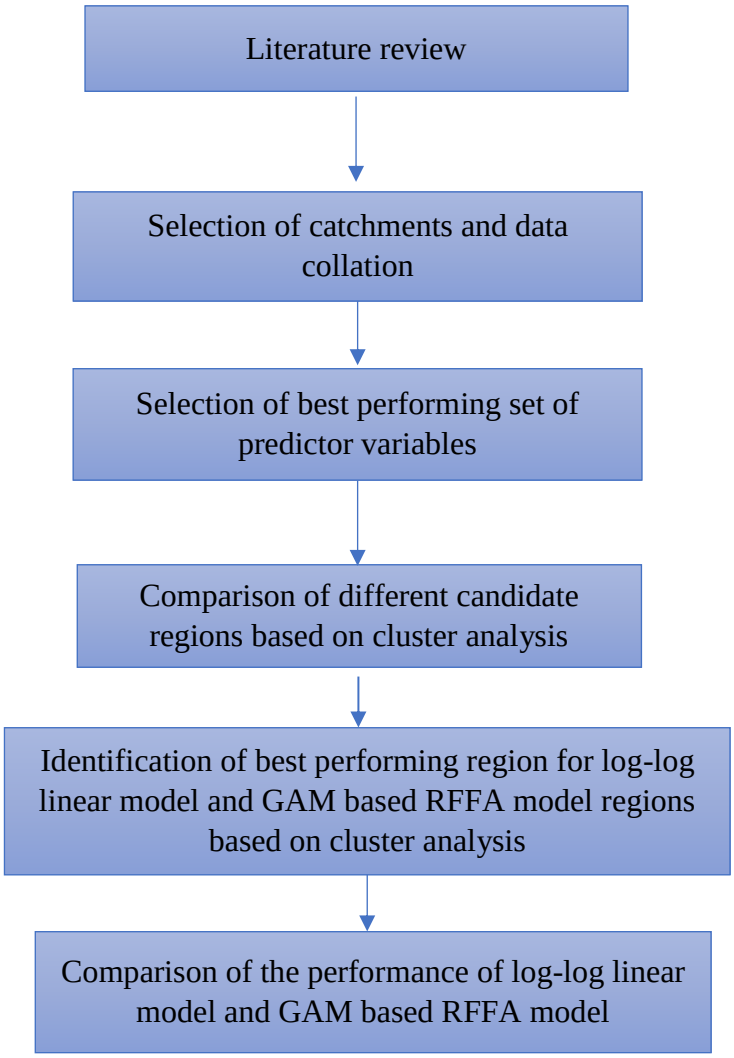

Figure 1.1 Flow chart showing major tasks in this research

1.5. Outline of the thesis

The research undertaken in this study is presented in this thesis in eight chapters and four appendices, as outlined below.

Chapter 1 presents a brief introduction to the overall study, includes a background of the proposed research. This chapter also presents the needs for this research, research questions being examined and the main research tasks undertaken to answer the identified research questions.

Chapter 2 contains a critical review on RFFA techniques with a particular emphasis on GAM, log-log linear model and cluster analysis. At the beginning, various methods of flood estimation are discussed. The review of linear RFFA methods including rational method, index flood method and quantile regression technique are then presented. The GAM is then discussed with a particular emphasis on their applications to hydrology. The assumptions, limitations, advantages and disadvantages of each of the RFFA methods are discussed. The current state of knowledge in RFFA is ascertained and the scopes of further research are identified.

Chapter 3 presents the study area and data collation including data exploration and correlation analysis. The methods of streamflow data preparation are discussed which include gap filling, outlier detection, trend analysis and rating curve error analysis. Selection of catchment characteristics are then presented. The preparation of annual maximum flood series data is described thereafter. Estimation of flood quantiles for average recurrence intervals of 2, 5, 10, 20, 50 and 100 years for the selected gauged catchments by at-site flood frequency analysis is then presented. Finally, a summary of the catchment characteristics data is provided.

Chapter 4 presents the adopted methodologies i.e. GAM, log-log linear model and cluster analysis.

Chapter 5 presents the results of selecting the best set of predictor variables for the development of log-log linear model considering the combined and grouped datasets.

Chapter 6 presents results of selecting the best set of predictor variables for the development of GAM based RFFA models considering combined and grouped datasets.

Chapter 7 presents the comparison of GAM and log-log linear models.

Chapter 8 presents the summary of the research undertaken in this thesis, conclusions and recommendations for further research.

CHAPTER 2

REVIEW OF REGIONAL FLOOD FREQUENCY ANALYSIS METHODS

2.1. General

Regional flood frequency analysis (RFFA) refers to a generic method of design flood estimation at a target catchment (usually ungauged) by utilizing streamflow records pooled from several other catchments which have similar characteristics with the target catchment. There are many RFFA techniques ranging from simple approximate methods to complex intelligence-based techniques. The purpose of this chapter is to review the concepts of RFFA focusing on estimation of design floods in the range of average recurrence intervals (ARIs) of 2 – 100 years based on linear methods (e.g., quantile regression technique and index flood method) and nonlinear methods (e.g., generalized) additive model. At the beginning, basic issues on design flood estimation are discussed, which is followed by a detailed description of various RFFA methods (index flood method, quantile regression technique and generalised additive models). The model validation techniques are then presented, followed by a description of cluster analysis.

2.2. Basic issues

2.2.1 Design flood estimation methods

Design of water control structures, reservoir management, economic evaluation of flood protection projects, land use planning and management and flood insurance assessment rely on knowledge of the magnitude and frequency of floods, which is referred to as design flood (Srinivas et al., 2007). Often, estimation of design flood is not easy because of paucity of flood records at the sites of interest. The most common methods of design flood estimation include at-site flood frequency analysis (FFA) using observed peak discharge data and event based rainfall runoff modelling.

The design flood can be estimated more accurately for catchments where relatively long streamflow data is available; however, for ungauged catchments (where recorded streamflow data is unavailable or of limited length (less than 10 years) or of poor quality), accurate predictions of design floods remains a challenging task. Moreover, design flood estimates for ungauged catchments are generally associated with a large degree of uncertainty (Haddad and Rahman, 2012).

Error in design flood estimates can lead to undersized or oversized drainage systems, which are equally unacceptable for drainage design; the former results in frequent flooding which cause inconveniences to inhabitants. The latter produces an uneconomical design, which costs more money. Thus, for the design of an efficient and economic drainage system, it is important to estimate design floods accurately.

Selection of particular design flood estimation methods largely depend on the data availability and the purpose of the flood estimation. Lumb and James (1976), Feldman (1979), and James and Robinson (1986) broadly classified design flood estimation methods into two broad categories: streamflow-based methods and rainfall-based methods. These are discussed below and illustrated in Figure 2.1.

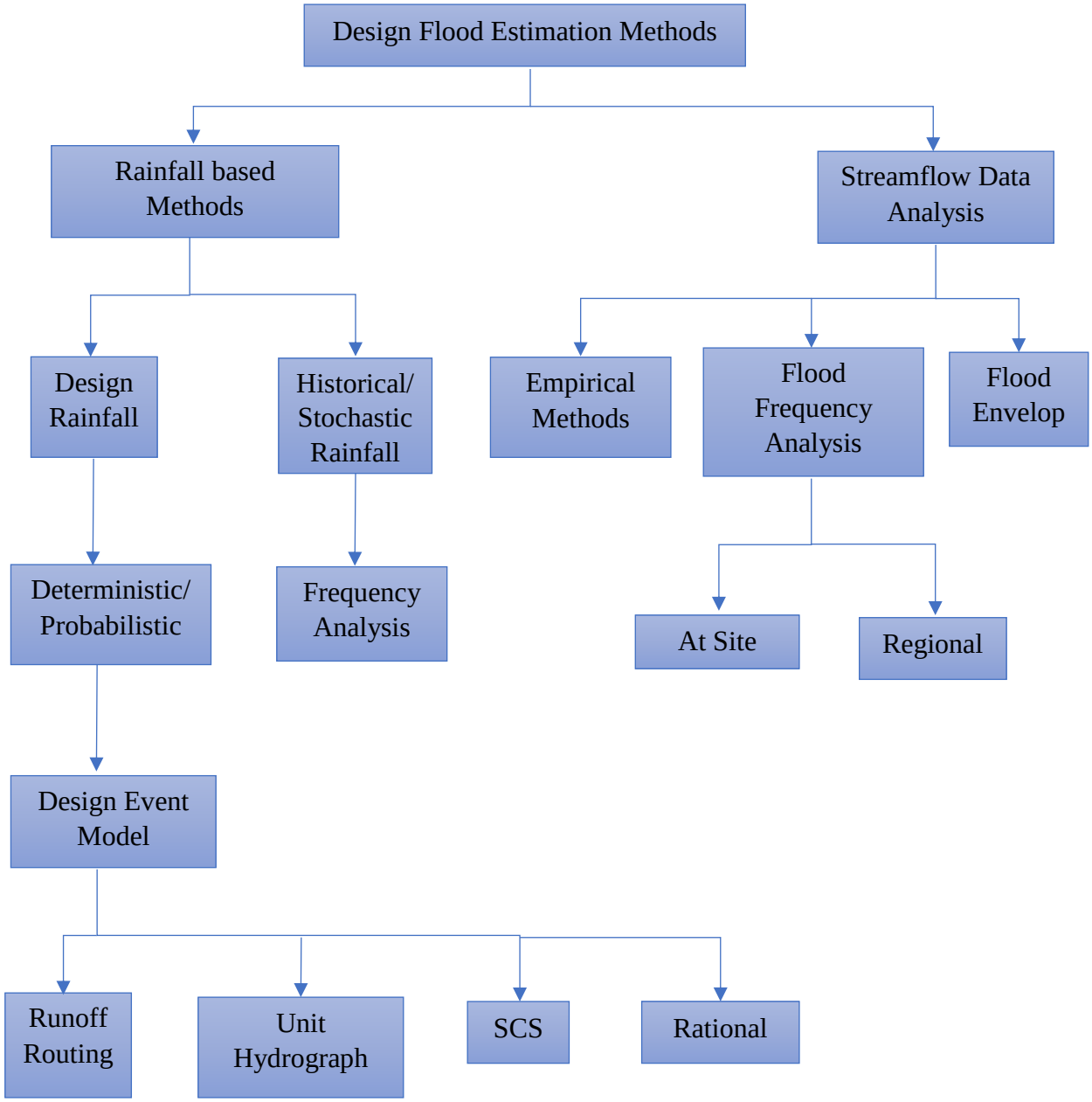

Figure 2.1 Various design flood estimation methods

2.2.2 At-site flood frequency analysis

At-site flood frequency analysis (FFA), a streamflow-based method, is the most direct method for estimating design floods utilizing the observed peak flow data. The main objective of this method is to develop a relationship between the flood magnitude and annual exceedance probability (AEP) through the use of probability distributions (Chow et al., 1988).

The prime advantage of FFA is that they provide a direct estimate of design floods based on gauged data. Peak flood records represent the integrated response of a catchment to storm events and thus are not subject to the potential for bias that can affect rainfall-based procedures. Furthermore, FFA is quick to apply compared to rainfall-based procedures and have the ability to provide estimates of uncertainty associated with the size of sample and gauging errors. These represent very considerable advantages, and thus it is not surprising that FFA is an important tool for the practicing hydrologists.

However, there are some practical disadvantages with FFA. The available peak flood records may not be representative of the conditions relevant to the problem of interest: changing land-use, urbanisation, upstream regulation, and non-stationary climate are the likely factors that may confound efforts to characterise flood risk. The length of available record may also limit the utility of the flood estimates for the rarer quantiles of interest. Peak flow records are obtained from the conversion of stage data and there may be considerable uncertainty about the reliability of the rating curve when extrapolated to the largest recorded events. In addition, gauges may be relocated, survey datum has been altered, and channel conditions may change, and hence different rating curves are applicable to different periods of historical data. There is also uncertainty associated with the choice of probability distribution which is not reflected in the width of derived confidence limits: the true probability distribution is unknown and it may be that different models may fit the observed data equally well yet diverge markedly when used to estimate quantiles beyond the period of record.

Perhaps the most obvious limitation of FFA is that it relies upon the availability of recorded flood data. This is a particular limitation in urban drainage design as there are so few gauged records of any utility in developed catchments. But the availability of representative records

is also often a limitation in rural catchments, either because of changed upstream conditions or because the site of interest may be remote from the closest gauging station.

FFA methods are most relevant to the estimation of peak flows for very frequent to rare floods. FFA methods can also be applied to other flood characteristics (e.g. flood volume over given duration), but this involves additional assumptions. Peak-over-threshold analysis is most relevant to the estimation of flood exceedances that occur several times a year, up to floods more frequent than around 10% AEP. For rarer events, the use of an annual maximum series is preferred, and with good quality information FFA methods are suited to the estimation of rare floods with AEPs of 2% to 1%. The use of regional flood data provides valuable information that can be used to help parameterise the shape of the flood distribution, and thus where feasible it is desirable to use at-site/regional flood frequency methods. The use of regional information can support the estimation of flood risks beyond 1% AEP and can greatly increase the confidence of estimates obtained using information at a single site.

2.2.3 Regional flood frequency analysis

Regional flood frequency analysis (RFFA) entails estimating design floods at an ungauged site by utilizing flood records pooled from several other catchments, which are similar to the ungauged site of interest. The process of identifying similar catchments for pooling peak flow information is known as regionalization. Research in this area is active over past four decades with new and intriguing findings constantly being reported.

RFFA method can enhance particular site estimates using regional relationships, especially for parameters like skew, which is more prone to sampling error and data extremes. Moreover, regional relationships optimize the effect of outliers which can lead to more reliable extrapolation of flood frequency curve of rarer frequencies. RFFA also enhances the design flood estimates at gauged sites where data may be limited and where direct flood frequency analysis is not feasible.

Various RFFA methods have been adopted in the past such as Rational Method, Probabilistic Rational Method (PRM), Index Flood Method, Quantile Regression Technique, Parameter Regression Technique, and artificial intelligence-based methods (Aziz et al., 2014; Aziz et al., 2015; Bates et al., 1998; Rahman. et al., 2011)

The Rational Method was first introduced by Mulvaney (1851) to estimate peak discharge, which is generally regarded as a deterministic model. However, ARR 1987 recommended a probabilistic form of the Rational Method, known as Probabilistic Rational Method (PRM) for Victoria and Eastern New South Wales (NSW). The PRM in ARR 1987 was based on the studies by Pilgrim (1982), Pilgrim and McDermott (1982) and Adams (1984). The application of the PRM in ARR 1987 requires a contour map of runoff coefficient. The runoff coefficient is assumed to vary smoothly over geographic space; however, a sharp variation in the runoff coefficients has been found even within a close proximity indicating discontinuities at catchment boundaries (Pirozzi et al., 2009; Rahman et al., 2008; Rahman and Hollerbach, 2003)

RRFA procedures generally involve the use of regression models to estimate the parameters of probability models (or the flood quantiles) using physical and meteorological characteristics, although simpler scaling functions can sometimes be used for local analyses. Rahman et al. (2015) provided details of a regional flood frequency estimation (RFFE) model for different Australian regions in which the three parameters of the log-Pearson Type 3 model are estimated from catchment characteristics using a Bayesian regression approach. This RFFE model has been incorporated in ARR 2016. The RFFE model provides a quick means to estimate design floods for AEPs ranging between 50% to 1%. The prime advantage of this technique is that it provides estimates of design floods (with uncertainty) using readily available information at ungauged sites; the estimates can also be combined with at-site analyses to help improve the accuracy of the estimated design floods. The prime disadvantage of the technique is that this is only applicable to the range of catchment characteristics used in development of the model, and this largely excludes urbanised catchments and those influenced by upstream impoundments (or other sources of major modification). For such catchments, it will be necessary to consider the use of rainfall-based methods. The RFFE model is quick to apply and provides a formal assessment of uncertainty, and thus is well suited to provide independent estimates for comparison with other design flood estimation approaches.

2.3. Different methods of RFFA

2.3.1. Index flood method

The index flood method is commonly used to develop a flood frequency curve that relates flood magnitude to flood AEP. This method involves scaling a dimensionless flood frequency curve by the index flood. The index flood is a middle-sized flood for which the mean or median of the flood data series is typically used. When the catchment of interest is ungauged, statistical models, such as multiple regressions, are often used to relate the index flood to catchment descriptors.

The index flood method was developed by the US Geological Survey (Dalrymple, 1960) and is based on the technique which relates to the hydrologically similar region. The method extracts data from gauged catchments within a defined region for calculation of parameters for a dimensionless flood frequency curve. The "index flood" of the catchment of interest then scales the curve.

If q_T is the dimensionless growth factor, μ_i is the index flood for site i, then the estimate of the T year flood event at site i, Q_T^i can be estimated by:

$$Q_T^i = \mu_i q_T \qquad \ldots(2.1)$$

The index flood, μ, is a middle-sized flood as the mean or median flood ($\bar{Q}$ and Q_{med}, respectively). The median flood, Q_{med}, is often preferred as it is a more robust measure than a mean, especially when the index flood must be estimated for a gauged catchment with a short record length. In case of ungauged catchments, the index flood is often estimated through some form of statistical modelling such as multiple regression.

Regression has long been used in hydrology to relate a desired flood quantile to catchment physiographic, geomorphologic and climate characteristics. The analysis is typically performed using the power-form equation:

$$Q_T = a x_1^{\beta 1} x_2^{\beta 2} x_3^{\beta 3} \ldots\ldots\ldots x_p^{\beta p} \qquad \ldots(2.2)$$

where Q_T is the flood quantile of interest, 'a' is constant, x_i is the i^{th} catchment characteristics, β_i is the i^{th} model parameter, and p is the number of catchment characteristics. In the present context, the quantile of interest is the median flood, which represents the index flood.

A significant amount of research has been conducted in regards to the index flood method both in the past and more recently. Dalrymple (1960) was one of the first researchers to develop an index flood technique which was used by the United States Geological Survey (USGS) prior to 1965. The method developed by Dalrymple (1960) was to relate annual maximum flood series to catchment areas for a particular region of interest. According to the assumption, the flood distribution at different sites was taken constant within a homogeneous region except for a site-specific scale or index flood factor. Homogeneity stands on the concept that the standardised peak floods from different sites in selected regions would follow the common probability distribution with identical parameter values. Relationships were then sought on geographical representation; the particular area was then divided into divisions based on similarity (Riggs, 1973).

The second part of Dalrymple's approach involved averaging the shapes of similar curves for the region to create one similar common curve; this method was relatively easy to implement as only one variable was required: which was catchment area. As this approach is an empirical one, a number of limitations have been identified:

- Arbitrary decisions are required at boundaries of regions with respect to mean annual flood and the shape of the frequency curve.

- There was no consideration of other important factors which have shown to be plausible/influential in the flood generation process(Riggs, 1973).

According to ARR 1987 (Pilgrim et al., 1987), the index flood method is not encouraged as adesign flood estimation technique for Australia. The assumption has been criticised on the grounds that it is heavily dependent on the idea of regional homogeneity which is not quite satisfactory in the case of Australian regional flood data. The coefficient of variation may vary approximately inversely in terms of catchment area, thus resulting in flatter frequency

curves for larger catchments. The scenario is particularly prominent in the case of humid catchments that differ greatly in size (Riggs, 1973; Smith, 1989).

The index flood method further developed in the late 1980s is a vast improvement to the past methodologies, which use regional average values of LCV and LSK with the at-site mean to fit a GEV or an alternative distribution (Hosking and Wallis, 1997). According to Hosking and Wallis (1997), this approach is effective for the relatively homogeneous region and where record lengths are relatively short. For a finer rating curve, a regional GEV shape parameter can be adopted based upon a regional average. The approach calls a pathway to solve the problems by increasing record lengths and regional homogeneity but at-site data was not long enough to define the shape parameter. Combination of at-site and regional estimators based on each estimator have been proposed as a solution.

Index flood method has been discouraged due to heterogeneity and complexities among Australian catchments. Results show certain discrepancies which is concerning due to concurrent errors in further applications. This provides the ground to further experimentation on other methods where assumptions of homogeneity might be relaxed by considering the spatial variability from site to site within a region.

2.3.2. Quantile regression technique

Regression technique is a simple approach that allows the use of different distributions for different sites in the region. This model develops a transfer function to define a direct relationship between at-site quantiles (outputs) and physio-meteorological variables (predictors or inputs). These techniques have been well suited to ungauged catchment simulations because of their ease of implementation, their rapidity and their good performance. In this regard, numerous models were proposed for RFFA using different transfer functions, including the linear regression model (e.g., Di Prinzio et al., 2011; Holder, 1985; Pandey and Nguyen, 1999; Phien et al., 1990), the generalized linear model (e.g.,Nelder and Baker, 1972), the generalized additive model (Chebana. et al., 2014) and artificial neural networks (Abrahart et al., 2007; Shu and Ouarda, 2007).

The major limitation of regression-based method is that they generally provide only the mean or the central part of at-site flood quantiles. As a result, most of the regression technique

gives the conditional mean of the quantile at ungauged sites considering the physiographic variables (Ouali et al., 2016; Ouarda et al., 2016; Pandey and Nguyen, 1999; Wazneh et al., 2013). Hence, estimated quantiles at gauged sites are commonly used to calibrate the transfer function of the regression model and are not the total representation of full hydrological time series observations.

The USGS adopted an empirical quantile regression method in which a large number of gauged catchments are selected from a region and flow quantiles are estimated from streamflow data, which are then regressed against a set of climatic and catchment characteristic variables that govern the flood generation process. The quantile regression method can be expressed as follows:

$$Q_T = aB^bC^cD^d \quad \text{...(2.3)}$$

where *B, C, D*, … are climatic and catchment characteristics variables (predictors) and Q_T is the flood magnitude with *T* year ARI, and *a, b, c, d*, … are regression coefficients.

This method does not require the assumption of a constant coefficient of variation (*Cv*) of annual maximum flood series in the region unlike an index flood method. It has been noted that the method can give design flood estimates that do not vary smoothly with ARI; however, hydrologic judgement can then be used to make a slight adjustment to the flood frequency curve so that flood estimates increase smoothly with ARI (Rahman, 2005).

Most regional QRTs are based on the methodology published by the USGS. Generally, this method uses a number of gauged catchments in a selected region from which the historical flood records are collected and used in a FFA to provide flood quantiles. Catchment characteristics are then collected for the same gauged catchments. The flood quantiles and catchment characteristics are then used in a regression analysis, which provides an equation that best describes the relationship between the two sets of data. Providing the gauged catchments used in the development of the equations reflect the variability in hydrological behaviour of the catchments in a given region; the equations can then be adopted as a regional flood frequency method.

QRT is particularly applicable for the small to middle-sized catchments where usually data is scarce. For example, if we consider the case of Queensland, it can be observed that there are numerous small catchments which consist of very complex nature of hydrologic and hydraulic characteristics. Therefore, this requires an approach to assess design floods in ungauged catchments using easily-measured parameters for routing drainage design projects.

In basic terms, the regression analysis attempts to allocate a proportion of the design flood peak to a particular catchment characteristic. The characteristics used in the regression are required to be hydrologically significant. That is, values must be able to be directly related to either the generation or reduction of rainfall runoff. The parameters should also be easily measured for ungauged catchments to ensure the method is able to be applied as a part of a desktop study.

Catchment characteristics that have been used in QRT studies include catchment area or shape, stream length and slope, vegetation type and quantity, soil type, rainfall depth and intensity, and in some cases, average temperature and catchment elevation. It is also important to note that there are possible inaccuracies in available data, so complex and less significant catchment characteristics may be adding to complexity without adding to the model performance for ungauged catchments. Therefore, only the most dominant characteristics should be adopted.

The USGS flood estimation methods generally use either ordinary least squares (OLS) or more recently the generalised least squares (GLS) method of regression. While the final prediction equations appear similar between the two methods, the GLS is a more complex model than OLS, which is reasonably straightforward in comparison. The GLS method as described by Stedinger (1983) is a regression technique that takes into account the correlation between, as well as differences in, the variability and reliability of the flow estimates used as dependent or response, variables. Whereas the OLS method assumes the model residual is normally distributed, each station is weighted equally, and each site is independent (uncorrelated) (Haddad and Rahman, 2012; Palmen and Weeks, 2011)

Rahman (2005) developed a QRT to test the accuracy of estimating design flood in small to medium sized ungauged catchments in south-east Australia. The study was conducted using

streamflow and catchment characteristics of data of 88 catchments of south-east Australia. The prediction equation for design floods was developed for 2, 5, 10, 20, 50 and 100 years of ARIs based on flood and catchment characteristics data of 88 small to medium sized catchments. A total of 12 explanatory (predictor) variables were adopted for the analyses: rainfall intensity of 12-hour duration and 2-year ARI (I12_2, mm/h), mean annual rainfall (rain, mm); mean annual rain days (rdays), mean annual Class A pan evaporation (evap, mm); catchment area (area, km^2); lemniscate shape, a measure of the rotundity of a catchment (shape); slope of the central 75% of the mainstream (slope, m/km); river bed elevation at the gauging station (elev, m); maximum elevation difference in the basin (relief, m); stream density (sden, km/km^2), which is the length of stream lines divided by the catchment area; fraction of basin covered by medium to dense forest (forest); and fraction quaternary sediment area (qsa). The developed prediction equations satisfied the underlying model assumptions very well and included hydrologically meaningful predictor variables that are readily obtainable. An independent test indicated that these prediction equations are able to provide reasonably accurate design flood estimates in the study area for small to medium-sized ungauged catchments.

Instead of classical quantile regression approaches, Ouali et al. (2016) proposed a quantile regression model that directly gives the conditional quantile for regional frequency analysis, avoiding using at-site estimated quantiles in the calibration process. The proposed model is able to integrate all the given hydrological information into the calibration step with very short station data record, which is an advantage in the case of poorly gauged catchments. The developed quantile regression model is applied on a dataset representing 151 hydrometric stations from the province of Quebec and compared with a classical regression model. Monte Carlo simulation method has been used to quantify the at-site estimation error and to assess the impact of record length on model accuracy. Application of this test to the annual maximum streamflow series for each gauged station indicates that three stations of 151 are found to be nonstationary at a significance level of 1%. Given the small percentage of rejected stations (2%), and to maximize sources of information; these stations have been retained in this study. In a nutshell, the model has proven to be a feasible model for regional flood estimation.

Different types of regression analysis

There are several methods to estimate regression coefficients including ordinary least squares (OLS), generalised least squares (GLS) and weighted least squares (WLS) methods.

Ordinary least Square (OLS) Method:

OLS method is widely adopted in regression analysis. This is considered as one of the simplest methods for estimation of regression coefficients. It attempts to find the best fitting regression coefficients by minimising the sum of squared residuals. The OLS model can be expressed as:

$$Y = X\beta + e \quad \ldots(2.4)$$

where **Y** is a ($n \times 1$) matrix of flow characteristics at N sites, **X** is a ($n \times k$) matrix of catchment characteristics augmented by a column of ones, **β** is a (n x 1) vector of regression parameters and **e** is an (n x 1) vector of random errors assumed to be normally distributed with zero mean and the covariance matrix assumed to be of the form $\mathbf{I}_N\sigma^2$, where $\mathbf{I}_N$ is a N–dimensional identity matrix. The OLS estimate of **β** is:

$$\beta_{ols} = (X'X)^{-1}X'\hat{Y} \quad \ldots(2.5)$$

The sampling covariance matrix based on the above assumptions can be expressed as:

$$Var(\hat{\beta}_{ols}) = \sigma^2 (X'X)^{-1} \quad \ldots(2.6)$$

The OLS estimator is generally used by hydrologists to estimate the parameters β in Equation 2.5. The accuracy of estimation by OLS in RFFA by QRT depends on several assumptions:

• The annual maximum flow at each of the sites are not correlated;

• The record lengths should be equal for all the sites; and

• The flood quantiles of gauged catchments should have equal variance.

These assumptions are very unlikely to be satisfied for hydrological regression analysis. In order to overcome the problem that has arisen from the OLS regression, Stedinger and Tasker (1985) proposed the GLS regression procedure which can result in remarkable improvements in the precision with which the parameters of regional hydrologic regression models can be estimated, in particular when the record length varies widely from site to site.

Generalised least squares (GLS) regression

Regression using hydrological data violates the assumption of OLS procedure that the residual errors associated with the individual observations are homoscedastic and independently distributed (Stedinger and Tasker, 1985). Variations in streamflow record length and cross-correlation among concurrent flows, resulting in estimation of *T* year events which is likely to vary in precision. Moreover, from the former studies it is found that, OLS estimates of the standard error of prediction and the estimated parameters are highly biased. GLS regression method is an effective way to deal with these problems.

Stedinger and Tasker (1985) used Monte Carlo simulation to show the superiority of the GLS procedure to derive empirical relationships between streamflow statistics and physiographic basin characteristics. A further extension of the GLS method was presented by Tasker and Stedinger (1989) which included the realities and complexities of regional hydrological data sets that were not addressed in the Monte Carlo simulation studies. These extensions incorporated (1) a more realistic model of the underlying model error; (2) smoothed estimates of cross correlation of flows; (3) procedures for including historical flow data; (4) diagnostic statistics describing leverage and influence for GLS regression. Therefore, it is preferable to develop GLS regression model employed by Stedinger and Tasker (1985) integrating these new extensions especially in regards to identifying the realistic model error associated with the GLS analysis. The GLS procedure as described by Stedinger and Tasker (1985) and Tasker and Stedinger (1989) require an estimate of the covariance matrix of residual errors $\hat{\Sigma}(Y)$ whose elements are organised in a matrix as follows:

$$\hat{\Sigma}(Y) = \begin{cases} \frac{\sigma_i^2}{n_i}\left[1 + K_T^2\frac{(\kappa-1)}{4}\right] \; for\; (i=j) \\ \rho_{ij}\frac{m_{ij}\hat{\sigma}_i\hat{\sigma}_j}{n_i n_j}\left[1 + \rho_{ij}K_T^2\frac{(\kappa-1)}{4}\right] \; for\; (i\neq j) \end{cases} \quad \ldots(2.7)$$

where $\hat{\sigma}_i$ is an estimate of the standard deviation of the observed flows at site *i*, K_T is the *T* year frequency factor for the flow distribution, κ is the kurtosis of the flow distribution, n_i is the record length at site *i*, m_i is the concurrent record length at sites *i* and *j*, and ρ_{ij} is an estimate of the cross correlation of concurrent flows at sites *i* and *j*.

Reis et al. (2005) upgraded the GLS regional regression model developed by Stedinger and Tasker (1985) by introducing a Bayesian approach to parameter estimation for hydrological assessments. From results in Reis et al. (2005) it is found that for cases with small model error variance comparing to sampling error of the at-site estimates, the Bayesian estimator provides a more reasonable estimate of the model error variance than the Method of Moments (MOM) and Maximum Likelihood (ML) estimators. This paper by Reis et al. (2005) also show regression statistics for WLS and GLS models including pseudo analysis of variance, a pseudo R^2, error variance ratio (EVR) and variance inflation ratio (VIR), and leverage and influence. Results obtained from OLS, WLS and GLS procedures were compared. Results from the OLS procedure provided were too scattered because it did not differ between the variance due to the model error and the variance due to the sampling error. The GLS method was found to provide the best result because the cross correlation between concurrent flows proved to be important. Both leverage and influence statistics were very useful in identifying stations that did have a significant impact on the analysis. In Australia, GLS regression has been applied in RFFA by Haddad and Rahman (2012).

Weighted least squares(WLS)

Tasker (1980) and Stedinger and Tasker (1985) developed the WLS procedure which accounts for sampling error in each Yi but not their cross correlation. The WLS β estimator is;

$$\hat{\beta}_{WLS} = (X^T \hat{W} X)^{-1} X^T \hat{W}^{-1} \hat{Y} \quad \ldots(2.8)$$

where $w_{ij} = [\wedge(\gamma^2)_{ii}]^{-1} \quad i=j$,

$w_{ij} = 0$ *otherwise*

Assuming $=[\wedge(\gamma^2)_{ii}]^{-1}$ is indeed *W* (which is the case if $\rho_{ij} = 0$ for all $i \neq j$), the covariance is

$$\hat{\beta}_{WLS} = (X^T \hat{W} X)^{-1} \quad \text{...(2.9)}$$

As with GLS, a difficulty encountered with the WLS estimation procedure is that the β estimator is defined in terms of the unknown model error variance $\gamma 2$. Two estimators of $\gamma 2$ are considered here for use with a WLS algorithm.

Tasker (1980) proposed a method of moments γ^2 estimator for use with WLS procedures. His estimator is based on a correction to the residual mean square error s_r^2.

In this instance the basic model is x_i

$$\hat{Y}_i = \beta_0 + \beta_1 ln\, A_i + \hat{\varepsilon}_i \quad \text{…(2.10)}$$

Where

$$Var\,[\hat{\varepsilon}_i] = \gamma^2 + Var[\hat{Y}_i]$$

$$Var[\hat{Y}_i] = E[(\hat{Y}_i - Y_i)^2]$$

As a result, for $\rho_{ij} = 0$ $(i \neq j)$

$$E[s_r^2] \cong \gamma^2 + \frac{1}{N} \sum_{i=0}^{N} Var[\hat{Y}_i] \quad \text{…(2.11)}$$

Thus, a method of moment's estimator of γ^2 for the WLS model when $\hat{Y}_i = \bar{x}_i + K_T s_i$ would be

$$\hat{Y}_{WLS-MM1} = s_r^2 - \frac{1}{N} \sum_{i=0}^{N} (1 + K_T^2/2)(s_i^2/n_i) \quad \text{…(2.12)}$$

The model error variance can be estimated by Tasker's (1980) method of moment's estimator $\hat{Y}_{WLS-MM1}$ in Equation 2.12, or by Stedinger and Tasker's (1985) method of moment's estimator $YWLS{-}MM1$ obtained by solving Equation 2.12.

There may be some difficulties in case of using WLS with hydrological data as it needs the estimation of the covariance matrix of residual errors. The covariance matrix is a precision function which is associated with sampling errors in the statistical estimations. The

discussion in the works by Tasker (1980) denotes difficulties associated with the estimation of this matrix.

2.3.3. Challenges regarding log transformation of regression variables

Most of the existing regression techniques are based upon the assumption that the model can be linearized by the logarithmic transformation. However, the danger with the logarithmic transformation is that unusually small observations are given greatly increased weights. This makes the estimated parameters biased in real flow domain, although they may be unbiased in log-flow domain (McCuen et al., 1990). Some previous efforts have been made to correct the transformation bias by modifying the intercept term of the model. However, as indicated by Miller (1984), correction of bias through the modification of the intercept term may eliminate only a portion of the total bias because other parameters of the model are not considered at all. On the contrary, Koch and Smillie (1986) reported high sensitivity of bias correction to the normality assumption and cautioned the use of bias correction techniques outside of the normality assumption. Cohn et al. (1989) reported that neglecting bias might produce significant under-prediction and that incorrect bias correction may lead to severe over-prediction. Alternatively, a model with an additive error could be employed, where the parameters are estimated directly using the real flows using the desired objective function. However, for additive model, there is no unanimity in the type of objective function to be used to determine the parameters.

2.3.4. GAM based method

The application of more general non-linear methods such as the generalized additive model (GAM) (Hastie and Tibshirani, 1987; Wood, 2006) has increased in recent years due to the development of new statistical tools and computer programs (e.g.,Kauermann and Opsomer, 2003; Morlini, 2006; Schindeler et al., 2009; Wood, 2003). GAMs have been applied successfully in environmental studies (e.g.,Wen et al., 2011; Wood and Augustin, 2002) in renewable energy assessment (e.g. Ouarda et al., 2016) and also in public health and epidemiological research (Bayentin et al., 2010; Clifford et al., 2011; Leitte et al., 2009; Vieira et al., 2009). There have been a number of applications of GAM in meteorology, e.g. Guan et al. (2009) applied GAM to predict temperature in mountainous regions and

Bertaccini et al. (2012) applied it to examine the impacts of traffic and meteorology on air quality.

In hydrology, there have only been limited applications of GAM. Tisseuil et al. (2010) applied generalized linear model (GLM), GAM, aggregated boosted trees (ABT) and multi-layer perceptron neural networks (ANN) for statistical downscaling of general circulation model outputs to local-scale river flows. They found that the non-linear models GAM, ABT and ANN generally outperformed the linear GLM when simulating fortnightly flow percentiles.

Morton and Henderson (2008) applied GAM to estimate nonlinear trends in water quality in the presence of serially correlated errors. They noted that GAM produced more reliable results and it could estimate the variance structure more accurately. In a recent study, Asquith et al. (2013) applied the generalized additive regression modelling approach to develop prediction equations to estimate discharge and mean velocity from predictor variables at ungauged stream locations in Texas, US. Asquith et al. (2013) noted that the incorporation of smooth functions is the strength of GAMs over simpler multilinear regression since appropriate smooth functions can accommodate otherwise difficult to linearly model components of a prediction model. In their study, the developed GAM-based non-linear models were found to provide more accurate prediction. Wang et al. (2015) modelled summer rainfall from 21 rainfall stations in the Luanhe River basin in China using non-stationary Gamma distributions by means of GAM. Galiano et al. (2015) adopted GAM to fit non-stationary frequency distributions to model droughts in south eastern Spain. Shortridge et al. (2015) adopted GAM to simulate monthly streamflow in five highly-seasonal rivers in Ethiopia.

In RFFA, the application of GAM has not been well investigated. In one study, Chebana et al. (2014) compared a number of RFFA methods (both linear and non-linear) using a dataset of 151 hydrometrical stations from Quebec, Canada. They found that RFFA models using GAM outperformed the linear models including the most widely adopted log-linear regression model. They noted that smooth curves in GAM allowed for a more realistic understanding of the physical relationship between dependent and predictor variables in RFFA. Rahman et al.

(2018) tested the applicability of GAM model in RFFA using NSW data and found promising results.

GAM allows for the inclusion and presentation of nonlinear effects of predictor variables on response variable. It is known that catchment rainfall and runoff hydrologic process is generally non-linear; for example, a larger rainfall on drier catchment produces smaller runoff compared to a wetter catchment. Hence, the application of GAM in predicting flood discharge at ungauged catchments is relevant. Moreover, GAM adopts nonparametric smooth functions to link the dependent and predictor variables, which makes GAM more flexible in capturing relationships between the dependent and predictor variables. In summary, GAM allows accounting for possible nonlinearities in regional flood models that cannot be achieved using linear models or through simple variable transformations such as log or power.

2.3.5. Formation of region by cluster analysis

Cluster analysis is the method that assists in finding patterns or groups in the data. The individual groups according to catchment characteristics are formed through cluster analysis, and thus hydrological homogeneous areas can be delineated. The regional estimation method that is often a set of regression models is developed for each cluster/group.

Clustering algorithms are generally categorised under two different categories – partitional and hierarchical. Partitional clustering algorithms divide the data set into non-overlapping groups and algorithms, k-mean, bisecting k-mean, k-modes, etc., fall under this category. Partitional clustering algorithms employ an iterative approach to group the data into a pre-determined k number of clusters by minimising a cost function. Whereas, hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom.

A number of methods of cluster analysis with different distance measures are used (e.g., Mosley, 1981; Tasker, 1982b; Acreman and Sinclair, 1986; Burn, 1989; Hughes and James, 1989; Roald, 1989; Nathan and McMahon, 1990; Burn and Boorman, 1993). One problem in cluster analysis is that it generates different groupings with different methods of cluster analysis. The question then arises which of these groupings is to be selected as the ‘acceptable grouping’. In selecting the ‘acceptable grouping’ the criterion could be that there

is no chaining effect in the final clusters and there should be well defined grouping in the final sets of clusters/groupings.

To overcome the problem arising from different dimensional units of the variables in cluster analysis, the variables are generally standardized. The variables can be transformed to z-scores (mean = 0 and standard deviation = 1).

2.3.6. The hierarchical cluster analysis

There are numerous ways in which clusters can be formed. Hierarchical clustering is one of the most straightforward methods. A key component of the analysis is repeated calculation of distance measured between objects, and between clusters once objects begin to be grouped into clusters. The outcome is represented graphically which is known as a dendrogram. The drawback of hierarchical clustering algorithms is that the resulting clusters are usually not optimal because the feature vectors committed to a cluster in the early stages cannot move to another cluster. Because the goal of the cluster analysis is to form similar groups of figure-skating judges, so to measure a similarity or distance, a criterion needs to be selected. This distance is a measure of how far apart two objects are, while similarity measures how similar two objects are. For cases that are alike, distance measures are smaller and similarity measures are larger. Some, like the Euclidean distance, are suitable for only continuous variables, while others are suitable for only categorical variables. There are also many specialized measures for binary variables. Some common distance measures are:

• Block;

• Euclid;

• Seuclid;

• Correlation;

• Cosine;

• Chebychev;

• Minkowski; and

• Power.

K-**means clustering**

K-means clustering is a partitioning method. The function k-means partitions data into k mutually exclusive clusters, and returns the index of the cluster to which it has assigned each observation. Unlike hierarchical clustering, k-means clustering operates on actual observations (rather than the larger set of dissimilarity measures), and creates a single level of clusters. The distinction mean that k-means clustering is often more suitable than hierarchical clustering for large amounts of data.

2.3.7. Model validation in regression analysis for hydrological assessments

Validation is an important tool for hydrological regression analysis considering the accuracy of the prediction model. In RFFA, multiple regression is the tool for the derivation of the best set of predictor variables, which is best suited or most optimal for inclusion in regression equation avoiding overfitting or under fitting. It is important to develop regression model as a dependable solution for the purpose of making reliable predictions for ungauged catchments.

Validation methods are often used to test the models' performance in hydrologic regression analysis. In this method, a fixed percentage of the data (e.g. 10%, 20%, 30%) is set apart during building the model, while the rest of the dataset is used as the training data for model. Then the developed model is tested on the left-out dataset which was not used for model building. This data set is termed as validation data set.

The validation procedure helps not only to find out the appropriate model according to its prediction ability but also evaluating the prediction ability of the model for ungauged catchments at the same time (Burn, 1990).

K fold cross validation

K fold cross validation is a well-known approach for hydrological assessments and validation methods. This approach randomly divides the set of observations into k groups or folds which

are of equal sizes considering first fold as the validation set and rest of the data as training set. The procedure is considered a good approach, considering it repeats the whole procedure for k times resulting better accuracy.

There have been several studies in regards to k fold cross validation in hydrological applications (Burn, 1990; De Michele and Rooso, 2002; Rao and Srinivas, 2006) .

2.4. Summary

This chapter provides a brief review of design flood estimation methods such as FFA and RFFA. This also reviews index flood method, QRT, GAM and cluster analysis for RFFA. The fundamental concepts, mathematical equations and input data requirements for each of these methods are presented in this chapter. The k fold validation technique is also described, which allows an independent testing of the developed models/prediction equations.

CHAPTER 3

SELECTION OF STUDY AREA AND DATA PREPARATION

3.1. General

This thesis focuses on design flood estimation in ungauged catchments using generalised additive models (GAM). Regional flood frequency analysis (RFFA) methods are based on the streamflow and catchment characteristics data of a set of selected gauged catchments in a region. It is important that appropriate set of catchments are selected and data is prepared following standard procedures. This chapter presents a selection of study area and catchments, collation of streamflow and catchment characteristics data used in this research.

3.2. Selection of study area

The proposed study selects the State of Victoria as the study area since it has a good number of stream gauging stations with good quality data as compared to other Australian states. The following factors were considered in order to select the study catchments. The locations of the selected study catchments are shown in Figure 3.1.

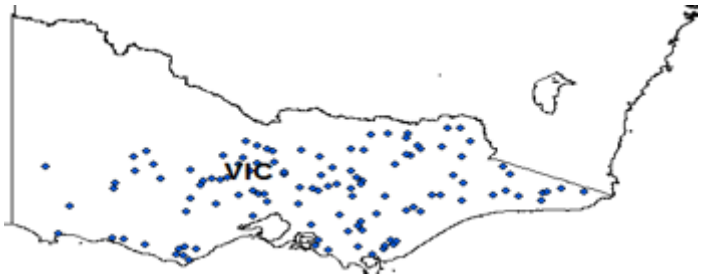

Figure 3.1 Locations of the selected study area and catchments in Victoria, Australia

3.3. Selection of study catchments

The following factors were considered in making the initial selection of study catchments.

Catchment Area: Catchment area is the most frequently adopted morphometric characteristic in RFFA, since it has a direct impact on the possible flood magnitude from a

given catchment and storm event. One of the reasons why the area variable has been so useful in statistical hydrology is its association with other significant morphometric characteristics like slope, stream length and stream order.

Record Length*:* The streamflow record at a stream gauging location should be long enough to characterise the underlying probability distribution with reasonable accuracy. In most practical situations, streamflow records at many gauging stations in a given study area are not long enough and hence a balancing act is required between obtaining a sufficient number of stations (which captures greater spatial information) and a reasonably long record length (which enhances accuracy of at-site flood frequency analysis). The selection of a cut-off record length appears to be difficult as this can affect the total number of stations available in a study area. However, for this study, the stations having a minimum of 10 years of annual instantaneous maximum flow records were selected initially as 'candidate stations'.

Regulation*:* Ideally, the selected streams should be unregulated, since major regulation affects the rainfall-runoff relationship significantly (e.g. storage effects). Streams with minor regulation, such as small farm dams and diversion weirs, may be included because this type of regulation is unlikely to have a significant effect on annual maximum floods (AMF). Gauging stations on streams subject to major upstream regulation were not included in this study.

Urbanisation*:* Urbanisation can affect flood behaviour dramatically (e.g. decreased infiltration losses and increased flow velocity). Therefore, catchments with more than 10% of the area affected by urbanisation were not included in the study.

Land-use Change*:* Major land-use changes, such as the clearing of forests and changing agricultural practices notably modify the flood generation mechanisms and make streamflow records heterogeneous over the period of record length. Catchments which have undergone major land-use changes over the period of streamflow records were not included in the data set.

Quality of Data*:* Most of the statistical analyses of flood flow data assumes that the available streamflow data is essentially error free; at some stations this assumption may be grossly violated. Stations graded as 'poor quality' or with specific comments by the gauging

authority regarding quality of the data were assessed in detail; if they were deemed 'low quality', they were excluded.

Based on the above criteria, 114 stations were selected. The geographical distribution of the candidate stations can be seen in Figure 3.2. It is interesting to note that there is a lack of stations in the Northwest of Victoria. It is not surprising, as there is usually little surface runoff during most years in this region and there is lack of a well-defined stream network in this region.

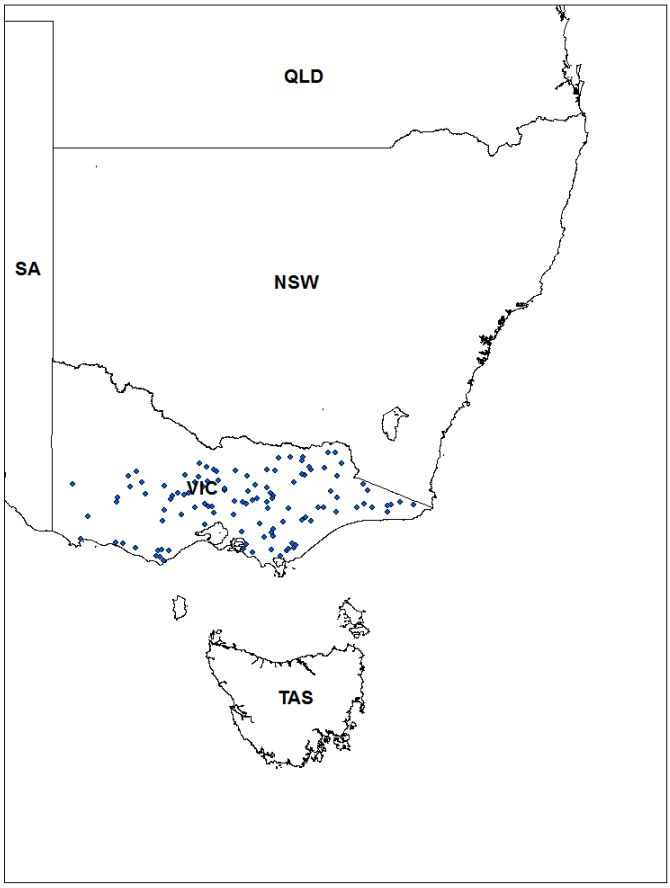

Figure 3.2 Geographical distributions of the selected study catchments

3.4. Selection of catchment characteristics

To identify the most relevant catchment characteristics in RFFA is a complex task. Moreover, most of the catchment characteristics are highly correlated, thus the presence of many of these in the prediction model might give rise to problems with the statistical analysis: such as

introducing multi-collinearity and not being able to provide much other extra useful information.

According to Rahman (1997), an initial selection of candidate characteristics should be based on an evaluation and success of catchment characteristics used in past RFFA studies, as there is no objective method for selecting catchment characteristics. Therefore, common catchment/climatic characteristics from the past studies are used as the reference and selection for a given study to increase the overall validity of the present study. In Rahman (1997) this aspect was considered in detail from over 20 previous studies to develop a reasonable starting point. But, in RFFA, the significance of characteristics may differ from region to region, and hence, no general inference about the significance of a particular catchment characteristic can be made for a given region based on the findings of other studies.

In this research, the following considerations were adopted in selecting the catchment characteristics:

- The characteristics play a significant role in flood generation.

- These are well defined and easily derived from simple physical interpretation

- These are not highly correlated.

On the basis of the above considerations, the following 8 catchment characteristics are selected for this study.

Rainfall Intensity: Rainfall intensity is one of the most significant climate characteristics in RFFA analysis. There is no doubt that it is significant in the flood generation process. It is also quite easy to obtain.

The use of rainfall intensity requires the selection of an appropriate duration and average recurrence interval (ARI). It seems to be logical to use rainfall intensity with duration equal to the time of concentration (t_c), as applied in the rational method. However, the time of concentration (t_c) differs for the selected catchments in a study area due to variability in size and shape; i.e. it is virtually impossible to select a storm having equal time of concentration, which is representative of every catchment in this study. Therefore, it was decided to include

design rainfall intensities with a 6-hour duration and 2-year return period in this study ($I_{6,2}$, mm/h). The basic design rainfall intensities data for the selected catchments were obtained from ARR Project 5 (Rahman et al., 2015).

Mean Annual rainfall: Mean annual rainfall has been adopted in many previous studies. Mean annual rainfall has been considered in this research due to its impacts on some catchment properties (e.g. vegetation cover and wetness index), although it may not have a direct influence or a link with flood peaks. Additionally, it is simple and readily available, therefore it is used as a predictor variable in this study. The mean annual rainfall data was obtained from the Australian Bureau of Meteorology CD. For all the catchments, the mean annual rainfall value for the rainfall station closest to the centroid of each catchment was extracted.

Mean Annual Potential Evapotranspiration: Mean annual evapotranspiration is the third influential climatic characteristic considered in the flood generation process. Evapotranspiration does not affect the flood peak directly but can have a secondary effect by being a surrogate for other catchment characteristics. Evapotranspiration can be defined as the water lost from a water body through the combined effects of evaporation and transpiration from catchment vegetation. In this study, mean annual areal potential evapotranspiration data was used as it is a loss component in rainfall runoff modelling. The data used was obtained from the Australian Bureau of Meteorology and previously used in ARR Project 5.

Catchment Area: Catchment area is the most frequently adopted morphometric characteristic in RFFA as mentioned earlier and hence it has been adopted in this study.

Catchment Shape: Catchment shape has also a direct influence on flood peak generation. Large narrow basins tend to have a slower response than round basins with a shorter distance. Moreover, the spatial and temporal uniformity of rainfall also depends on catchment shape. This has been used in this study, and is defined as the ratio of the shortest distance between the catchment outlet and centroid and square root of catchment area.

Slope: Slope is of vital importance in case of any gravitational flow. The steeper slope generates greater velocity of flow when other catchment characteristics are constant.

Overland slope influences the velocity of shallow surface flow; therefore, it is considered a more important factor for generation of streamflow in smaller catchments. For larger catchments, channel slope is relatively more important than overland slope. Slope has been found to be highly correlated with area and rainfall intensity in many instances. In the upper reach of a river, commonly located in mountainous zones, catchment areas are smaller, slopes steeper and rainfall heavier.

In this study, a slope measure called S1085 has been adopted. This excludes the extremes of a slope that can be found at either end of the mainstream. S1085 is defined as the ratio of the difference in elevation of the stream bed at 85% and 10% of its length from the catchment outlet, and 75% of the main stream length.

Stream Density: Stream density is defined as the total stream length divided by catchment area. The higher stream density denotes greater stream length with smaller area; hence, it is a measure of the closeness of the spacing of channels. High stream density results into a quicker response, and is more likely to occur in regions of higher impermeable sub surface material. On the other hand, low stream density tends to occur in highly permeable subsoil regions. Stream density has been adopted in this study.

Forest area: Vegetation reduces runoff by precipitation interception and transpiration. For a surface without a canopy or leaf litter layer, the interception loss is lower and overland flow travels more rapidly with less opportunity time for infiltration. Hence, Flavell (1983) found that losses from rainfall decrease with increased clearing and that the runoff coefficient of the rational method increases with increased clearing. Fraction forest cover (i.e. forested area divided by catchment area) has been included in this study.

3.5. Summary of catchment characteristics data

Data of the selected eight predictor variables are obtained from ARR Project 5 (Rahman et al., 2015). Descriptive statistics of these data are summarised in Table 3.1.

Table 3.1 Descriptive statistics of predictor variables of the selected 114 catchments from 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	S10,85	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

3.6. Streamflow data attributes

Catchment Area

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 distribution of catchment areas of the selected catchments is shown in Figure 3.3. The statistics of catchment areas of selected 114 catchments are summarised below:

- ✓ Majority of the catchments (81 catchments) fall into the category of 3 to 400 km^2.
- ✓ 23 catchments (20%) are in the range of 500 to 700 km^2; and
- ✓ 10 catchments (9%) are in the range of 700 to 1000 km^2.

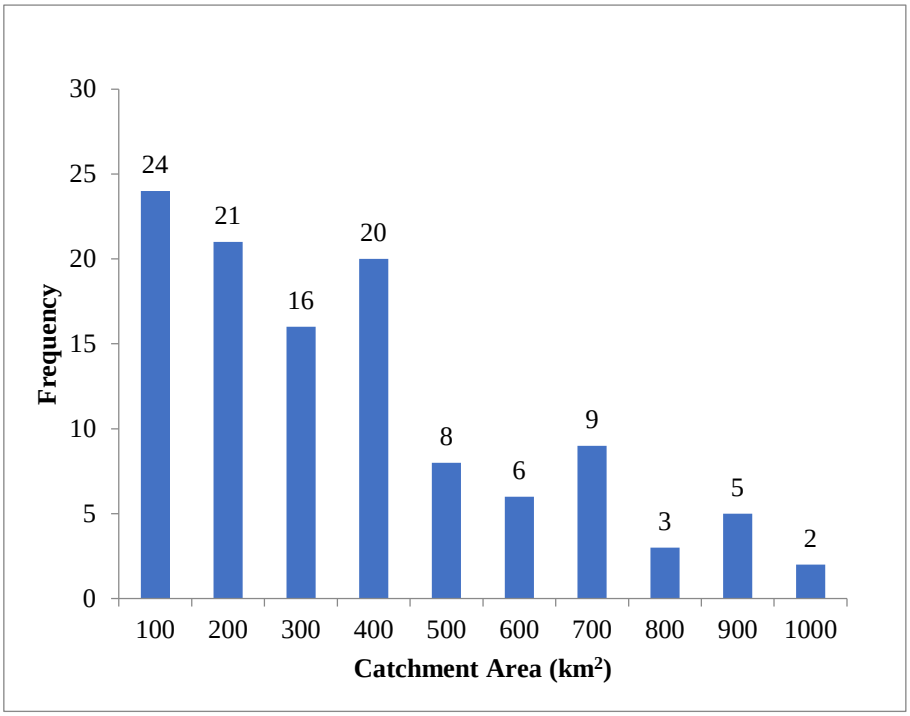

Figure 3.3 Histogram of catchment area of the selected 114 catchments

Record Length

The statistics of annual maximum flood record length is summarised below:

- ✓ Record lengths range from 26 years to 62 years, mean 38 years, median 39 years and standard deviation 5 years;
- ✓ 77 % of the stations have the record length of 34 to 42 years;
- ✓ 11% have the record length of 26 to 34 years; and
- ✓ 7% have the record length of 42 to 50 years.

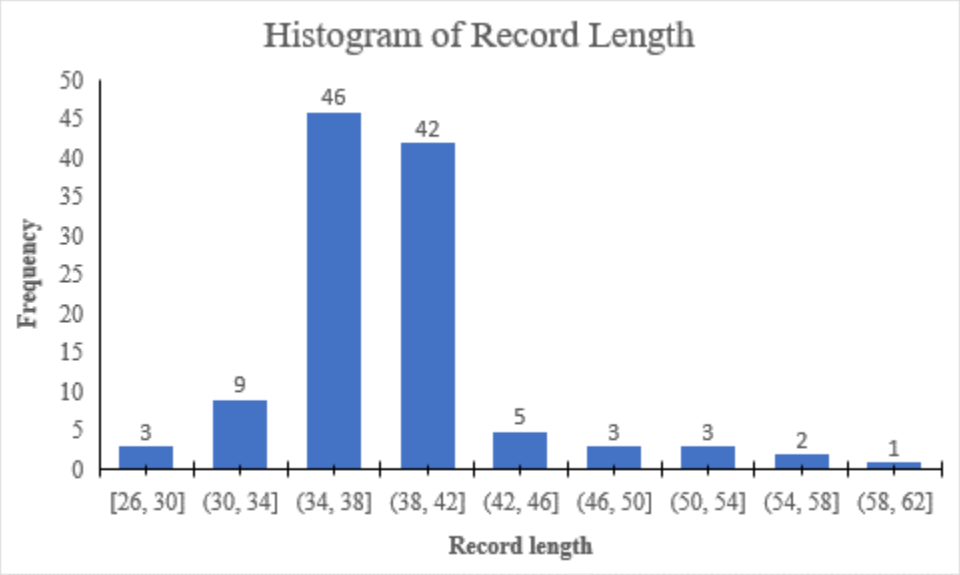

Figure 3.4 Histogram of Streamflow Record Length

3.7. Summary

A total of 114 catchments have been selected from Victoria, Australia for this study. The locations of these catchments are shown in Figure 3.2. The statistical check for streamflow data was made as described in Rahman et al. (2015). For each of the selected catchments, five catchment characteristics data have been extracted. This collection of data will now be applied in the following chapters to develop and test GAM based RFFA techniques.

CHAPTER 4

METHODOLOGY

4.1. General

This chapter describes the statistical techniques adopted in this study to develop regional flood frequency analysis (RFFA) models by using log-log linear models based on quantile regression technique (QRT) and generalised additive models (GAM). In RFFA, cluster analysis has been observed to be one of the most efficient methods to group the selected gauged stations into homogeneous groups based on catchment characteristics data; hence this has been adopted in this study. At the outset, a flow chart (Figure 4.2) is provided which summaries the statistical procedures and methodologies adopted in this thesis. At the beginning, log-log linear model is described, which is implemented by a backward stepwise regression procedure. A discussion is then presented on the QRT (the basic theory of this has been introduced in Chapter 2); further emphasis is given here on the model fitting and estimation. Thereafter, a brief discussion on GAM is provided, followed by the description of clustering algorithm and methods. Finally, this chapter discusses the model validation procedure.

4.2. Methods adopted in this study

The overall methodologies adopted in this study are illustrated in Figure 4.1.

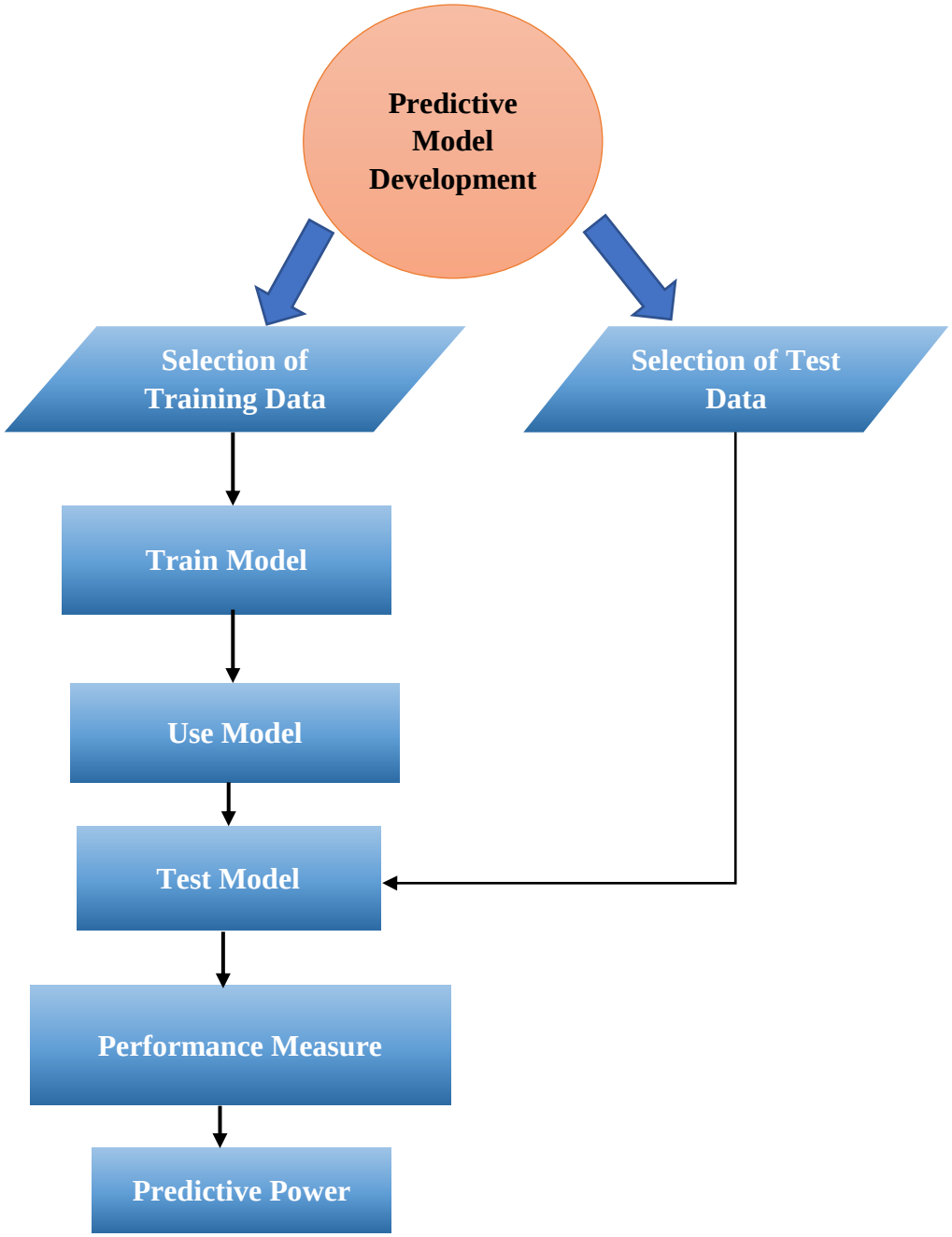

Figure 4.1 Predictive Techniques Explained

The RFFA techniques developed in this thesis are based on log-log linear regression and Generalized Additive Model. The features, fundamental concepts, mathematical equations and input data requirements for each of these methods are discussed below (Figure 4.2 provides a summary of RFFA methods).

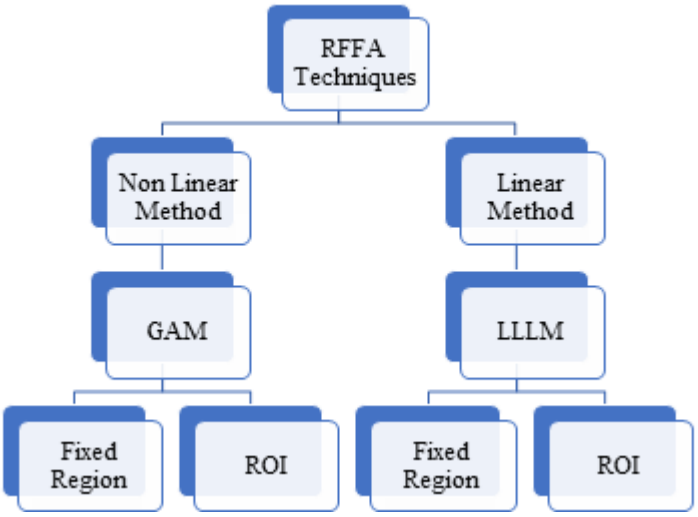

Figure 4.2 RFFA methods (LLLM stands for Log-log linear model, ROI stands for Region of influence and GAM stands for Generalised Additive Model)

4.2.1. Log-log linear model development

The statistics of flood flow largely depend on the interrelationship between flood statistics and climatic and physiographic factors. In this regard, regression analysis is widely used to develop prediction equations for flow statistics based on the data from a group of gauged catchments. These prediction equations are then used to predict flow statistics from the ungauged catchment in the study region. In a comprehensive study by the US Interagency Work Group on Flood Frequency Estimation at ungauged Sites, regression based methods of flood regionalization were found to be the most consistent and reproducible procedures for estimating flood quantiles for ungauged sites in the USA (Newton and Herrin, 1982)

The most commonly used relation between the flow statistics (e.g. flood quantile Q_T of return period *T* years) and the catchment characteristics (A_1, A_2,…,A_n) is the power-form function (Thomas and Benson, 1970) in the form:

$$Q_T = \alpha_0 A_1^{\alpha_1} A_2^{\alpha_2} \ldots . A_n^{\alpha_n} \varepsilon_0 \quad \ldots(4.1)$$

in which α_0, α_1,…, α_n are the coefficients of prediction equation, ε_0 is the multiplicative error term and *n* is the number of catchment characteristics. Alternatively, if the error term (ε_0) is assumed to be additive then the power-form function becomes (McCuen et al., 1990):

$$Q_T = \alpha_0 A_1^{\alpha_1} A_2^{\alpha_2} \ldots . A_n^{\alpha_n} + \varepsilon_0 \quad \ldots(4.2)$$

For both cases, the regression coefficients/model parameters are not known and have to be estimated using observed flow statistics data and regional catchment characteristics. If the error term is multiplicative (Eq. 4.1), then the power-form model can be linearised by a logarithmic transformation and the parameters of the linearised model can be estimated by a linear regression technique. Taking log on both sides, Eq. (4.1) can be expressed as:

$$\log(Q_T) = \log(\alpha_0) + \alpha_1 \log(A_1) + \cdots \alpha_n \log(A_{n1}) + \log(\varepsilon_{0)} \quad \ldots(4.3)$$

or in matrix form:

$$Y = X\beta + e \quad \ldots(4.4)$$

in which Y is the vector of flood statistics (quantile) from m sites ($Y = \log(Q_T)$), β is the vector of regression coefficients ($\beta = \alpha_0, \alpha_1, \ldots, \alpha_n$), X is the matrix of the physiographic characteristics or the explanatory variables ($X = \log(A_1)$) and e is the matrix of the error ($e = \log(\varepsilon_0)$). However, if the model error is additive (i.e. Eq. 4.2), it is not possible to linearise the power-form model by a logarithmic transformation and the model coefficients need to be estimated by some nonlinear optimisation method.

Log-log linear model is one of the most popular forms of linear regression analysis adopted in RFFA. In QRT, prediction equations for flood quantiles Q_2, Q_5, Q_{10}, Q_{20}, Q_{50}, Q_{100} and Q_{200} are to be developed using the mathematical assumption based on multiple linear regression analysis. For this analysis, a program was written in the statistical programming language *R*. This program produced prediction equations based on the interrelations and correlations between the dependent and predictor variables. However, both user intervention and mathematical and hydrological judgements are required to select the best form of prediction equations from the regression analyses.

The log linear function in R software is based on a backward variable selection procedure. The significance of a predictor variable is tested by checking the significance level, which must be smaller than or equal to 0.10. The goodness-of-fit of the model is assessed by coefficient of determination (R^2). Once the initial prediction equations are produced, they are then investigated for model assumptions such as outliers, normality of residuals, goodness-of-fit and influential data points. The residuals must be normally distributed and uncorrelated as per ordinary least squares (OLS) method, which is widely used in RFFA.

4.2.2. Generalized additive models

Generalized additive models (GAM) were first proposed by Hastie and Tibshirani (1987). These models assume that the mean of the response (dependent) variable depends on an additive predictor through a link function. GAM uses non-linear functions of each of the predictor variables, while maintaining additivity. Like generalized linear models (GLMs), GAM permits the response probability distribution to be from any member of the exponential family of distributions. The only difference between GAMs and GLMs is that the GAMs allow for unknown smooth functions in the linear predictor.

Mathematically speaking, GAM is an additive modelling technique where the impact of predictive variables is captured through smooth functions, which depends on the underlying patterns in the data, which could be nonlinear.

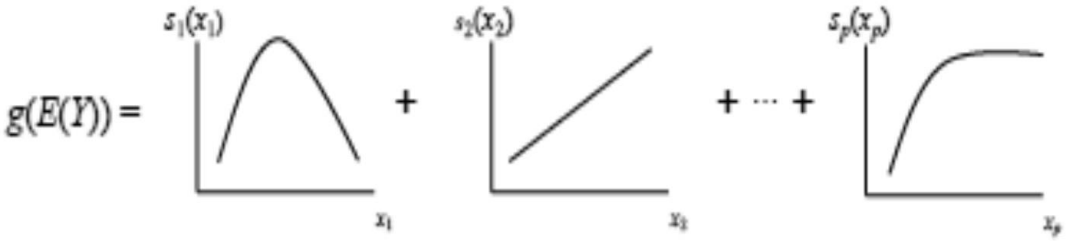

Figure 4.3 Visual Interpretation of GAM

We can write the GAM as:

$$g\big(E(Y)\big) = \alpha + s_1(x_1) + \cdots + s_P(x_P) \quad \ldots(4.5)$$

where Y is the dependent variable (i.e. what we are trying to predict, here Q_T), $E(Y)$ denotes the expected value, and $g(E(Y))$ denotes the link function that links the expected value to the predictor variables $x_1, \ldots, x_p$. The terms $s_1(x_1), \ldots, s_p(x_p)$ denote smooth, nonparametric functions.

In general, a GAM has the below form:

$$g(\mu_i) = \boldsymbol{X}_i^* \beta + \sum_{j=1}^{m} f_j\,(x_{ij}) \quad \ldots(4.6)$$

where

$\mu_i \equiv \mathrm{E}(Y_i)$ and $Y_i \sim$ an exponential family distribution;

Y_i is a response variable, $\mathbf{X}_i^*$ is the i^{th} row of the model matrix for the strictly parametric model components; and f_j are smooth functions of the covariates x_j.

In the context of regression models, the terminology 'nonparametric' means that the shape of predictor functions can be fully determined by the data as opposed to parametric functions that are defined by a typically small set of parameters. This allows for more flexible estimation of the underlying predictive patterns without knowing upfront what these patterns look like.

GAMs can also contain parametric terms as well as two-dimensional smoothers. Moreover, like GLM, GAM supports multiple link functions. For example, when *Y* is binary, we would use the logit link given by

$$g\,(E(Y\,)) = \log\frac{P(Y=1)}{P(Y=0)} \qquad \ldots(4.7)$$

GAM allows for rather flexible specification of the dependence of the response variables on the covariates, but by specifying the model only in terms of 'smooth functions', rather than detailed parametric relationships, it generally performs better than the conventional linear regression methods.

4.2.2.1. Interpretation of the model

GAM deals with highly on-linear and non-monotonic relationships between the response and the set of explanatory variables whereas linear predictor variables are interpreted in terms of a sum of smooth functions of predictor variables. To control the predictability with GAM models, it is important to define the smooth functions with varying degrees of smoothness.

Different types of smooth functions

A smoother is a tool for summarising the trend of a dependent variable *Y* as a function of one or more independent variables $X_1,\ldots, X_{p.}$ It is termed as smoother because it produces an estimate of the trend that is less variable than *Y* itself. The estimation product from smoother is termed as smooth function.

Smoother is very useful in statistical analysis. Firstly, it helps to pick up the trend from the plot easily. Secondly, it estimates the dependence of the mean of *Y* on the predictor. The most important property of smoother is its non-parametric nature; hence, the smooth function is

also known as non-parametric function. It does not assume a rigid form for the dependence of Y on $X_1,\ldots, X_p$. This is the biggest difference between GAM and GLM. It allows an 'approximation' with sum of functions (these functions have separate input variables), not just with one unknown function only. That is why it is the building block of the GAM algorithm.

Univariate smooth functions

The representation of smooth functions can be introduced by considering a model containing one smooth function of one covariate:

$$y_i = f(x_i) + \varepsilon_i \qquad \ldots(4.8)$$

where y_i is a response variable, x_i is a covariate, f is a smooth function and ε_i is the error term.

Mostly, there are three classes of smoothers used in GAM:

- Local regression
- Smoothing splines
- Regression splines

Among the smoothers, regression splines are the most practical one and frequently used due to computational ease and quick simulation. Additionally, regression splines can be written as a linear combination of basic functions that do not depend on the dependent variable *Y*, which is convenient for prediction and estimation.

Regression splines

Regression splines are more flexible than polynomials and step functions, and in fact, are an extension of the two. The main advantage of regression splines is that they can be expressed

as a linear combination of a finite set of "basis" functions that do not depend on the dependent variable *Y*, which is practical for prediction and estimation.

They involve dividing the range of variable *Y* into *K* distinct regions. Within each region, a polynomial function is fitted to the data. However, these polynomials are constrained so that they join smoothly at the regional boundaries or "knots". If the interval is divided into enough regions, an extremely flexible fit can be achieved.

To estimate *f*, using linear and logistic regressions, *f* should be represented in such a way that Eq 4.8 becomes a linear model. This can be done by choosing a "basis", defining the space of functions of which *f* (or a close approximation to it) is an element. Choosing a "basis", refers to choosing some "basis" functions, which will be treated as completely known: if $b_i\,(x)$ is the i^{th} such "basis" function, then *f* is assumed to have a representation:

$$f(x) = \sum_{i=}^{q} b_i\,(x)\beta_i \qquad \ldots(4.9)$$

where $b_i\,(x)$ are "basis" functions, *b* is the model matrix of "basis" functions and β = [β_1 :β_2 :….: β_p] are the coefficients. The number of "basis" functions depends on the number of *inner knots* – a set of ordered, distinct values of x_j – as well as the order of the spline. Specifically, if we let *m* denoting the number of inner knots, the number of basis functions is given by $K = p + 1 + m$.

Polynomial regression

Instead of fitting a high-degree polynomial over the entire range of *X*, *piecewise polynomial regression* involves fitting separate low-degree polynomials over different regions of *X*. For example, a piecewise cubic polynomial works by fitting a cubic regression model of the form

$$y_1 = \beta_o + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \epsilon_1 \qquad \ldots(4.10)$$

Here the coefficients β_o, β_1, β_2, and β_3 differ in different parts of the range of *X*. The points where the coefficients change is called knots.

For example, a piecewise cubic with no knots is just a standard cubic polynomial, as in Eq 4.10 with $d = 3$. A piecewise cubic polynomial with a single knot at a point c takes the form:

$$y_1 = \begin{cases} \beta_o + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \epsilon_1 & if\ x_1 < c \\ \beta_o + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \epsilon_1 & if\ x_1 \geq c \end{cases} \quad \ldots(4.11)$$

In other words, we fit two different polynomial functions to the data, one on the subset of the observations with $x_i < c$, and one on the subset of the observations with $x_i \geq c$. The first polynomial function has coefficients β_{01}, β_{11}, β_{21}, β_{31}, and the second has coefficients β_{02}, β_{12}, β_{22}, β_{32}. Each of these polynomial functions can be fitted using least squares applied to simple functions of the original predictor. Using more knots leads to more flexible piecewise polynomial. Generally, one does not need to worry too much about knot placement. Quantiles seem to work well in most cases (although more than three knots are usually required).

Smoothing splines

A smoothing spline is simply a natural cubic spline with knots at every unique value of x_i. Rather than using a nearest-neighbour moving window, it aims to estimate smooth functions by minimising the penalized sum of squares by fixing knots at each of the data points. The general algorithm of fitting a smooth curve uses a set of data and it aims to have RSS = $\sum_{i=1}^{n}(y_i - g(x_i))^2$ to be small.

Smoothing splines have a major drawback, it is not practical to have knots at every data point when dealing with large models. Moreover, having knots at every data point is only justified in the calculations where wiggly functions are measured with small values of λ.

However, it is important to put constraints on function $g(x_i)$ to avoid overfitting of data. The trade-off between model fit predictive modelling and smoothness is controlled by the non-negative smoothing parameter, λ, which is called the tuning parameter.

In fitting a smoothing spline, it is required to select the number or location of the knots—there will be a knot at each training observation, $x_1, \ldots, x_n$. Additionally, it is a prerequisite to choosing the value of λ. It should come as no surprise that one possible solution to this problem is cross-validation. A natural approach is to find the function g that minimises:

$$\sum_{i=1}^{n}\left(y_i - g(x_i)\right)^2 + \lambda \int g''(t)^2 dt \quad \ldots(4.12)$$

The function g that does minimisation is known as *smoothing spline.* The term $\sum_{i=1}^{n}(y_i - g(x_i))^2$is a *loss function* that encourages loss *g to fit the data well,* and the term $\lambda \int g''(t)^2 dt$ is a penalty term that penalizes variability in g. The tuning parameter λ controls the roughness of the smoothing spline, and hence the *effective degrees of freedom*. It is possible to show that as λ increases from 0 to ∞, the *effective degrees of freedom*, which we write df_λ, decrease from *n* to 2. The larger the value of λ, the smoother g will be. When λ = 0, then the penalty term in Eq 4.12 has no effect, and so the function g will be very jumpy and will exactly interpolate the training observations. When $\lambda \rightarrow \infty$, g will be perfectly smooth—it will just be a straight line that passes as closely as possible to the training points. In fact, in this case, g will be the linear least squares line, since the loss function in Eq 4.12 amounts to minimizing the residual sum of squares. For an intermediate value of λ, g will approximate the training observations but will be somewhat smoother. Thus λ controls the bias-variance trade-off of the smoothing spline.

Local regression

Local regression (loess) is an approach for fitting flexible non-linear functions, which involves computing the fit at a target point x_0 using only the regression on the nearby training observations. This belongs to the class of nearest neighbourhood-based smoothers. In order to appreciate loess, it is important to understand the most simplistic member of this family: the running mean smoother.

Running mean smoothers are symmetric, moving averages. Smoothing is achieved by sliding a window based on the nearest neighbours across the data and computing the average of Y at each step. The level of smoothness is determined by the width of the window. While appealing due to their simplicity, running mean smoothers have two major issues: they are not very smooth and they perform poorly at the boundaries of the data. This is a problem, which is dealt with more sophisticated choices, such as loess.

For example, to produce a loess-smoothed value for target data point *x*, loess involves the following steps:

1. Determine smoothness using the span parameter. For example, if span = 0.6, each symmetric sliding neighbourhood will contain 60% of the data (30% to the left and 30% to the right).

2. Calculate $d_i = (x_i - x)/h$ where h is the width of the neighbourhood. Create weights using the tri-cube function $w_i = (1 - d_{3i})^3$, if x_i is inside the neighbourhood, and 0 elsewhere.

3. Fit a weighted regression with *Y* as the dependent variable using the weights from step 3. The fitted value at target data point *x* is the smoothed value.

Below is a loess smoother applied to the simulated data, loess function in R with a span of 0.6. As we can see, loess overcomes the issues with the running mean smoother. The idea of local regression can be generalised in many different ways. In a setting with multiple features $X_1, X_2, \ldots X_p$ one very useful generalisation involves fitting a multiple linear regression model that is global.

Local regression attempts to fit models that are local in a pair of variables X_1 and X_2, rather than one. We can simply use two-dimensional neighbourhood, and fit bivariate linear regression models using the observations that are near each target point in two-dimensional space. Theoretically, the same approach can be implemented in higher dimensions using linear regressions fit to p-dimensional neighbourhoods. However, local regression can perform poorly if p is much larger than about 3 or 4 because there will generally be very few training observations close to x_0.

Estimation of GAM model parameters

GAMs consist of *multiple* smoothing functions. Thus, when estimating GAMs, the goal is to *simultaneously* estimate all smoothers along with the parametric terms (if any) in the model, while factoring in the covariance between the smoothers. There are two ways of doing this:

- Local scoring algorithm.

- Solving GAM as a large GLM with penalised iterative reweighted least squares (PIRLS).

In general, the local scoring algorithm is more flexible considering the flexibility to use any type of smoother in the model whereas the GLM approach only works for regression splines. However, the local scoring algorithm is computationally more expensive and it does not lend itself as nicely to automated selection of smoothing parameters as the GLM approach.

When fitting a GAM, the choice of smoothing parameters i.e., the parameters that control the smoothness of the predictive functions is key for the aesthetics and fit of the model. We can choose to pre-select the smoothing parameters or we may choose to estimate the smoothing parameters from the data. There are two ways of estimating the smoothing parameter for a logistic GAM:

• Generalized cross validation criteria (GCV); and

• Mixed model approach via restricted maximum likelihood (REML).

Generalized cross validation criteria

The generalized cross-validation (GCV) statistic (Golub et al., 1979) does not require iterative refitting of the model to different data subsets. The formula for this statistic is the *ith* training set outcome:

$$GCV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_1-\hat{y}_i}{1-\frac{df}{n}}\right)^2 \quad \ldots(4.13)$$

where y_i is the *ith* item in the training set outcome, $\hat{y}_i$ is the model prediction of that outcome, and *df* is the degrees of freedom of the model.

The strategy is to remove one data point at a time, fit a smoother to the remaining data, and then fit off the smoother against the entire dataset. The goal is to pick the *j* term that minimises the average error across all the *n* validations.

In fact, for a logistic GAM, we can use the GCV statistic:

$$\mathrm{GCV} = \frac{n\left\|\sqrt{W}(z - B'\beta)\right\|^2}{(n - \mathrm{tr}(H))^2} \quad \ldots(4.14)$$

where **H** is the hat matrix and **B** is the model matrix consisting of "basis" functions. This statistic essentially calculates the error of the model and adjusts for the degrees of freedom and is a linear transformation of the AIC statistic. Hence, we can use this statistic for model comparison in general, not just selection of smoothing parameters.

REML is only applicable if GAM is treated as a large GLM. Generally, the REML approach converges faster than GCV, and GCV tends to under-smooth.

4.2.3. Formation of regions in RFFA

Identification of homogeneous regions is a difficult task in RFFA, particularly in Australia which has a highly variable hydrology. The aim is to form groups of streamflow gauging sites that approximately satisfy the homogeneity criteria. In order to identify groups of catchments of similar hydrologic characteristics, cluster analysis is a widely adopted method, which is also used in this research.

Cluster analysis

Clustering refers to a very broad set of techniques for finding *subgroups*, or *clusters*, in a data set. The objective of clustering the observations of a data set is to seek partitioning of observations into distinct groups so that the observations within each group are quite similar to each other (in relation to some attributes of the data), while observations in different groups are quite different from each other.

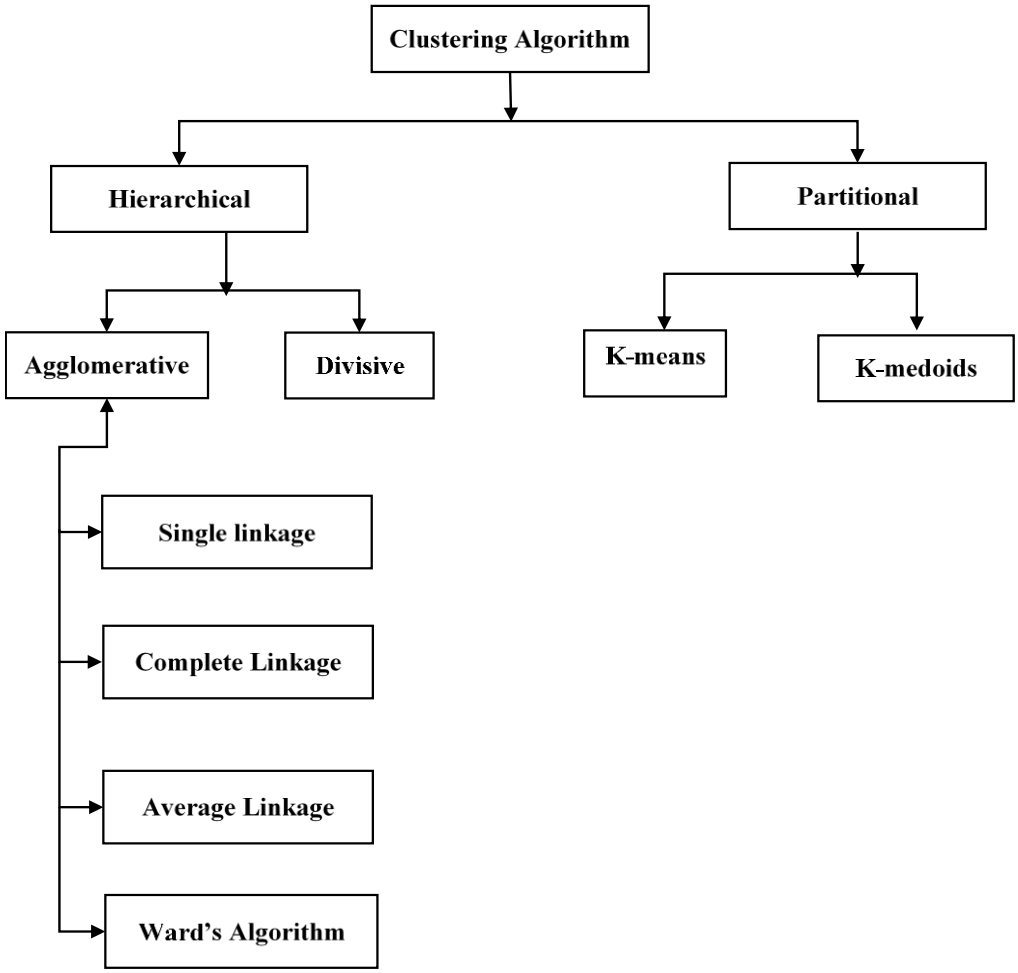

Figure 4.4 Different Clustering Techniques

Clusters are formed with sites having similar site characteristics. When the regions are intended for use in RFFA, some special considerations apply to cluster analysis. Most clustering algorithms can be classified into two categories (Jain and Dubes, 1988): hierarchical clustering and partitional clustering.

Hierarchical clustering procedures provide a nested sequence of partitions, whereas partitional clustering procedures generate a single partition of the data in an attempt to recover the natural grouping present in the data. In this subsection, a brief description of these clustering procedures is presented.

Hierarchical clustering

The hierarchical clustering process (both agglomerative and divisive) can be represented as a nested sequence or tree, called *dendrogram*, which shows how the clusters that are formed at the various steps of the process are related. Hierarchical clustering algorithms can be subdivided into two categories: Agglomerative and Divisive.

The agglomerative hierarchical clustering begins with singleton clusters and proceeds successively by merging smaller clusters into larger ones. For a given set of *N* feature vectors, the agglomerative hierarchical clustering procedures begin with *N singleton clusters*. The singleton clusters are those that consist of only one feature vector. A distance measure such as the Euclidean is chosen to evaluate the dissimilarity between any two clusters. The clusters that are least dissimilar are found and merged. This provides *N*-2 singleton clusters and a cluster with two feature vectors. The process of identifying and merging two closest clusters is repeated till the desired number of clusters is obtained.

Algorithms that are representative of the agglomerative hierarchical method of clustering include: (i) single linkage or nearest neighbour; (ii) complete linkage or furthest neighbour; (iii) average linkage; and (iv)Ward's algorithm. These algorithms differ from each other by the strategy used for defining nearest neighbour to a chosen cluster. Clusters with the smallest distance between them are merged.

Different linkage algorithms for agglomerative hierarchical clustering

The algorithms begin with *N* singleton clusters each comprising a rescaled feature vector. Among the *N* singleton clusters, two closest clusters $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are identified and merged to form a new cluster [$\boldsymbol{x}_i$, $\boldsymbol{x}_j$].

In the *single linkage* algorithm, distance between two non-singleton clusters [$\boldsymbol{x}_i$, $\boldsymbol{x}_j$] and any other singleton cluster $\boldsymbol{x}_k$ is the smaller of the distances between $\boldsymbol{x}_i$ and $\boldsymbol{x}_k$,or $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$. In general, the distance between two non-singleton clusters is the smallest of the distances between all possible pairs of feature vectors in the two clusters. This algorithm tends to form a small number of large clusters, with remaining small outlying clusters on the fringes of the

space of site characteristics and is not likely to yield good regions for regional flood frequency analysis (Hosking and Wallis, 1997; Rao and Srinivas, 2006).

In the *Complete linkage* algorithm, between the new cluster [$\boldsymbol{x}_i$, $\boldsymbol{x}_j$] and any other singleton cluster $\boldsymbol{x}_k$ is the greater of the distances between $\boldsymbol{x}_i$ and $\boldsymbol{x}_k$, or $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$. In general, the distance between two non-singleton clusters is the largest of the distances between all possible pairs of feature vectors in the two clusters. This algorithm tends to form small, tightly bound clusters. It is usually not suitable for the application to large data sets.

In the a*verage linkage* algorithm, the distance between two clusters is defined as average distance between them. There are several methods available for computing the average distance. These include unweighted pair-group average, weighted pair group average, unweighted pair group centroid and weighted pair group centroid.

Unweighted pair-group average (*UPGA*): The distance between two clusters is defined as average distance between all pairs of feature vectors, each of which is in one of the two clusters.

Weighted pair-group average (*WPGA*): This method is identical to the *UPGA*, except that in the computations, the size of the respective clusters (i.e., the number of feature vectors contained in them) is used as a weight. This method is preferred when the cluster sizes are suspected to be greatly uneven.

Unweighted pair-group centroid (*UPGC*): The distance between two clusters is defined as the distance between their centroids. The centroid of a cluster is the mean vector of all the feature vectors contained in the cluster. In this method, if two clusters to be merged are very different in their size, the centroid of the cluster resulting from the merger tends to be closer to the centroid of the larger cluster.

Weighted pair-group centroid (*WPGC*): This method is identical to the *UPGC*, except that feature vectors are weighted in proportion to the size of clusters.

Ward's algorithm (Ward, 1963) is a frequently used technique for regionalisation studies in hydrology and climatology (Acreman and Sinclair, 1986; Hosking and Wallis, 1997;

Kalkstein and Corrigan, 1986; Nathan and McMahon, 1990; Willmott and Vernon, 1980; Winkler, 1985).

The objective function, *W*, of Ward's algorithm (Ward Jr, 1963) minimizes the sum of squares of deviations of the feature vectors from the centroid of their respective clusters. It is based on the assumption that if two clusters are merged, the resulting loss of information, or change in the value of objective function, will depend only on the relationship between the two merged clusters and not on the relationships with any other clusters. The governing equation of Ward's algorithm is written as:

$$W = \sum_{k=1}^{K} \sum_{j=1}^{n} \sum_{i=1}^{N_k} \left(x_{ij}^{k} - x_{.j}^{k}\right)^2 \quad \ldots(4.15)$$

Divisive hierarchical clustering

The divisive hierarchical clustering begins with one large cluster comprising all the N feature vectors and proceeds by splitting them into smaller clusters. The feature vector that has the greatest dissimilarity to other vectors of the cluster is then identified and separated to form a splinter group. The dissimilarity values of the remaining feature vectors in the original cluster are then examined to determine if any additional vectors are to be added to the splinter group. This step divides the original cluster into two parts. The larger cluster is subjected to the aforementioned procedure in the next step. The process continues until a stopping criterion (such as the requested number of clusters) is achieved. The algorithm terminates when the desired number of clusters is obtained. If no stopping criterion is specified, the algorithm terminates when clusters resulting from the analysis are all singleton clusters. Description of divisive clustering algorithms can be found in Murtagh (1983), Guenoche et al. (1991). Savaresi et al. (2002) discussed strategies for the selection of a cluster to be split in divisive clustering algorithms. The divisive clustering methods are yet to be applied in regionalization studies.

Divisive hierarchical clustering algorithms always split clusters. In contrast, agglomerative algorithms always merge clusters.

While hierarchical clustering procedures are not influenced by initialization and local minima, partitional clustering procedures are influenced by initial guesses (e.g. number of clusters, cluster centres, etc.). The partitional clustering procedures are dynamic in the sense that feature vectors can move from one cluster to another to minimize the objective function. In contrast, the feature vectors committed to a cluster in the early stages cannot move to another in hierarchical clustering procedures.

Steps in regionalisation by cluster analysis

The steps in cluster analysis for RRFA applications are noted below:

1. **Selection of attributes:** It is important to select the attributes influencing the flood responses in the study region. Therefore, data exploration of various predictor variables to identify the attributes is carried out in this step.

2. **Preparing feature vectors**: The data available for each attribute are rescaled to nullify differences in their variance and relative magnitude. The rescaling may involve transforming the values of attributes by appropriate transformation function (such as logarithmic) and dividing the transformed values by standard deviation. Each feature vector consists of rescaled (dimensionless) attributes of a catchment.

3. **Forming clusters:** This step involves selection of a clustering algorithm to partition feature vectors prepared in step 2 into disjoint or overlapping clusters. The catchments represented by feature vectors in a cluster constitute a region for flood frequency analysis. In general, distance (or dissimilarity) measure and a clustering criterion characterize a clustering algorithm.

4. **Selecting optimum number of regions:** The clusters formed in step 3 are interpreted visually and by using cluster validity indices to determine the optimum number of regions.

5. **Visual interpretation:** Clusters obtained in step 3 are visually interpreted by plotting them in the geographical space of the study region to identify stable regions. The

stable regions do not change their configuration drastically with a change in the number of clusters formed by the clustering algorithm.

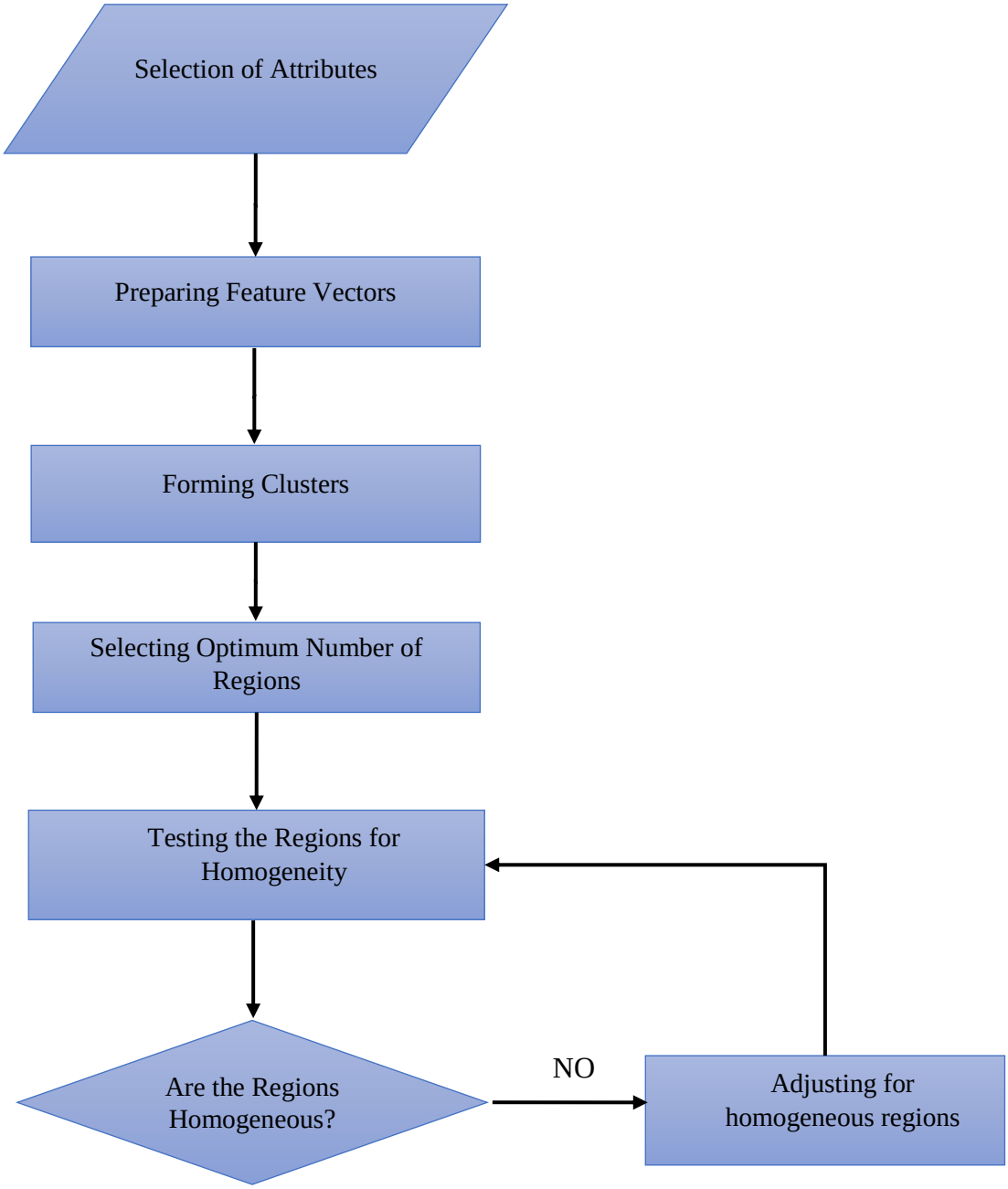

Figure 4.5 Steps in Regionalization using Cluster Analysis

Dissimilarity measures for computing distance between cluster centroids, or feature vectors:

Distance measure:	**Equation:**	
Euclidean	$\sqrt{\sum_{k=1}^{n}\left(x_{ik}-x_{jk}\right)^2}$	…(4.16)
Squared Euclidian	$\sum_{k=1}^{n}\left(x_{ik}-x_{jk}\right)^2$	…(4.17)
Mahalonobis distance	$\sqrt{\left(x_i-x_j\right)^T\sum^{-1}\left(x_i-x_j\right)}$	…(4.18)
Manhattan or City Block	$\sum_{k=1}^{n}\left\|x_{ik}-x_{jk}\right\|$	…(4.19)
Canberra	$\sum_{k=1}^{n}\frac{\left\|x_{ik}-x_{jk}\right\|}{\left\|x_{ik}\right\|+\left\|x_{jk}\right\|}$	…(4.20)
Chebychev	$\max_{1\le k\le n}\left\|x_{ik}-x_{jk}\right\|$	…(4.21)
Cosine	$1-\frac{\sum_{k=1}^{n}x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{n}x_{ik}^2\sum_{k=1}^{n}x_{jk}^2}}$	…(4.22)
Minkowski	$\left(\sum_{k=1}^{n}\left\|x_{ik}-x_{jk}\right\|^t\right)^{\frac{1}{t}}$	…(4.23)

n: number of attributes; x_{ik}: attribute k of feature vector x_i in cluster 1; x_{jk} : attribute k of feature vector x_j in cluster 2; In Mahalanobis distance measure, T is transpose of matrix, and Σ is covariance matrix. If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. t denotes the order of Minkowski distance.

Partitional clustering methods

In partitional clustering procedures, an attempt is made to recover the natural grouping present in the data through a single partition. These procedures are subdivided into K-means and K-medoids methods.

In K-means method (Ball and Hall 1965; MacQueen, 1967), each cluster is represented by its centroid, which is mean (weighted or unweighted average) of feature vectors within the cluster. This method is known for its efficiency in clustering large data sets with numerical

attributes. However, it has limitations in clustering categorical data (Ralambondrainy, 1995; Huang and Ng, 2003). Further, the method is sensitive to the presence of outliers.

4.2.4. Cross validation

Resampling or cross validation is a crucial part of predictive analysis in recent days. This is a method for accuracy checking and evaluating model performance for certain datasets through a recurrent procedure of drawing samples from a selected dataset and refitting the model of interest on each sample. This is a complex procedure, which involves multiple iterations of the same statistical method using different subsets of the training data; it therefore was a computationally expensive and time-consuming procedure in its earlier days. However with the advances of computational capacity in the present, it has become a prerequisite for predictive model development.

The two most commonly used resampling methods are cross validation and bootstrapping. Cross validation is based on the concept of data training of certain set of whole datasets and testing the trained model using the rest part of data set. It is mostly used to assess the test error.

Cross validation can be used to estimate the test error incorporated with the particular statistical learning method with the purpose of evaluating its performance or *model assessment*, or to select the appropriate level of flexibility, which is known as *model selection*.

Bootstrapping is a lengthy procedure comprising multiple random sampling from training dataset and replacing into the samples. This method is a complicated and time-consuming procedure, which is generally used to evaluate the level of accuracy due to a parameter estimation or of a given statistical method.

The concept of selection of the particular statistical method depends on test error; it is chosen if the selected statistical method gives low test error for the given dataset. Test error refers to the average error associated with the predictions of the response on a new observation from using a particular statistical method. This method is chosen in this study considering the lowest test error.

K-fold cross validation

In this study, K-fold cross validation is chosen to evaluate the RFFA model performance. K fold cross validation allows a randomly separate set of observations into *k* groups or folds which are approximately of equal size, and fits the model using the rest of the samples except the first subset or *fold.* The held out dataset is used in order to validate the statistical model through generating predictions using the statistical model based on the test dataset.

This procedure is repeated for *k* times. The mean and standard error values of *k* number of trials are summarised and used subsequently to evaluate the performance of the relationship between the tuning parameter(s) and model utility. The *k*-fold CV estimate is computed by averaging these values:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i \quad \ldots(4.24)$$

The choice of *k* is usually 5 or 10, which depends on dataset. The difference in size between the training set and the resampling subsets gets smaller as the *k* increases. The *bias* of the technique becomes smaller (i.e., the bias is smaller for *k* = 10 than *k* = 5) with difference decrease. In this context, the bias is the difference between the estimated and true values of performance.

The advantage of this method is its flexibility. It does not matter how the data gets divided. Moreover, it has the provision to control the training and test dataset length and number of trial.

In this study, the total dataset consists of 114 catchments; therefore, 10-fold cross validation has been chosen which is reasonable considering the length of the dataset.

The following statistical measures noted below are used to check the suitability and performance of the prediction model, which are:

$$\text{Relative Error (RE)} = \text{Median}\left[abs\left(\frac{Q_{pred}-Q_{obs}}{Q_{obs}}\right)\right] \quad \ldots(4.25)$$

$$\text{Ratio} = \frac{Q_{pred}}{Q_{obs}} \quad \ldots(4.26)$$

Where Q_{obs}=observed flood quantile at each site

Q_{pred} = predicted flood quantile at each site from regional prediction equation.

The relative error and ratio give an indication of the overall performance of the regional prediction model. The model gets better with the minimum value of relative error.

The average value of the $\frac{Q_{pred}}{Q_{obs}}$ provides an indication of the degree of bias of the prediction model. It helps to understand whether there is any systematic overestimation or underestimation prevailing. A value of one indicates good average agreement between Q_{pred} and Q_{obs} as both of the values are random variables. If the ratio value is found in the range of 0.5 to 2, it might be regarded as desirable in RFFA. A value lower than 0.5 might be considered as an underestimated value and value higher than 2 might be considered as overestimated one. Both the Q_{pred} and Q_{obs} values are associated with uncertainties with them; hence, these methods are considered a reasonable guide for checking the accuracy as far as practical application is concerned where a certain level of risk is accepted.

The relative error and ratio values are examined through boxplots. Boxplot is a widely used graphical tool introduced by Tukey (1977). It is a simple plot of five sample quantities: the minimum value; the lower quartile, $q_{0.25}$; the median, $q_{0.5}$; the upper quartile, $q_{0.75}$; and the maximum value. The boxplots can be used to show the location of the median and the associated dispersion of the data at specific probability levels. It is a very useful tool in regards to the cases where there is a high degree of variation in RE values.

4.3. Summary

This chapter provides a description of the statistical and mathematical tools adopted in this study. These include log-log linear model, GAM, cluster analysis, cross validation and evaluation statistics. The fundamental concepts, mathematical equations and input data requirements for each of these methods have been presented in this chapter.

CHAPTER 5

DEVELOPMENT OF LOG-LOG LINEAR MODEL

5.1. General

This chapter focuses on the development of new prediction equations for regional flood estimation using log-log linear model in Quantile Regression Technique (QRT) framework. Six average recurrence intervals (ARIs) (2, 5, 10, 20, 50 and 100 years) are considered. In forming regions, both the fixed region and region of influence (ROI) approaches are adopted. To assess the performance of the developed prediction equations, a 10-fold cross validation approach is adopted for the total catchment flood data.

5.2. Log transformation of variables

Log transformation is generally made on both flood quantiles (dependent variables) and catchment characteristics (independent variables) dataset to change the scale of the variables to achieve linearity or near-linearity. This is very common in RFFA.

5.2.1. Development of prediction equations using log-log linear method

Log-log linear regression analysis was carried out considering all the 114 catchments from Victoria as a single group. The location of the catchment has been shown in Chapter 3 (Figure 3.1). The prediction equations are developed following a backward stepwise regression approach.

The data from 114 catchments has been log transformed in order to develop the log-log linear model. Log transformation has been done both on flood quantiles (dependent variables) and catchment characteristics (independent variables). Linear regression analysis has been done using the dataset and backward stepwise procedure has been followed to choose the particular catchment characteristics for model development. The diagnostic statistics for the model relevant to different ARIs has been given in Table 5.1. The detailed results for Q_2 is provided below. Results of the remaining ARIs can be found in Appendix B (Figure B.1 to B.15).

For Q_2 model, three catchment characteristics are found to be statistically significant from log-log linear regression analysis which are catchment area (area), rainfall intensity ($I_{6,2}$) and

stream density (*sden*). The important properties of residuals are shown in Figure 5.1, 5.2 and 5.3.

Figure 5.1 represents the standardised residual vs fitted predicted value graph for Q_2 model. From this plot it can be observed that most of the residuals are scattered around the 0 line (black dotted line) which indicates that there are no trends in the residuals. The results show slight heterogeneity of variances near fitted value of 1.5. Overall, it indicates slight heteroscedasticity between predicted value and residuals; however, it appears to be linear.

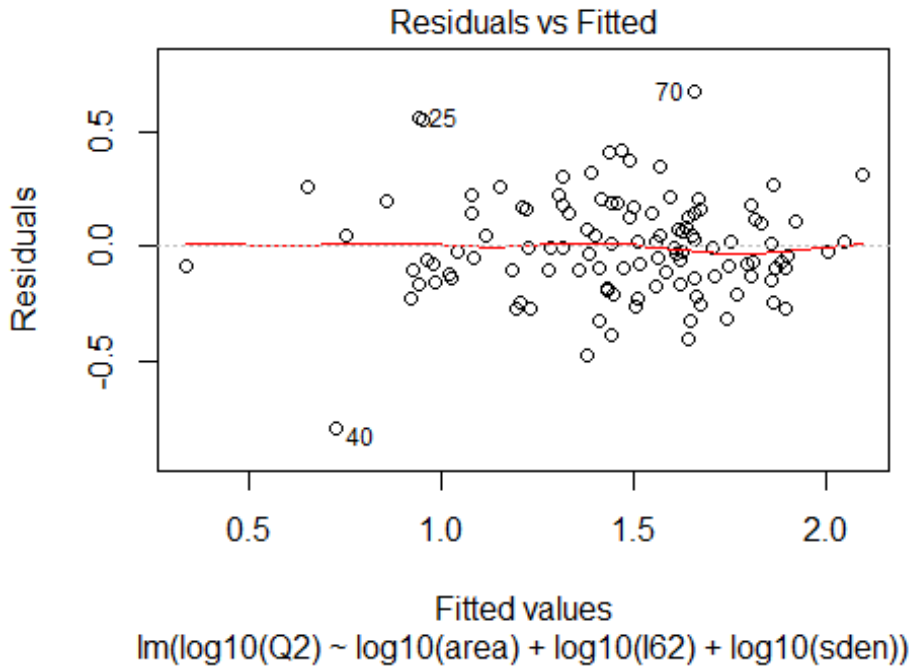

Figure 5.1 Standardised residual vs fitted predicted value for the log-log linear model for combined group for Q_2

Figure 5.2 represents the normal Q-Q plot for the standardised residuals. The plot shows that the standardised residuals follow normal distribution. Most of the points are plotted around the trend line which indicates that the standardised residuals are near normally distributed.

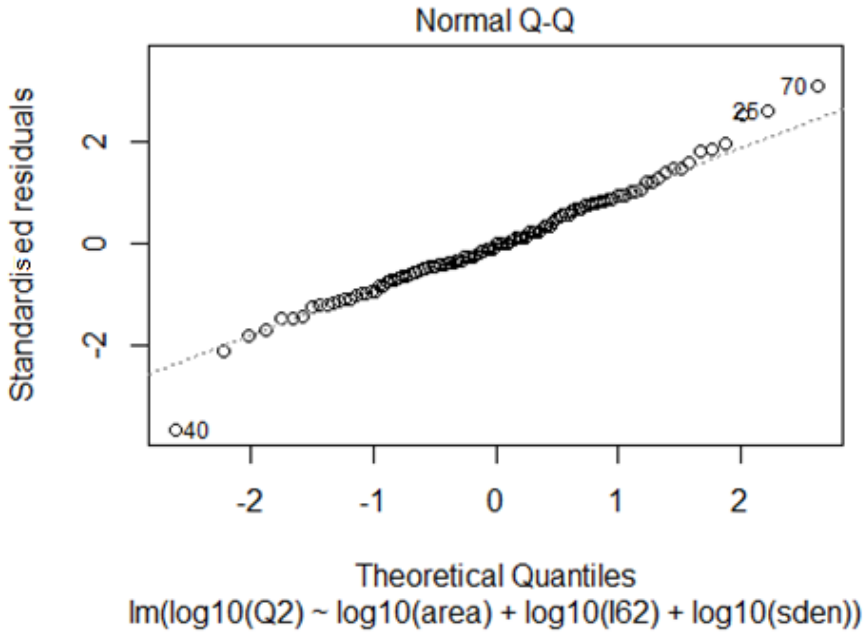

Figure 5.2 Normal Q-Q plot for the standardised residuals for the log-log linear model for combined group for Q_2

Figure 5.3 represents the scale-location plot between predicted values and standardised residual for Q_2 model. The plot exhibits a slight deviation from the red smooth line which indicates a slight heteroscedasticity in variances.

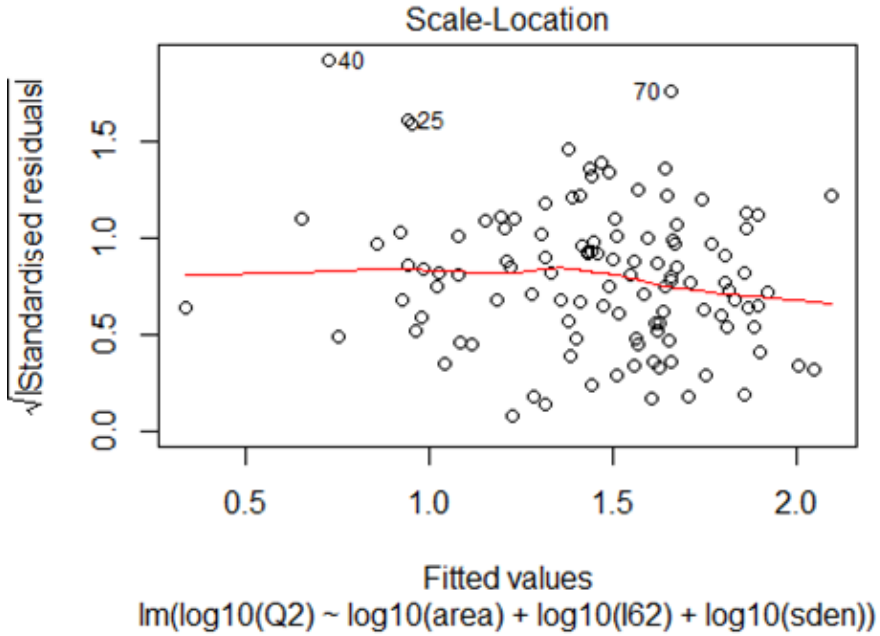

Figure 5.3 Scale-location plot between predicted values and standardised residuals for the log-log linear model for combined group for Q_2

Model Development Statistics

Table 5.1 shows the overall model statistics for the 6 different ARIs. The major determinants are coefficient of determination (R^2), *p*-statistics and standard error of estimate (*SEE*). From Table 5.1 it is found that the R^2 values of log-log linear model range from 0.69 to 0.53 respectively for Q_2 to Q_{100}. The R^2 values are found particularly smaller for higher ARIs which indicates towards the larger variance of prediction in estimation of higher ARI floods. All the R^2 values are quite reasonable and indicates a good linear fit for the prediction equations.

The SEE vary from 0.22 to 0.32 respectively for Q_2 to Q_{100}. The lowest value of SEE is found for Q_2 and highest is found for Q_{100}. This indicates that the percentage of error increase with higher ARIs.

The predictor variables for individual models are selected considering *p*-statistics for respective models. The predictor variables selected in the final model with the *p*-statistics value of ≤ 0.10. Table 5.1 contains all the selected predicted values for individual models along with respective *p*-statistics. It reveals that the *area* and $I_{6,2}$ appear to be most important variables for estimating Q for log-log linear model. These two variables are common with all the prediction equations. The next most important predictor variable is found as *rain* which appears in every prediction model except for Q_2 and Q_5. Only for Q_2, *sden* is selected whereas *rain* is absent as predictor variable. For Q_5, both *rain* and *sden* are selected as predictor variable. Overall, the prediction equations show consistency in selection of independent variables except for Q_2 and Q_5.

The developed prediction equations given below:

$$\log Q_2 = -2.42 + 0.68\log(area) + 1.48\log(I_{6,2}) + 0.39\log(sden) \quad \ldots(5.1)$$

$$\log Q_5 = -1.60 + 0.68\log(area) + 1.74\log(I_{6,2}) - 0.29(rain) + 0.31\log(sden) \quad \ldots(5.2)$$

$$\log Q_{10} = -1.25 + 0.66\log(area) + 2.14\log(I_{6,2}) + 2.30\log(rain) \quad \ldots(5.3)$$

$$\log Q_{20} = -1.00 + 0.66\log(area) + 2.30\log(I_{6,2}) - 0.66\log(rain) \quad \ldots(5.4)$$

$$\log Q_{50} = -0.79 + 0.66\log(area) + 2.45\log(I_{6,2}) - 0.76\log(rain) \quad \ldots(5.5)$$

$$\log Q_{100} = -0.70 + 0.66\log(area) + 2.54\log(I_{6,2}) - 0.81\log(rain) \quad \ldots(5.6)$$

Table 5.1 Model statistics for log-log linear model of combined group

Equation	Predictor variables	Regression Coefficient (β)	Standard Error	Standard Error of Estimate (*SEE*)	R^2	*p* value	D.F
log Q_2	(constant)	-2.42	0.52	0.22	0.69	9.0E-06	110
	log (*area*)	0.68	0.04			< 2e-16	
	log ($I_{6,2}$)	1.48	0.33			1.6E-05	
	log (*sden*)	0.39	0.15			1.4E-02	
log Q_5	(constant)	-1.60	0.57	0.23	0.67	6.3E-03	109
	log (*area*)	0.68	0.05			< 2e-16	
	log ($I_{6,2}$)	1.74	0.41			4.6E-05	
	log (*rain*)	-0.29	0.19			1.2E-01	
	log (*sden*)	0.31	0.16			6.2E-02	
log Q_{10}	(constant)	-1.25	0.62	0.25	0.63	4.7E-02	110
	log (*area*)	0.66	0.05			< 2e-16	
	log ($I_{6,2}$)	2.14	0.43			3.0E-06	
	log (*rain*)	-0.53	0.20			8.3E-03	
log Q_{20}	(constant)	-1.00	0.66	0.27	0.61	1.4E-01	110
	log (*area*)	0.66	0.05			< 2e-16	
	log ($I_{6,2}$)	2.30	0.46			2.7E-06	
	log (*rain*)	-0.66	0.21			2.5E-03	
log Q_{50}	(constant)	-0.79	0.73	0.30	0.57	2.8E-01	110
	log (*area*)	0.66	0.06			< 2e-16	
	log ($I_{6,2}$)	2.45	0.51			4.5E-06	
	log (*rain*)	-0.76	0.23			1.4E-03	
log Q_{100}	(constant)	-0.70	0.78	0.32	0.53	3.7E-01	110
	log (*area*)	0.66	0.06			< 2e-16	
	log ($I_{6,2}$)	2.54	0.54			8.5E-06	
	log (*rain*)	-0.81	0.25			1.5E-03	

The log-log linear models are evaluated based on the following criteria (see Chapter 4 for details):

- Q_{pred}/Q_{obs} ratio
- Plot of Q_{obs} and Q_{pred}
- Median relative error (RE)

5.2.2. Adequacy of developed log-log linear model

To assess the model fit, the plot of Q_{obs} and Q_{pred} , Q_{pred}/Q_{obs} ratio and median relative error values are used. Here the data for all the 114 catchments are used in developing the model. Figure 5.4 shows Q_{obs} and Q_{pred} plot for Q_{20}. Most of the catchments are within a narrow range of scatter from the 45-degree line except for a few outliers. Overall, the plot shows a good match between Q_{obs} and Q_{pred}. The Q_{obs} and Q_{pred} plots for all the six ARIs are shown in Figures B.1 to B.6. It is found from these plots that the degree of scatter in Q_{obs} and Q_{pred} values are remarkably smaller for Q_2, Q_5 and Q_{10} as compared to Q_{20}, Q_{50} and Q_{100}. This indicates that the model error increases with increasing ARI, which is as expected.

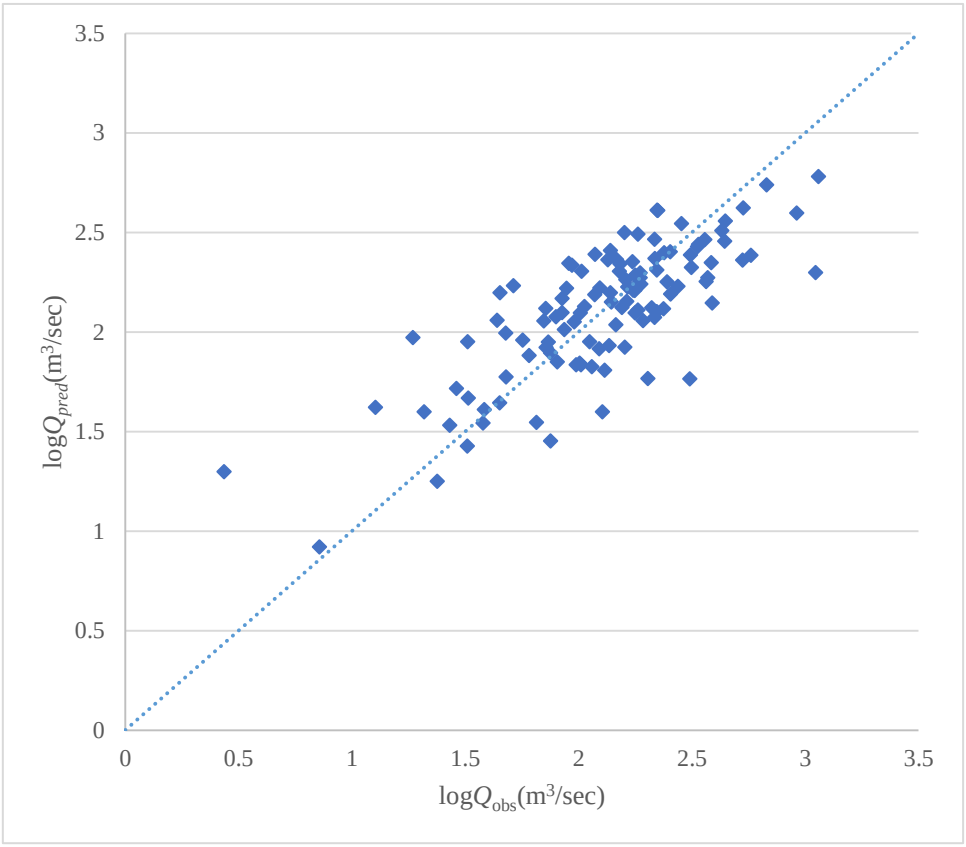

Figure 5.4 Comparison of observed and predicted flood quantiles for log-log linear model of combined group for Q_{20}

Figure 5.5 shows the boxplots of RE values for the log-log linear model Q_{20}. From this figure, it is revealed that the median RE values (represented by the black line within a box)

match with the 0 – 0 line very well for ARIs of 5 and 20 years, and quite reasonably for ARIs of 10 and 50 years. For ARI of 2 years, some underestimations are noticed. For ARI of 100 years, the underestimation is remarkable. In terms of the RE band, which is represented by the total spread of the box, ARI of 2 years shows the lowest spread. The RE band for 10 years ARI is very similar to that of 2 years ARI. The RE spreads for ARIs of 5, 20, 50 and 100 years are much higher than ARIs of 2 and 10 years. The RE band for 100 years ARI is more than double to that of 2 and 10 years. These results show that in terms of RE, the best result is achieved for 10 years ARI, followed by 2 years ARIs. This demonstrates that higher ARI flood quantiles are associated with a greater degree of uncertainty as represented by a higher degree of spread in the RE. This is very similar to the findings by Haddad and Rahman (2012) and Rahman et al. (2011).

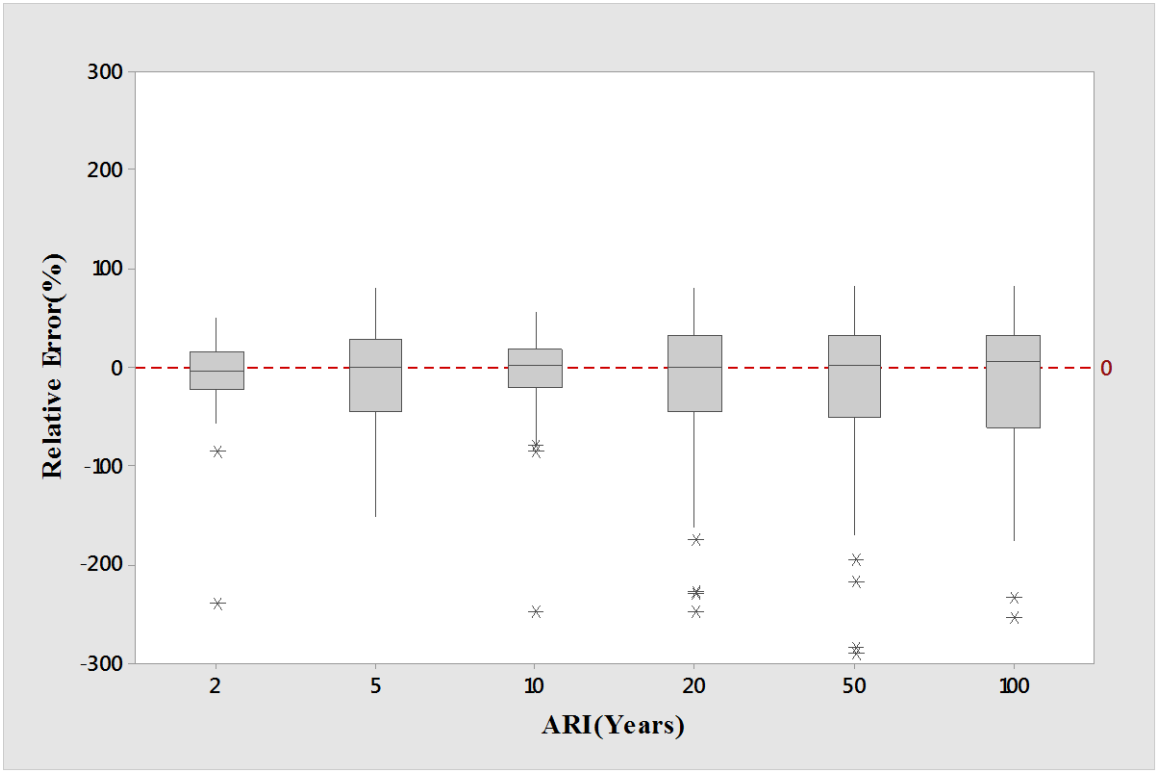

Figure 5.5 Boxplots of relative error RE values for log-log linear model of combined group

Figure 5.6 presents the boxplot of the Q_{pred}/Q_{obs} ratio values of the selected 114 catchments for the log-log linear model. It is found that the median Q_{pred}/Q_{obs} ratio values (represented by

the thick black lines within a box) are located closer to 1 – 1 line (the horizontal line in Figure 5.6), in particular for ARIs of 2, 5,10, 20 and 50 years (the best agreement is for ARI of 20 years). However, for ARI of 100 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance below the 1 – 1 line, and for ARI of 2 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance above the 1 – 1 line. In terms of the spread of the Q_{pred}/Q_{obs} ratio values, ARI of 2 years exhibits the lowest spread followed by an ARI of 10 years. Furthermore, the spreads of the Q_{obs}/Q_{pred} ratio values for 50 and 100 years are very similar, which are remarkably larger than 2 and 10 years.

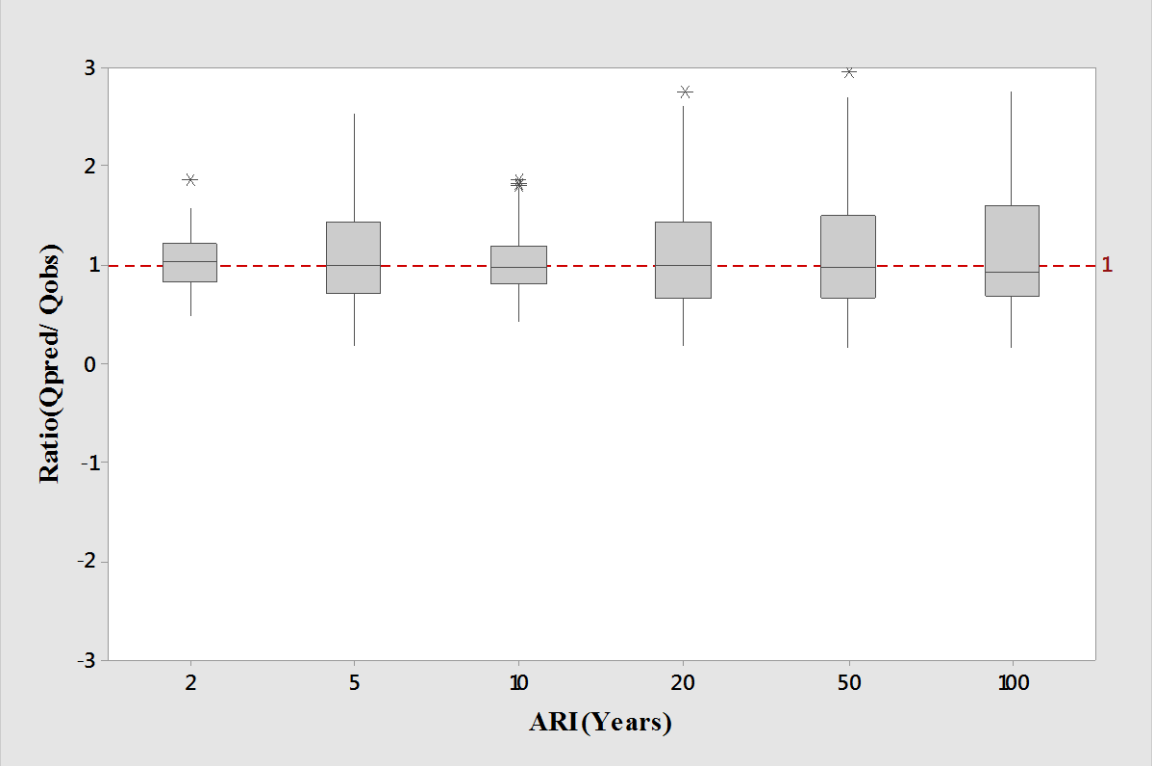

Figure 5.6 Boxplots of Q_{pred}/Q_{obs} ratio values for log-log linear model of combined group

5.3. Regions based on catchment characteristics data

Cluster analysis is carried out to identify groups of catchments in catchment characteristic data space. Both hierarchical and partitioned clustering are carried out in this study. Eight catchment characteristics variables are adopted to form groups by cluster analysis (see Table 3.1).

5.3.1. Cluster analysis

In the cluster analysis, the variables are standardised and are given equal weights. The hierarchical clustering is used with a combination of Wards-Manhattan method, as discussed in Chapter 4. The groups formed by hierarchical clustering are illustrated through the dendogram in Figure 5.7. The best results obtained from cluster analysis are summarised in Table 5.2, which delivers two groupings: A1 (79 stations) and A2 (35 stations) from Wards-Manhattan clustering and B1 (67 stations) and B2 (47 stations) from K-Means clustering (Appendix A).

It should be noted that A1 is the biggest cluster group containing 69 % of the catchments and B1 contains the remaining 31 % of the catchments. The A1 group has two sub-clusters, however, they have not been used in model testing in this thesis. The B1 contains 58 % of the catchments while B2 contains the remaining 42 % of the catchments. Further sub-division of B1 and B2 groups have not been considered.

Table 5.2 Groups Formed by Cluster Analysis

Method	Total no. of stations	Grouping	Grouping
Wards-Manhattan Cluster combination	114	79 (A1)	35(A2)
K-Means Cluster	114	67 (B1)	47 (B2)

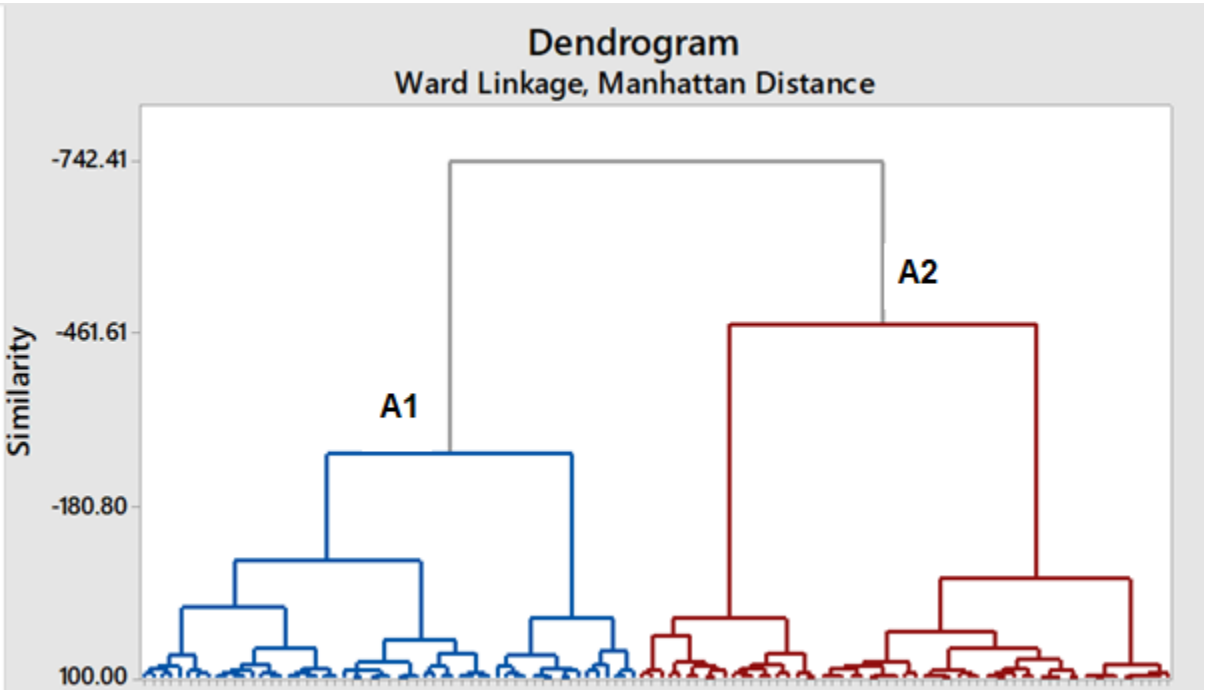

Figure 5.7 Dendogram Using Ward Linkage Manhattan Distance Between Groups

5.3.2. Evaluation of log-log linear models (clustering group A1)

The model statistics for log-log linear model for A1 is listed below in Table 5.3. The R^2 values are ranged from 0.74 to 0.62 respectively for Q_2 to Q_{100} which indicates toward the lower accuracy of predictions for higher ARIs. Although, the R^2 values follow a decreasing trend from lower to higher ARIs, the largest to smallest value of R^2 does not have a large variation for this model. All the R^2 values seem to be quite reasonable and indicate a good linear fit for the prediction equations.

The SEE values vary from 0.21 to 0.29 respectively for Q_2 to Q_{100}. The lowest value of SEE is found for Q_2 and highest is found for Q_{100}. Larger SEE values indicate toward the associated percentage of error increase with higher ARIs.

The predictor variables for individual models are selected considering the *p*-statistics value for the respective model. The final predictor variables are chosen for each prediction model where the *p*-statistics value is ≤ 0.10. All the predictor variables for log-log linear model for clustering group A1 is listed in Table 5.3 along with the respective *p*-statistics.

From Table 5.3, *area* and $I_{6,2}$ are found as the most feasible predictor variables for estimation of Q for log-log linear model for clustering group A1. These two common variables are present for all the prediction models developed for the 6 ARIs. The next important variables found are *rain* and S1085, which are found in all the prediction models except for Q_2. Only for Q_2, *forest* is selected as a predictor variable whereas *rain* and S1085 is absent. Overall, the prediction equations show consistency in selection of predictor variables except for Q_2.

The developed model equations are:

$$\log Q_2 = -2.08 + 0.69\log(area) + 1.26\log(I_{6,2}) - 0.25\log(forest) \quad \ldots(5.7)$$

$$\log Q_5 = -.78 + 0.56\log(area) + 2.022\log(I_{6,2}) - 0.48(rain) - 0.34\log(\text{S1085}) \ldots(5.8)$$

$$\log Q_{10} = -.49 + 0.56\log(area) + 2.31\log(I_{6,2}) - 0.66\log(rain) - 0.36\log(\text{S1085}) \ldots(5.9)$$

$$\log Q_{20} = -.32 + 0.55\log(area) + 2.52\log(I_{6,2}) - 0.77\log(rain) - 0.37\log(\text{S1085}) \ldots(5.10)$$

$$\log Q_{50} = -0.21 + 0.55\log(area) + 2.73\log(I_{6,2}) - 0.86\log(rain) - .38\log(\text{S1085}) \ldots(5.11)$$

$$\log Q_{100} = -0.16 + 0.55\log(area) + 2.85\log(I_{6,2}) - 0.91\log(rain) - .38\log(\text{S1085}) \ldots(5.12)$$

Table 5.3 Model statistics for log-log linear model of clustering group A1

Equation	Predictor variables	Regression Coefficient (β)	Standard Error	Standard Error of Estimate (*SEE*)	R^2	*p* value	D.F
log Q_2	(constant)	-2.08	0.67	0.21	0.74	0.00278	75
	log (*area*)	0.69	0.05			< 2e-16	
	log ($I_{6,2}$)	1.26	0.42			0.00379	
	log (*forest*)	-0.25	0.10			0.01372	
log Q_5	(constant)	-0.78	0.79	0.23	0.72	3.25E-01	74
	log (*area*)	0.56	0.06			2.93E-13	
	log ($I_{6,2}$)	2.02	0.48			7.56E-05	
	log (*rain*)	-0.48	0.21			2.75E-02	
	log (*S*1085)	-0.34	0.12			5.40E-03	
log Q_{10}	(constant)	-0.49	0.84	0.24	0.70	0.56199	74
	log (*area*)	0.56	0.07			3.35E-12	
	log ($I_{6,2}$)	2.31	0.51			2.30E-05	
	log (*rain*)	-0.66	0.22			0.00444	
	log (S1085)	-0.36	0.13			0.00524	
log Q_{20}	(constant)	-0.32	0.89	0.26	0.68	7.19E-01	74
	log (area)	0.55	0.07			3.34E-11	
	log ($I_{6,2}$)	2.52	0.54			1.44E-05	
	log (*rain*)	-0.77	0.24			0.00174	
	log (s1085)	-0.37	0.13			0.00646	
log Q_{50}	(constant)	-0.21	0.96	0.28	0.65	8.32E-01	74
	log (*area*)	0.55	0.08			5.67E-10	
	log ($I_{6,2}$)	2.73	0.59			1.48E-05	
	log (*rain)*	-0.87	0.26			0.00122	
	log (s1085)	-0.38	0.14			0.00998	
log Q_{100}	(constant)	-0.16	1.03	0.29	0.62	0.873	74
	log (*area*)	0.55	0.08			4.12E-09	
	log ($I_{6,2}$)	2.85	0.63			2.04E-05	
	log (*rain*)	-0.91	0.27			0.0014	
	log (*S*1085)	-0.38	0.15			0.0145	

Adequacy Checking of Model

For each of the groups in cluster analyses, a log-log linear regression model is developed. To assess the model performance, the plot of Q_{obs} and Q_{pred} , Q_{pred}/Q_{obs} ratio and median RE

values are examined for log-log linear model for clustering group A1 (Figures 5.8, 5.9 and 5.10).

Figure 5.8 illustrates the scatter plot of observed and predicted flood quantiles for clustering group A1 for Q_{20}. The remaining scatter plots of predicted and observed flood quantiles for clustering group A1 for all the ARIs are included in Appendix C (Figures C.6 to C.10). Most of these plots generally represent a good agreement between the predicted and observed flood quantiles; however, there are some under estimations by the higher discharges, in particular for ARIs of 20, 50 and 100 years. Most of the catchments are within a narrow range of scatter from the 45-degree line except for a few outliers. The variability of scatter from the gradient line is found particularly larger for higher discharges. Overall, the log-log linear model for A1 clustering group shows reasonable performance with respect to Q_{pred} and Q_{obs} plots.

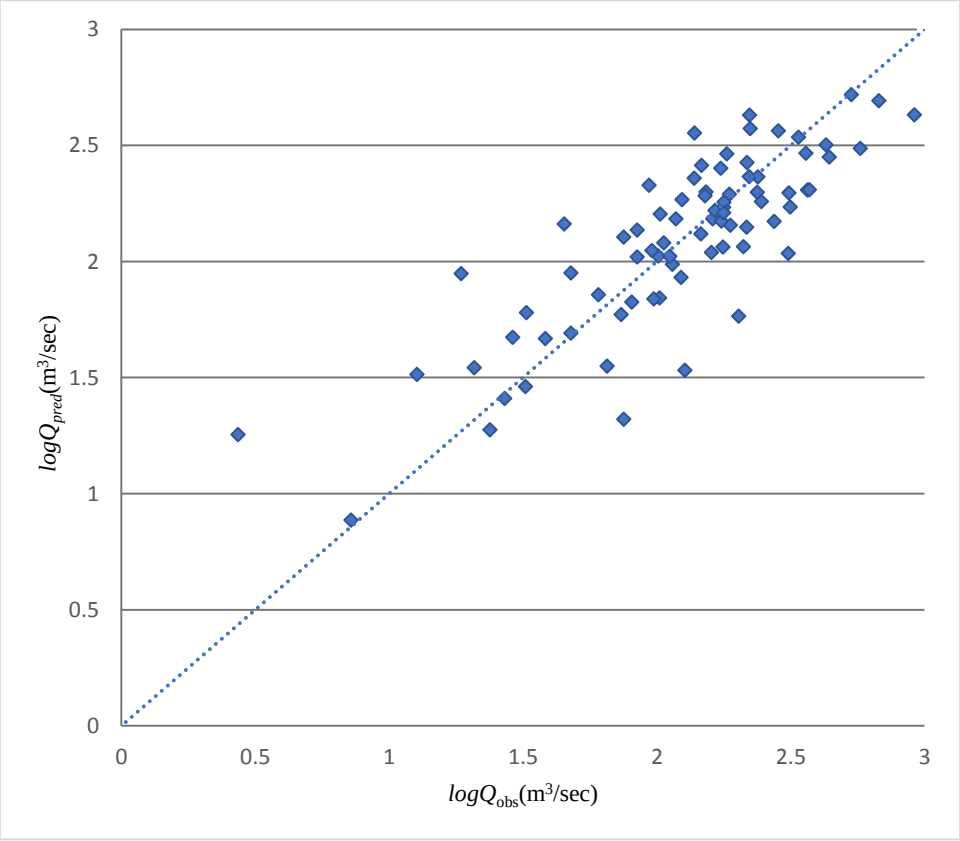

Figure 5.8 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A1 for Q_{20}

Figure 5.9 shows the boxplots of RE values for the log-log linear model for clustering group A1. The median RE values match with the 0 – 0 line very well for ARIs of 50 and 100 years. For ARI of 2 years, a small degree of underestimation is noticed. For ARIs of 5, 10 and 20 years, a small degree of overestimation is noticed. In terms of the RE band, ARIs of 2 and 5 years show the lowest spread, which is slightly lower than RE band of 10 years ARI. The ARIs of 20, 50 and 100 years show a much higher spread. According to RE band, it is revealed that the performance of log-log linear model for cluster group A1 is relatively poor for higher ARIs (i.e. 50 and 100 years).

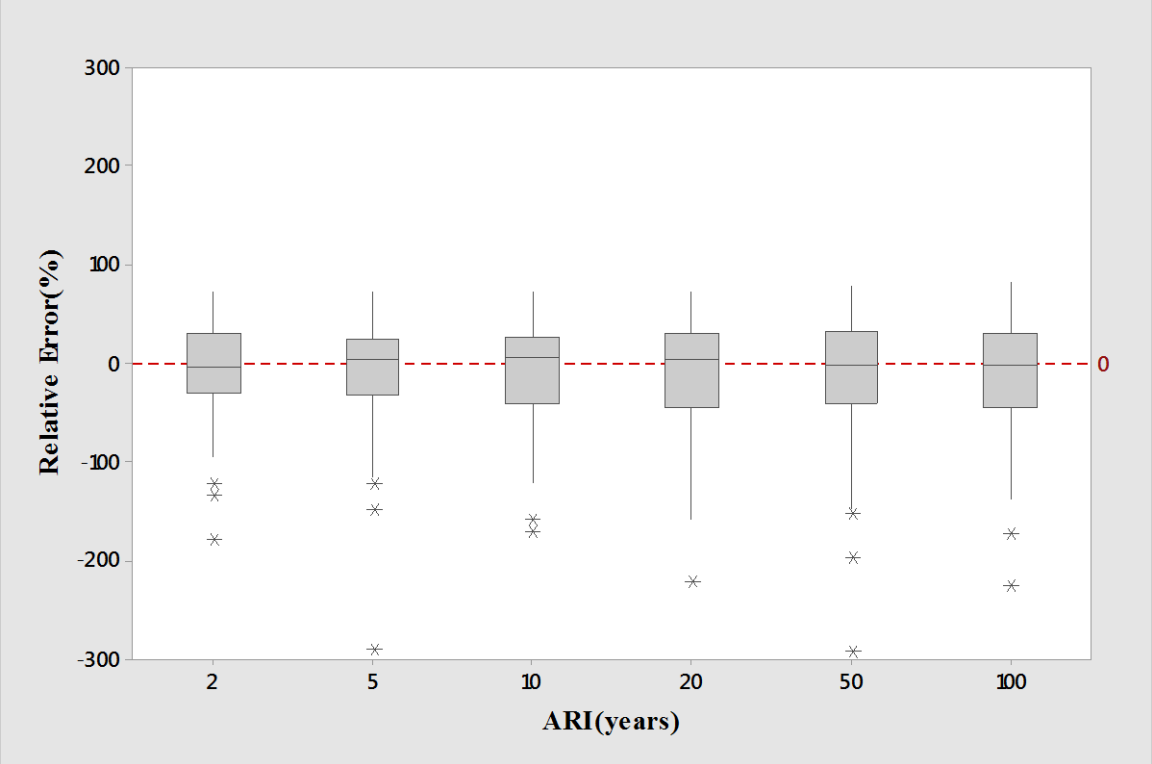

Figure 5.9 Boxplots of RE values for log-log linear model of clustering group A1

Figure 5.10 presents the boxplots of the Q_{pred}/Q_{obs} ratio values for clustering group A1. It is found that the median Q_{pred}/Q_{obs} ratio values are located closer to 1 – 1 line, in particular for ARIs of 50 and 100 years (the best agreement is for ARI of 100 years). However, for ARI of 10 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance below the 1 – 1 line and for ARI of 2 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance above the 1 – 1 line. These results indicate noticeable underestimations and overestimations of the predicted flood quantiles by the log-log linear model for 2 and 10 years ARIs. In terms of the spread of the Q_{pred}/Q_{obs} ratio values, ARI of 5 years exhibits the lowest spread, followed by

ARI of 10 years. The spreads for ARIs of 10, 20, 50 and 100 years are very similar, which are slightly larger than that of ARIs of 2 and 5 years.

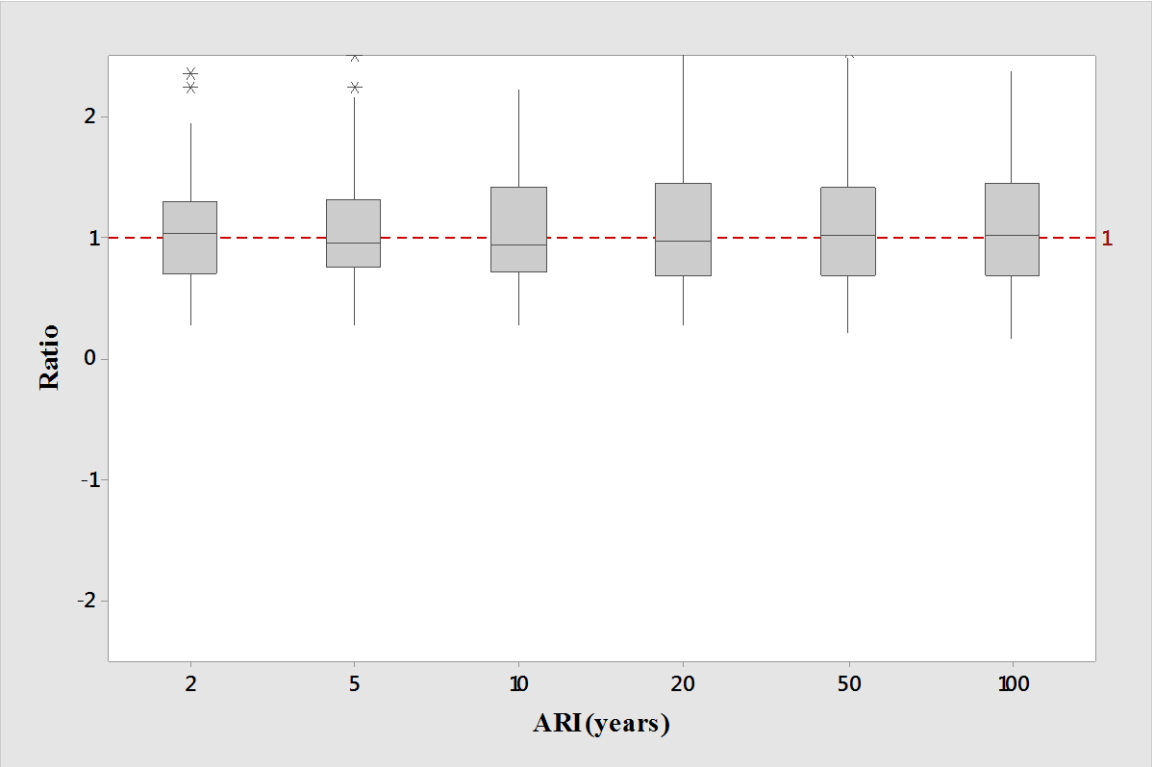

Figure 5.10 Boxplots of Q_{pred}/Q_{obs} ratio values for log-log linear model of clustering group A1

5.3.3. Evaluation of log-log linear model performance (clustering group A2)

Model development

The model is developed considering the same determinants as before which are R^2, SEE and *p*-statistics. The model statistics for log-log linear model for A2 is found from Table 5.4 below. From the R^2 values, it is observed that the values range showing large variations from 0.69 to 0.27 respectively for Q_2 to Q_{100}. The large variations from lower to higher ARIs for this model indicate toward larger uncertainty associated with higher ARIs for this model. Moreover, particularly small R^2 values for higher ARIs (e.g., Q_{50} and Q_{100}) indicate towards the larger variance of prediction in estimation of higher ARI flows. Most of the R^2 values seem to be relatively low except for Q_2 and Q_5 which indicates towards poor prediction accuracy for higher ARIs for this model.

The SEE values vary from 0.19 to 0.34 respectively for Q_2 to Q_{100}. The lowest value of residual standard error was found for Q_2 and highest was found for Q_{100} which indicates towards the higher percentage of prediction error associated with higher ARIs.

The most important predictor variable found for the model is *area,* which is common in every prediction model. The second most important independent variable is found as *sden*, which is present in every model except for Q_{20}. Only for $Q_{2,}$ *rain* is found as a functioning predictor variable in final model. Overall the prediction models are found to be consistent in selection of predictor variables.

Table 5.4 Model statistics for log-log linear model of clustering group A2

Equation	Predictor variables	Regression Coefficient (β)	Standard Error	Standard Error of Estimate (*SEE*)	R^2	*p* value	D.F
log Q_2	(constant)	-4.77	1.01	0.19	0.69	5.00E-05	31
	log (*area*)	0.80	0.10			5.47E-09	
	log (*rain*)	1.47	0.34			1.39E-04	
	log (*sden*)	0.74	0.16			4.35E-05	
log Q_5	(constant)	-0.07	0.29	0.22	0.55	8.16E-01	32
	log (*area*)	0.74	0.11			2.90E-07	
	log (*sden*)	0.62	0.18			1.32E-03	
log Q_{10}	(constant)	0.14	0.33	0.24	0.48	6.74E-01	32
	log (*area*)	0.72	0.13			2.58E-06	
	log (*sden*)	0.58	0.20			5.77E-03	
log Q_{20}	(constant)	-3.44	1.06	0.31	0.43	2.08E-03	32
	log (*area*)	0.68	0.08			4.17E-11	
	log ($I_{6,2}$)	2.66	0.67			1.85E-04	
log Q_{50}	(constant)	0.47	0.42	0.31	0.32	2.77E-01	32
	log (*area*)	0.70	0.16			1.66E-04	
	log (*sden*)	0.48	0.26			6.80E-02	
log Q_{100}	(constant)	0.58	0.47	0.34	0.27	2.26E-01	32
	log (*area*)	0.70	0.18			5.46E-04	
	log (*sden*)	0.44	0.28			1.27E-01	

Overall, the model equations can be written as;

$$\log Q_2 = -4.77 + 0.80 \log(area) + 1.47 \log(rain) + .74\log(sden) \quad \ldots(5.13)$$

$$\log Q_5 = -.07 + 0.74 \log(area) + .62\log(sden) \quad \ldots(5.14)$$

$$\log Q_{10} = 0.14 + 0.72\log(area) + .58\log(sden) \quad \ldots(5.15)$$

$$\log Q_{20} = -3.44 + 0.68\log(area) + 2.66\log(I_{6,2}) \quad \ldots(5.16)$$

$$\log Q_{50} = .47 + 0.70\log(area) + .48\log(sden) \quad \ldots(5.17)$$

$$\log Q_{100} = .58 + 0.70\log(area) + .44\log(sden) \quad \ldots(5.18)$$

Adequacy checking of model

To assess the model performance, the plot of Q_{obs} and Q_{pred}, Q_{pred}/Q_{obs} ratio and median RE values are computed for clustering group A2 (Figures 5.11, 5.12 and 5.13).

Figure 5.11 shows a reasonable scatter between the observed and predicted flood quantiles for clustering group A2 for Q_{20}. Overall, the scatter around the 45-degree line in this figure is deemed to be reasonable for most of the catchments. The plots of observed and predicted flood quantiles for all the six return periods can be seen in Appendix C (Figures C.11 to C.15). Results for ARIs of 2 and 5 years (Figures C.11 to C.12, respectively) are relatively better as compared with other ARIs.

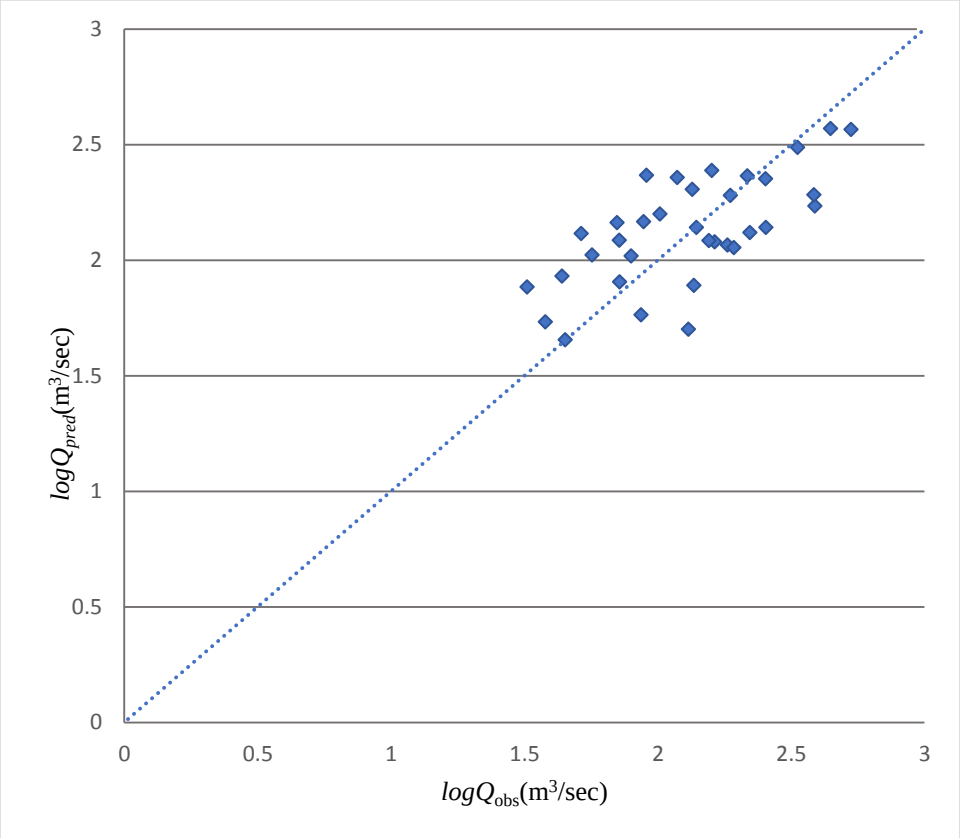

Figure 5.11 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A2 for Q_{20},

Figure 5.12 shows the boxplots of RE values for the log-log linear model for clustering group A2. The median RE values match with the 0 – 0 line very well for ARI of 2, 5, 10 and 20 years and reasonably well for ARIs of 50 and 100 years. For ARIs of 50 and 100 years, slight overestimations are noticed. In terms of the RE band, ARI of 2 years shows the lowest spread. The spread of RE increases with increasing ARI. The RE band for 100 years ARI is more than double to that of 2 and 5 years ARIs. These results show that in terms of RE, the overall best result is achieved for 2 years ARI. The results for higher ARIs (20, 50 and 100 years) are relatively poor, i.e. too high spread in RE values, indicating a higher model error.

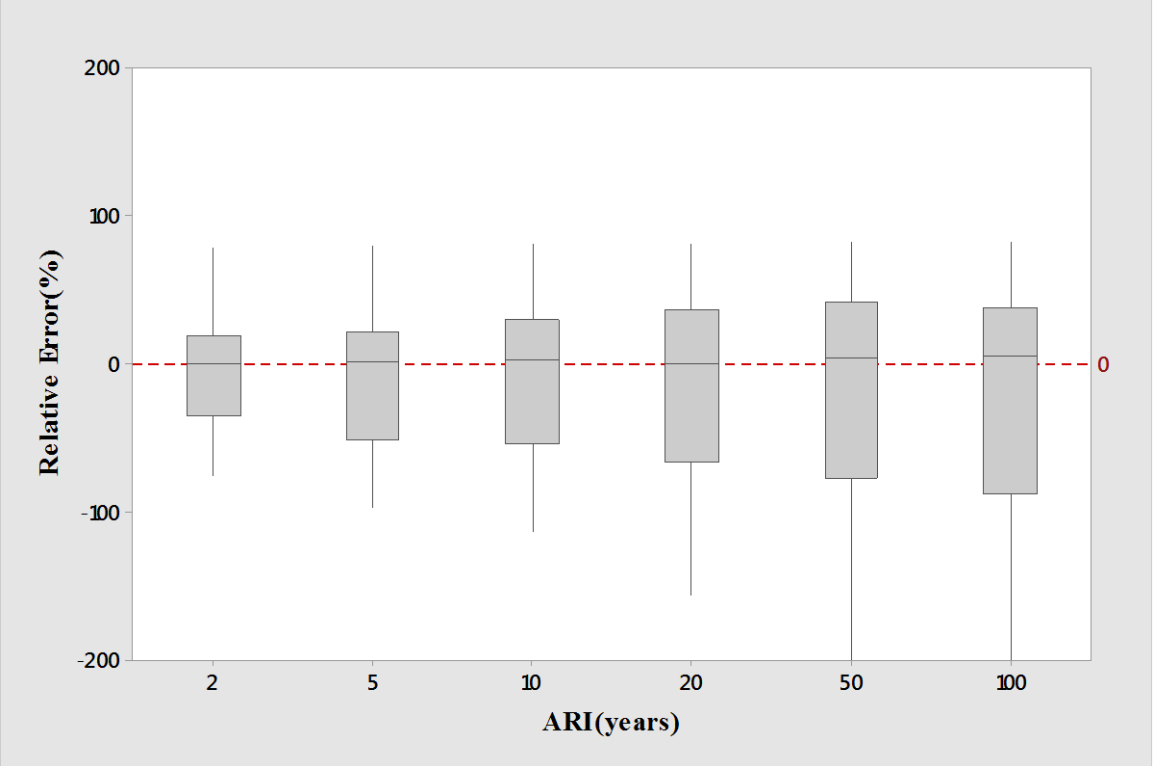

Figure 5.12 Boxplots of RE for log-log linear model of clustering group A2

Figure 5.13 presents the boxplots of the Q_{pred}/Q_{obs} ratio values for clustering group A2 for different ARIs. It is found that the median Q_{pred}/Q_{obs} ratio values are located closer to 1 – 1 line, in particular for ARIs of 2, 5 and 20 years. However, for ARIs of 10, 50 and 100 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance below the 1 – 1 line, indicating a negative bias. Also, most of the Q_{pred}/Q_{obs} ratio values for ARIs of 20, 50 and 100 years are located above the 1 – 1 line, indicating overestimation by the log-log model for many catchments. In terms of the spread of the Q_{pred}/Q_{obs} ratio values, ARI of 2 years exhibits the lowest spread, followed by ARIs of 5, 10, 20, 50 and 100 years. Furthermore, the spreads of the Q_{pred}/Q_{obs} ratio values for 50 and 100 years are very similar, which are remarkably larger than 2, 5 and 10 years. It indicates a comparatively higher range of overestimation of flood quantiles for larger ARI values for clustering group A2.

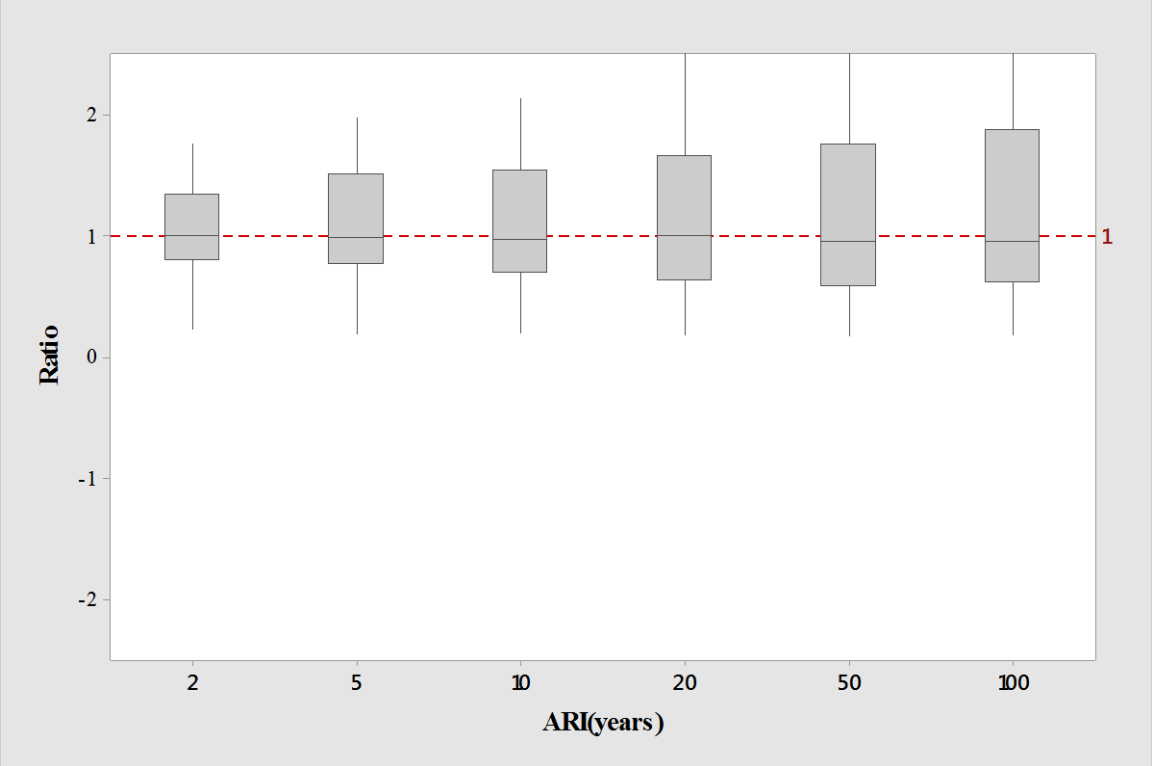

Figure 5.13 Boxplots of Q_{pred}/Q_{obs} ratio values for log-log linear model of clustering group A2

5.3.4. Evaluation of log-log linear model performance (clustering group B1)

Model development

The model statistics for log-log linear model for B1 are illustrated in Table 5.5 below. R^2 values range from 0.78 to 0.62 respectively for Q_2 to Q_{100} following a linear trend. R^2 values are found particularly smaller for higher ARIs which is not uncommon considering the associated uncertainties for prediction of flood quantiles for higher ARIs. All the R^2 values seem to be quite reasonable and indicate a good linear fit for the prediction equations.

The SEE values vary following a linear trend from 0.21 to 0.32 respectively for Q_2 to Q_{100}. The lowest value of residual standard error is found for Q_2 and highest is found for Q_{100}. The predictor variables selected for log-log linear model for group B1 are described in Table 5.5. These predictor variables are selected based on *p*-statistics value where p-statistics ≤ 0.10. The most statistically important predictor variables which are found in every prediction model are *area*, $I_{6,2}$ and S1085. The second most important predictor variable is found is *rain*, which is present in almost every log-log linear model for group B1 except for Q_2. For Q_2,

evap is found to be statistically significant whereas *rain* is absent. Overall, the prediction equations are rather consistent with the selection of independent variables except for Q_2.

Table 5.5 Model statistics for log-log linear model of clustering group B1

Equation	Predictor variables	Regression Coefficient (β)	Standard Error	Standard Error of Estimate (*SEE*)	R^2	*p* value	D.F
log Q_2	(constant)	-10.03	4.47	0.21	0.78	0.02846	62
	log (*area*)	0.54	0.06			4.12E-12	
	log ($I_{6,2}$)	1.53	0.49			0.00299	
	log (*evap)*	2.74	1.52			0.0758	
	log (S1085)	-0.31	0.13			0.01773	
log Q_5	(constant)	-1.15	1.02	0.23	0.74	2.64E-01	62
	log (*area*)	0.55	0.07			1.18E-10	
	log ($I_{6,2}$)	2.35	0.55			7.66E-05	
	log (*rain*)	-0.51	0.24			3.51E-02	
	log (S1085)	-0.36	0.14			1.48E-02	
log Q_{10}	(constant)	-0.80	1.10	0.25	0.71	0.47162	62
	log (*area*)	0.54	0.08			1.36E-09	
	log ($I_{6,2}$)	2.61	0.60			5.04E-05	
	log(*rain*)	-0.69	0.25			0.00834	
	log (S1085)	-0.38	0.16			0.0169	
log Q_{20}	(constant)	-0.57	1.18	0.27	0.69	6.34E-01	62
	log (*area*)	0.54	0.08			1.20E-08	
	log ($I_{6,2}$)	2.78	0.64			5.53E-05	
	log (*rain*)	-0.81	0.27			0.0043	
	log (S1085)	-0.40	0.17			0.0212	
log Q_{50}	(constant)	-0.37	1.29	0.30	0.65	0.77693	62
	log (*area*)	0.53	0.09			1.48E-07	
	log ($I_{6,2}$)	2.95	0.70			8.96E-05	
	log (*rain*)	-0.91	0.30			0.00359	
	log (S1085)	0.41	0.18			0.02989	
log Q_{100}	(constant)	-0.27	1.38	0.32	0.62	0.845199	62
	log (*area*)	0.53	0.10			7.69E-07	
	log ($I_{6,2}$)	3.03	0.75			0.000146	
	log (*rain*)	-0.95	0.32			0.004209	
	log (S1085)	-0.41	0.20			0.038807	

The model equations are given below:

$$\log Q_2 = -10.03 + 0.54\log(area) + 1.53\log(I_{6,2}) + 2.74\log(evap) - 0.31\log(S1085) \quad \ldots(5.19)$$

$$\log Q_5 = -1.15 + 0.55\log(area) + 2.35\log(I_{6,2}) - 0.51log(rain) - 0.36\log(S1085) \quad \ldots(5.20)$$

$$\log Q_{10} = -.80 + 0.54\log(area) + 2.61\log(I_{6,2}) - 0.69\log(rain) - 0.38\log(S1085) \quad \ldots(5.21)$$

$$\log Q_{20} = -.57 + 0.54\log(area) + 2.78\log(I_{6,2}) - 0.81\log(rain) - 0.40\log(S1085) \quad \ldots(5.22)$$

$$\log Q_{50} = -0.37 + 0.53\log(area) + 2.95\log(I_{6,2}) - 0.91\log(rain) - 0.41\log(S1085) \quad \ldots(5.23)$$

$$\log Q_{100} = -0.27 + 0.53\log(area) + 3.03\log(I_{6,2}) - 0.95\log(rain) - .41\log(S1085) \quad \ldots(5.24)$$

Adequacy Checking of Model

To assess the model adequacy, the plot of Q_{obs} and Q_{pred} , Q_{pred}/Q_{obs} ratio and median relative error values are examined for clustering group B1 (consisting of 67 catchments) (Figures 5.14, 5.15 and 5.16).

Figure 5.14 represents the plot of observed vs predicted flood quantiles for 20 years ARI. The plot overall shows a reasonable scatter between observed and predicted flood quantiles. Overall, the scatter around the 45-degree line in Figure 5.14 is seemed to be reasonable for most of the catchments. The plots for 2, 5, 10, 50 and 100 year ARIs can be seen in Appendix C (Figure C.16 to Figure C.20). The scatter in the observed vs predicted flood quantiles for these ARIs seem to be reasonable; however, the smaller ARIs represent a better match.

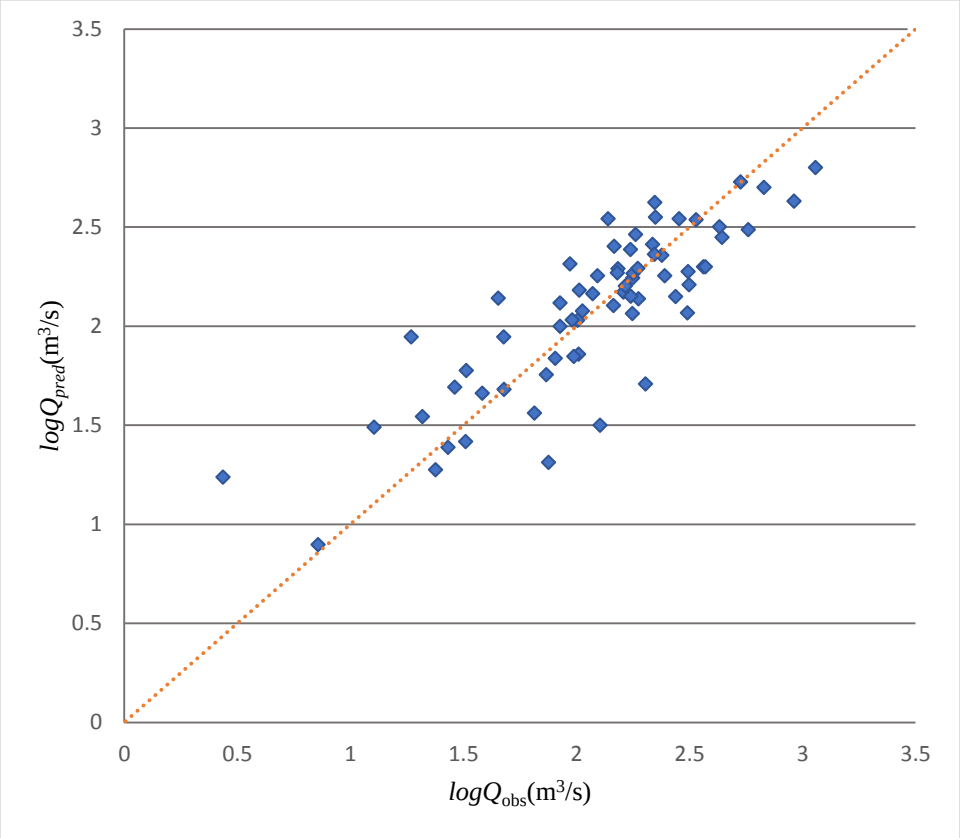

Figure 5.14 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B1 for Q_{20}

Figure 5.15 shows the boxplots of RE values for the log-log linear model for clustering group B1. The median RE values match with the 0 – 0 line very well for ARIs of 2 and 20 years, reasonably well for ARIs of 5, 10 and 50 years. For 100 years ARI, there is noticeable underestimation. In terms of the RE band, ARI of 5 years shows the lowest spread, which is very similar to that of ARIs of 2 and 10 years. The RE band for 100 years ARI is more than double to ARIs of 2 and 10 years. These results show that in terms of RE, the best overall result is achieved for 10 years ARI. According to RE band, it is revealed that the performance of log-log linear model based RFFA model is relatively poor for the higher ARIs (i.e. 50 to 100 years), which is as expected (Haddad and Rahman, 2012).

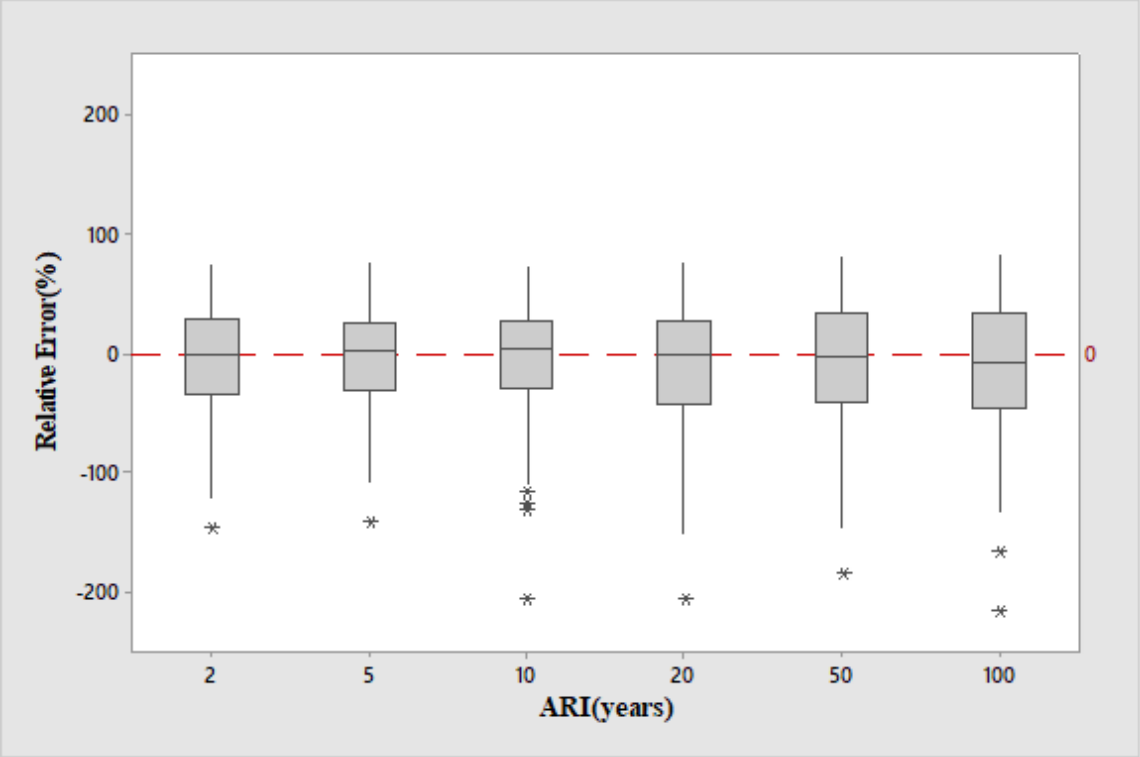

Figure 5.15 Boxplots of RE values for log-log linear model of clustering group B1

Figure 5.16 presents the boxplots of the Q_{pred}/Q_{obs} ratio values of clustering group B1 for all the six ARIs. It is found that the median Q_{pred}/Q_{obs} ratio values are located closer to 1 – 1 line, in particular for ARIs of 2, 5, and 20 years (the best agreement is for ARI of 2 and 20 years). However, for ARI of 100 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance above the 1 – 1 line and for ARI of 50 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance above the 1 – 1 line. In terms of the spread of the Q_{pred}/Q_{obs} s ratio values, ARIs of 2, 5 and 10 years exhibit very similar results. Furthermore, the spreads of the Q_{pred}/Q_{obs} ratio values for 50 and 100 years are very similar, which are remarkably larger than 2, 5 and 10 years.

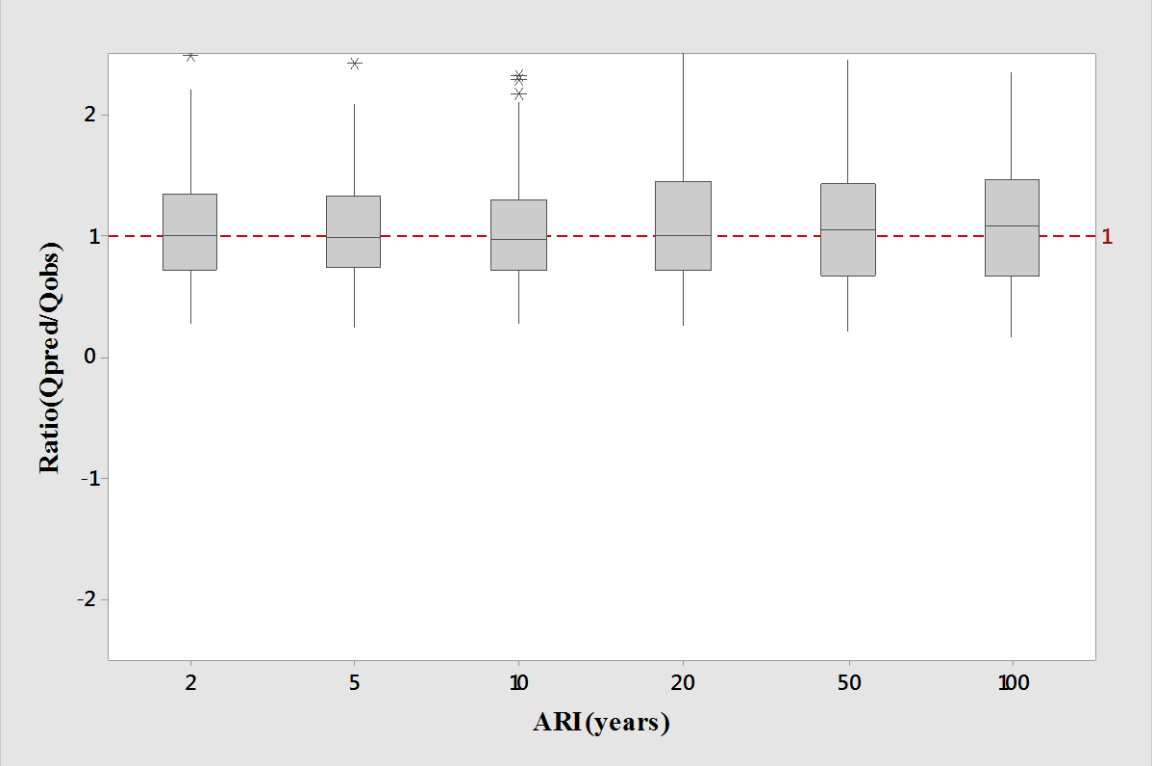

Figure 5.16 Boxplots of Q_{pred}/Q_{obs} ratio values for log-log linear model of clustering group B1

5.3.5. Evaluation of log-log linear model performance (clustering group B2)

Model development

Log-log linear prediction models are developed using 47 catchment data from group B2 for 6 different ARIs. Table 5.6 shows the model statistics for log-log linear model for B2 which includes the major determinants like R^2, *p*-statistics, *SEE etc*. R^2 values are found to range from 0.65 to 0.32 respectively for Q_2 to Q_{100} which indicates towards a large variation from higher to lower values. The R^2 value is found particularly smaller for higher ARIs which indicates towards the larger variance of prediction in estimation of higher ARI floods. This might be reasonable considering the smaller number of catchment datasets which drives toward larger uncertainties for higher ARIs.

The SEE varies from 0.20 to 0.30 respectively for Q_2 to Q_{100} following a linearly increasing trend. The lowest value of residual standard error was found for Q_2 and highest was found for Q_{100}. The increasing range of SEE value indicates towards the association of a larger percentage error with higher ARIs.

The predictor variables selected for log-log linear model for B2 are also available from Table 5.6 which are selected considering respective *p*-statistics value. It is found that *area* and *sden* are present for almost all the prediction models for different ARIs. The second most important independent variable is found to be *rain*, which is present only in prediction models of smaller ARIs like Q_2 and Q_5. Overall, the prediction equations show consistency in regards to selection of independent variables.

Table 5.6 Model statistics for log-log linear model of clustering group B2

Equation	Predictor variables	Regression Coefficient (β)	Standard Error	Standard Error of Estimate (*SEE*)	R^2	*p* value	D.F
log Q_2	(constant)	-4.18	0.83	0.20	0.65	8.89E-06	43
	log (*area*)	0.75	0.09			2.48E-10	
	log (*rain*)	1.31	0.27			1.96E-05	
	log (*sden*)	0.69	0.15			5.67E-05	
log Q_5	(constant)	-1.88	0.84	0.20	0.57	2.98E-02	43
	log (*area*)	0.70	0.09			1.63E-09	
	log (*rain*)	0.68	0.27			0.01782	
	log (*sden*)	0.63	0.16			0.000198	
log Q_{10}	(constant)	0.29	0.26	0.22	0.48	2.74E-01	44
	log (*area*)	0.67	0.10			5.47E-08	
	log (*sden*)	0.58	0.18			2.00E-03	
log Q_{20}	(constant)	0.43	0.29	0.24	0.42	0.14397	44
	log (*area*)	0.67	0.11			4.14E-07	
	log (*sden*)	0.54	0.19			0.00771	
log Q_{50}	(constant)	0.57	0.33	0.28	0.39	0.0919	44
	log (*area*)	0.67	0.13			4.15E-06	
	log (*sden*)	0.49	0.22			0.0309	
log Q_{100}	(constant)	0.64	0.36	0.30	0.32	0.0795	44
	log (*area*)	0.68	0.14			1.70E-05	
	log (*sden*)	0.45	0.24			6.66E-02	

Overall, the model equations can be written as:

$$\log Q_2 = -4.18 + 0.75\log(area) + 1.31\log(rain) + .69\log(sden) \quad \ldots(5.25)$$

$$\log Q_5 = -1.88 + 0.70\log(area) + .68\log(rain) + .63\log(sden) \quad \ldots(5.26)$$

$$\log Q_{10} = 0.29 + 0.67\log(area) + .58\log(sden) \quad \ldots(5.27)$$

$$\log Q_{20} = 0.44 + 0.67\log(area) + .54\log(sden) \quad \ldots(5.28)$$

$$\log Q_{50} = 0.57 + 0.67\log(area) + .49\log(sden) \quad \ldots(5.29)$$

$$\log Q_{100} = 0.64 + 0.68\log(area) + .45\log(sden) \quad \ldots(5.30)$$

Adequacy checking

To assess the model performance, the plots of Q_{obs} and Q_{pred}, Q_{pred}/Q_{obs} ratio and median relative error values (Figures 5.17, 5.18 and 5.19) are examined for clustering group B2 (consisting of 47 catchments).

Figure 5.17 shows Q_{obs} and Q_{pred} values for a 20 years return period. The figure shows an overall reasonable scatter between the observed and predicted flood quantiles. Overall, the scatter around the 45-degree line in Figure 1 is deemed reasonable for most of the test catchments. The Q_{obs} and Q_{pred} plots for the remaining return periods can be seen in Appendix C (Figure C.21 to Figure C.25); from these figures, it is found that the results are very similar for ARIs of 2, 5, 10 and 20 and 50 years.

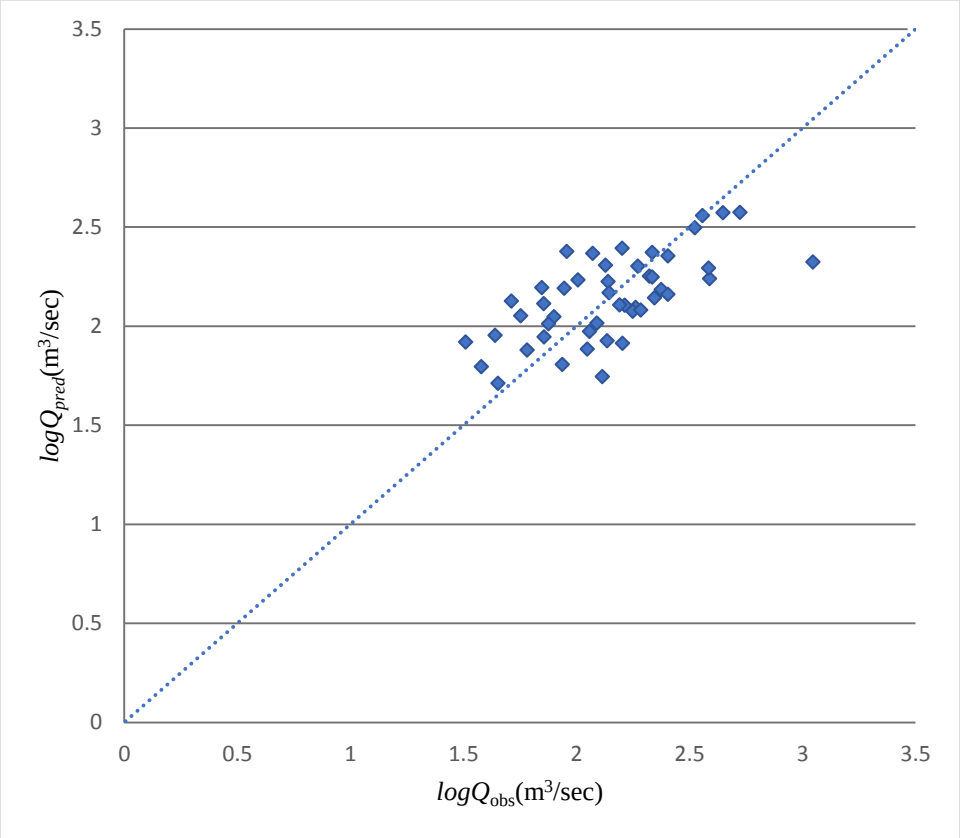

Figure 5.17 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B2 for Q_{20}

Figure 5.18 shows the boxplots of RE values for the log-log linear model for B2. The median RE values match with the 0 – 0 line very well for ARI of 5 years and reasonably well for ARIs of 2, 20 and 50 years. For ARIs of 2 and 100 years, a noticeable underestimation and overestimation are noticed, respectively. In terms of the RE band, ARI of 2 years shows the lowest spread, which is slightly lower than RE band of 5 years of ARI. The lower to higher range followed by ARIs of 2, 5, 100, 20, 50 and 100 years, respectively. The RE band for 100 years ARI is more than double to ARIs of 2 and 10 years. These results show that in terms of RE, the best result overall is achieved for 5 years ARI. According to RE band, it is revealed that the performance of log-log linear model is relatively poor for the higher ARIs (i.e. 50 to 100 years).

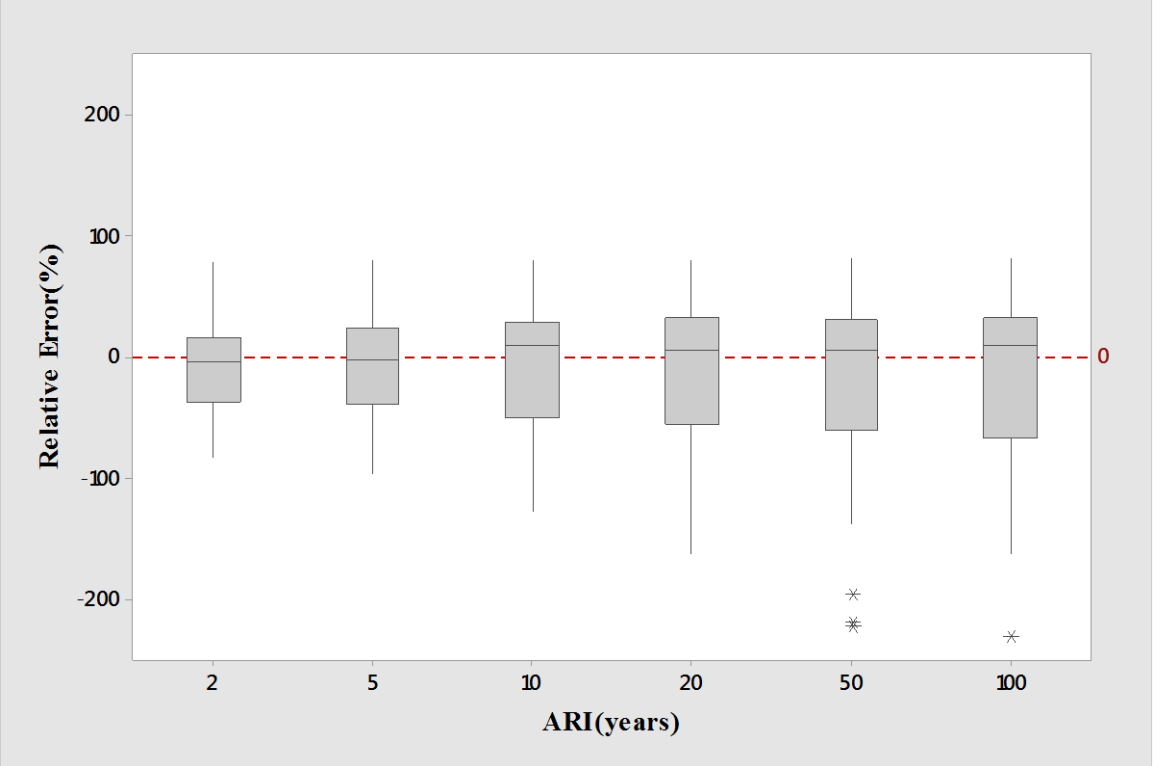

Figure 5.18 Boxplots of RE for log-log linear model of clustering group B2

Figure 5.19 presents the boxplots of the Q_{pred}/Q_{obs} ratio values for different ARIs. It is found that the median Q_{pred}/Q_{obs} ratio values are located closer to 1 – 1 line in particular for ARIs of 2 and 5 years (the best agreement is for ARI of 2 years). However, for ARI of 100 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance below the 1 – 1 line and for ARI of 2 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance below the 1 – 1 line. These results indicate a noticeable overall underestimation for 10 and years return periods. In terms of the spread of the Q_{pred}/Q_{obs} ratio values, ARI of 2 years exhibits the lowest spread followed by ARIs of 2, 5, 10, 20, 50 and 100 years. Furthermore, the spreads of the Q_{pred}/Q_{obs} ratio values for 50 and 100 years are very similar, which are remarkably larger than 2 and 5 years of ARIs.

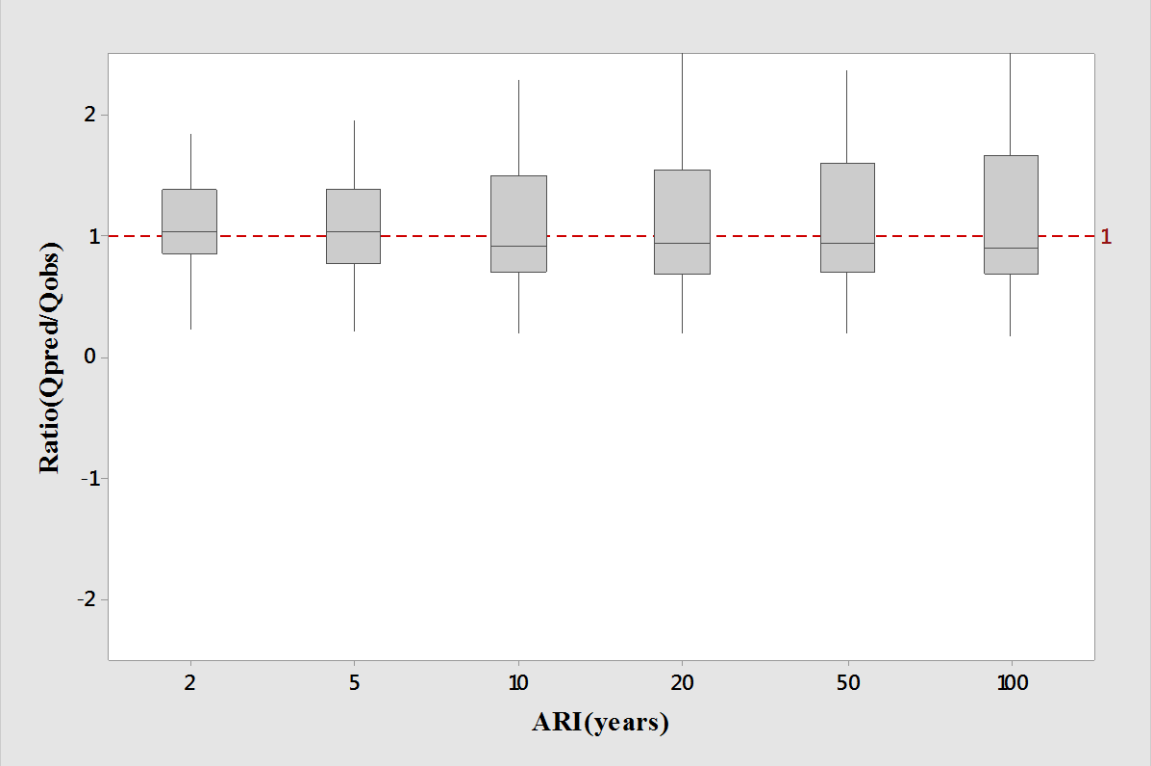

Figure 5.19 Boxplots of Q_{pred}/Q_{obs} ratio log-log linear model values of clustering group B2

5.4. Comparison of median RE and median Q_{pred}/Q_{obs} ratio values for the log-log linear model

5.4.1. Median RE

Table 5.7 shows the median RE values for all the log-log linear models developed in this chapter. In terms of median RE, groups A1, B1 and B2 show pretty consistent and reasonable results with similar range (approximately between 30~40%). The lowest value of median relative error is 18.75% which is for the combined group, and the highest median relative error is found for group A2 which is about 60%. Median RE values are considerably higher for Q_{50} and Q_{100} in all the clustering groups, which can be seen in Table 5.7 and Figure 5.20. Overall, clustering group A1 shows the best result among all the clustering groups. However, if both groups A1 and A2 are compared (generated by Wards-Manhattan cluster analysis method) against groups B1 and B2 (generated by K-means cluster analysis method), groups B1 and B2 perform better than groups A1 and A2. This shows that K-means that the cluster analysis method has generated better groups than the Wards-Manhattan cluster analysis method.

Table 5.7 Median RE values for combined data set and clustering groups

Flood quantile	Combined	Group (A1)	Group (A2)	Group (B1)	Group (B2)
Q_2	18.73	29.56	23.10	30.33	25.82
Q_5	32.88	28.60	34.69	28.20	31.97
Q_{10}	19.36	27.47	40.54	27.37	33.05
Q_{20}	34.51	30.74	43.02	29.37	36.69
Q_{50}	40.41	33.25	53.10	37.42	39.29
Q_{100}	40.99	37.05	59.94	37.00	42.63
Overall	31.15	31.11	42.40	31.61	34.91

Figure 5.20 illustrates the comparative performance for individual log-log linear models with respect to median RE. Overall, higher range of median RE values can be seen for group A2. The graphical representation also depicts that B1 and B2 produce relatively smaller median RE compared with group A2. Overall, Group A1 shows the smallest median RE values for most of the ARIs; however, for ARIs of 2 and 10 years, combined data set (i.e. all the 114 catchments forming one group) shows the smallest median RE values.

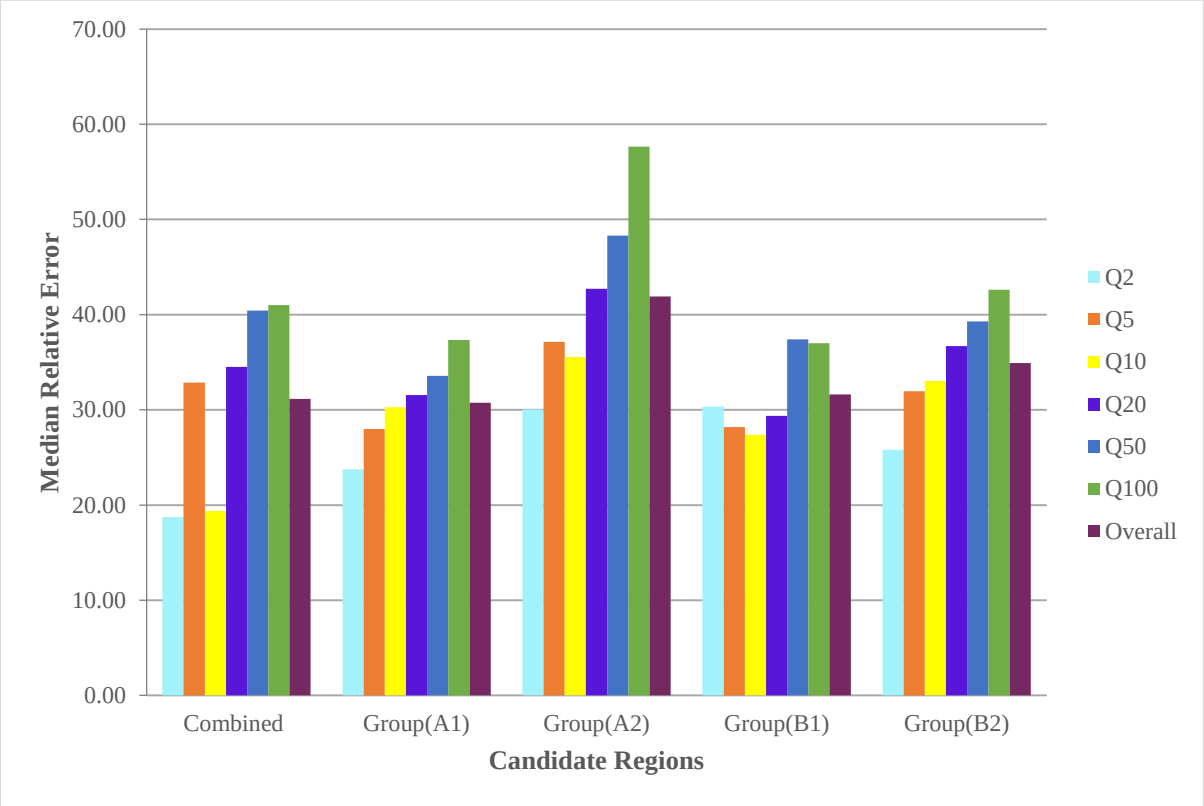

Figure 5.20 Median Relative Error values of log-log linear model based RFFA models based on combined data set and groupings based on cluster analysis

5.4.2. Median Q_{pred}/Q_{obs} ratio

Table 5.8 summarises the median Q_{pred}/Q_{obs} ratio values for the five different log-log linear models. For the combined dataset, the median Q_{pred}/Q_{obs} ratio values range from 0.94 to 1.03. Q_{pred}/Q_{obs} value for 100 years ARI for the combined dataset is found to be lowest (which is 0.94), exhibiting a notable underestimation for this ARI. The best result is obtained for Q_5 and Q_{20}, which is 1.00. In summary, the log-log linear model for the combined dataset shows a very good median Q_{pred}/Q_{obs} ratio value of 0.99, which puts it at rank 2 among all the four models (see Table 5.9), and consistent values of median Q_{pred}/Q_{obs} ratio values are also found for ARIs of 5 and 20 years.

Table 5.8 Median Q_{pred}/Q_{obs} ratio values for log-log linear model based on combined data set and groupings based on cluster analysis

Flood quantile	Combined group	Group (A1)	Group (A2)	Group (B1)	Group (B2)
Q_2	1.03	1.04	1.00	1.01	1.04
Q_5	1.00	0.95	0.99	0.98	1.03
Q_{10}	0.97	0.94	0.98	0.96	0.92
Q_{20}	1.00	0.97	1.01	1.01	0.94
Q_{50}	0.98	1.02	0.95	1.05	0.94
Q_{100}	0.94	1.02	0.95	1.09	0.90
Overall	0.99	0.99	0.98	1.01	0.96

In case of the clustering group A1, the median Q_{pred}/Q_{obs} ratio values range from 0.94 (Q_{10}) to 1.04 (Q_2); all the median Q_{pred}/Q_{obs} ratio values seem to be within an acceptable range. The overall median Q_{pred}/Q_{obs} ratio value for A1 shows a very good performance with the value of 0.99, which places it at rank 1 among the 5 group of the log-log linear models, and also with consistent values between the 6 ARIs.

In case of group A2, the flood quantiles seem to be underestimated with 0.95 for Q_{100} and Q_{50}. The rest of the flood quantiles are showing mostly underestimation of 1% to 5%. The overall median Q_{pred}/Q_{obs} ratio value for A2 is found to be reasonable (i.e. 0.98), thus placing it in rank 4 among the five clustering groups of the log-log linear model.

For the clustering group B1, the median Q_{pred}/Q_{obs} ratio varies from 0.96 to 1.09, which shows a large variation among ARIs. Most of the predictions are overestimated except for slight underestimation in the case of Q_5 and Q_{10} which show median Q_{pred}/Q_{obs} ratio values of 0.98 and 0.96, respectively. Overall, it shows a reasonable performance with a median Q_{pred}/Q_{obs} ratio value of 1.01 for clustering group B1, which ranks it at position 3 among the 5 clustering groups.

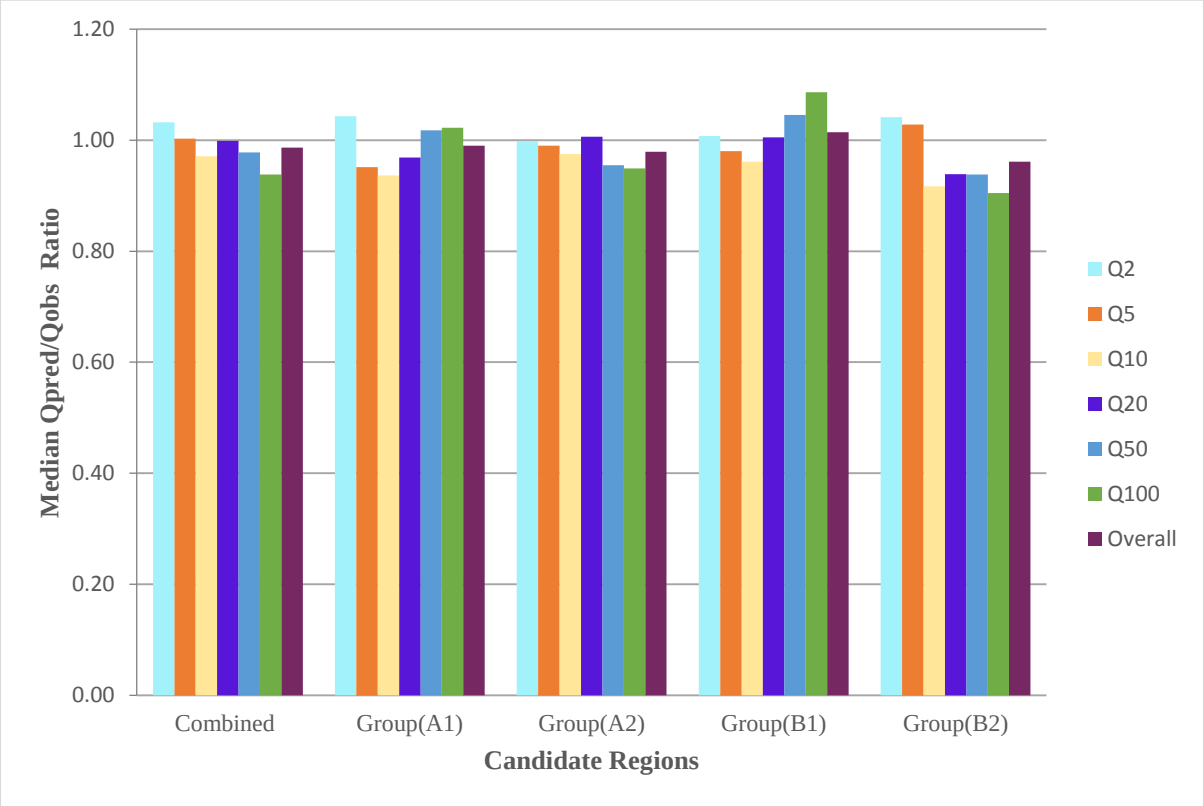

Figure 5.21 Median Q_{pred}/Q_{obs} values for log-log linear models based on combined data set and groupings based on cluster analysis

For B2 clustering group, slight underestimations are found for ARIs of 10 to 100 years (Table 5.8) and slight overestimations are noticed for ARIs of 2 and 5 years. This model shows a moderate range in terms of median Q_{pred}/Q_{obs} ratio value. Therefore, with the overall median Q_{pred}/Q_{obs} ratio value being 0.96, clustering group B2 ranks 5 among the 5 clustering groups.

5.4.3. Ranking of log-log linear models

Table 5.9 summarises the subjective rankings of the log-log linear models based on four clustering groups and combined data set with respect to median RE and Q_{pred}/Q_{obs} ratio values. From Table 5.9 it can be seen that the best performing log-log linear model is achieved for the clustering group A1 (consisting of 79 catchments), having rank 1 with respect to both the median RE and median Q_{pred}/Q_{obs} ratio value. For the combined dataset (when all the 114 catchments are placed in a single group), the log-log linear model receives rank 2 (with respect to median Q_{pred}/Q_{obs} ratio value) and rank 3 (with respect to median RE), and hence it shows a better log-log linear model than clustering groups A2, B1 and B2.

Hence, it can be concluded that the for log-log linear model found from the clustering group A1 is the best model.

Table 5.9 Ranking of log-log linear models

Criteria	Rank1	Rank 2	Rank 3	Rank 4	Rank5
Median RE	A1	A2	Combined	B1	B2
Median Q_{pred}/Q_{obs} ratio	A1	Combined	B1	A2	B2

5.5. Summary

In this chapter, the log-log linear model is developed based on the full dataset consisting of 114 catchments and 4 regions formed by cluster analysis (i.e. consisting of 79, 35, 67 and 47 catchments, respectively). The models are assessed based on three criteria: scatter plot of Q_{obs} vs Q_{pred} , median RE (%) and median Q_{pred}/Q_{obs} ratio values. Each of the developed log-log linear models based on these criteria is ranked in this chapter, and it is found that clustering group A1 (derived by Ward Manhattan cluster analysis) results in the best-performing model. This model needs to be compared with the GAM model in the next chapter.

CHAPTER 6

DEVELOPMENT OF GAM BASED RFFA TECHNIQUES

6.1. General

The chapter focuses on the development of a new technique of design flood estimation for ungauged catchments using generalised additive model (GAM). It describes the method of developing prediction equations (for 6 average recurrence intervals (ARIs), which are 2, 5, 10, 20, 50 and 100 years) by utilising GAM for 5 different groups of data (full dataset consisting of 114 catchments and 4 regions formed by cluster analysis as mentioned in Chapter 5). The developed prediction models are then tested to assess their relative accuracy in making predictions. Adequacy of the developed prediction models are assessed using three criteria: median Q_{pred}/Q_{obs} ratio, plot of Q_{obs} and Q_{pred} and median relative error (RE). Furthermore, this chapter compares the overall performance of the GAM models. Finally, the GAM models are compared with the log-log linear models for each of the ARIs.

6.2. GAM model development

The detail results for Q_2 GAM model are provided below. Additional results on the GAM models are provided in Appendix D.

For Q_2 model, four catchment characteristics are found to be statistically significant from GAM, which are *area*, $I_{6,2}$, *evap*, and *sden*. The important properties of model residuals are shown in Figures 6.1, 6.2 and 6.3.

Figure 6.1 represents the standardised residual vs fitted predicted values for the Q_2 model. From this plot, it can be observed that there are medium to large deviations of the residuals from 0-0 line, which indicates heteroscedasticity of prediction and residuals. The overall results show medium heterogeneity of variances for lower discharges and large scatter for higher ones. Overall, it indicates minor heteroscedasticity between the model predicted values and residuals.

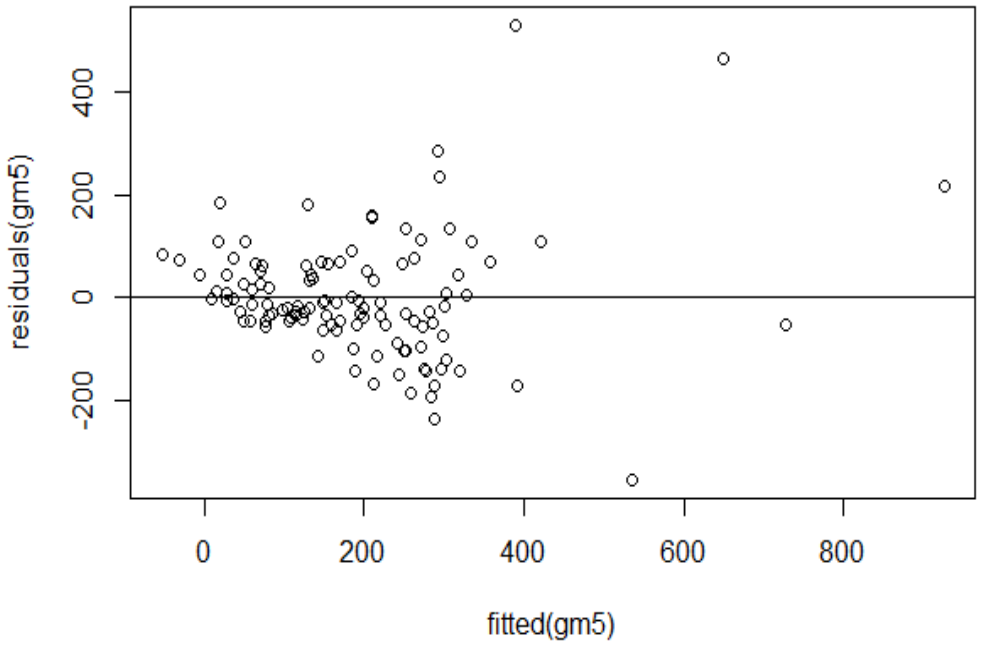

Figure 6.1 Fitted predicted value vs standardised residuals plot for GAM model of combined group

Figure 6.2 represents the normal Q-Q plot of the standardised residuals for the Q_2 GAM model. The plot shows a good agreement between the predicted values and the standardised residuals, which indicates that the residuals for Q_2 model generally follow a normal distribution except in the tails of the distribution.

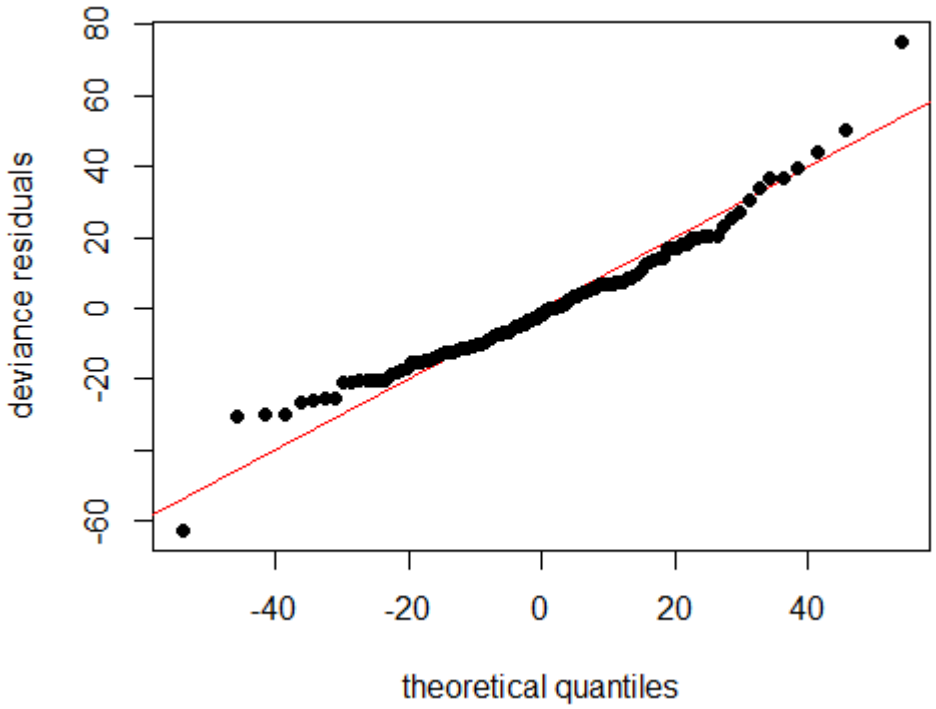

Figure 6.2 Normal Q-Q plot of the standardised residuals for GAM model for combined group for Q_2

Figure 5.3 represents the histogram of the standardised residuals, which indicates that the residuals are near normally distributed with a mean of zero, but there are a few outliers with values larger than -50 and +50.

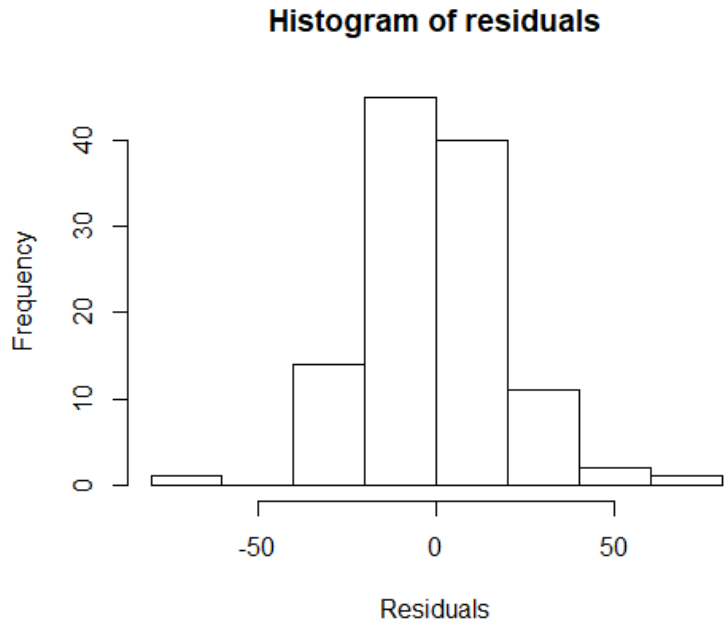

Figure 6.3 Histogram of the standardised residuals for GAM model for combined group for Q_2

Model statistics

Table 6.1 represents the overall GAM model (combined group) statistics for the 6 different ARIs. The major determinants are coefficient of determination (R^2), *p*-statistics and GCV score. From Table 6.1, it is found that the R^2 values range from 0.69 to 0.44; particularly, smaller R^2 values are found for the higher ARIs indicating a weaker model. The R^2 values for lower ARIs seem to be quite reasonable (0.62-0.69).

The GCV values vary from 501 to 82,994 for Q_2 to Q_{100}. The lowest value of GCV is found for Q_2 and the highest one is found for Q_{100}. This indicates that the cross validation error increases with increasing ARIs.

The predictor variables for the individual models are selected based on the *p*-statistics of the predictor variables. The criterion of including a predictor variable in the final model is $p \leq 0.10$. Table 6.1 contains all the selected predictor values for the models along with the respective *p*-statistics. The predictor variables *area*, $I_{6,2}$ and *evap* appear to be the most important variables for estimating flood quantiles using GAM, as these three variables are common in all the prediction equations. The next most important predictor variable is *rain*, which appears in all the prediction models except for Q_2. Another predictor variable, which is found statistically significant in Q_2, Q_5 and Q_{10} is *sden*. Overall, Q_{20}, Q_{50} and Q_{100} models show a consistency in the selection of predictor variables (with *area*, $I_{6,2}$ and *evap*). The general forms of the developed prediction equations using GAM are shown below:

$$\ln(Q_2) = \alpha + \mathrm{s}(area) + \mathrm{s}(I_{6,2}) + \mathrm{s}(evap) + \mathrm{s}(sden) \quad \ldots(6.1)$$

$$\ln(Q_5) = \alpha + \mathrm{s}(area) + \mathrm{s}(I_{6,2}) + \mathrm{s}(rain) + \mathrm{s}(evap) + \mathrm{s}(sden) \quad \ldots(6.2)$$

$$\ln(Q_{10}) = \alpha + \mathrm{s}(area) + \mathrm{s}(I_{6,2}) + \mathrm{s}(rain) + \mathrm{s}(evap) + \mathrm{s}(sden) \quad \ldots(6.3)$$

$$\ln(Q_{20}) = \alpha + \mathrm{s}(area) + \mathrm{s}(I_{6,2}) + \mathrm{s}(rain) + \mathrm{s}(evap) \quad \ldots(6.4)$$

$$\ln(Q_{50}) = \alpha + \mathrm{s}(area) + \mathrm{s}(I_{6,2}) + \mathrm{s}(rain) + \mathrm{s}(evap) \quad \ldots(6.5)$$

$$\ln(Q_{100}) = \alpha + \mathrm{s}(area) + \mathrm{s}(I_{6,2}) + \mathrm{s}(rain) + \mathrm{s}(evap) \quad \ldots(6.6)$$

Table 6.1 Important model statistics for GAM models of combined group

Flood quantile	Predictor variables	Deviance explained (%)	GCV	R^2	*F* value	*p* value
Q_2	*area*	73.70	501.61	0.69	30.199	4.13E-15
	$I_{6,2}$				5.37	7.39E-06
	evap				7.59	1.57E-06
	sden				6.07	0.00209
Q_5	*area*	71.3	3201.90	0.66	26.69	3.95E-13
	$I_{6,2}$				4.898	3.43E-05
	rain				3.073	0.0828
	evap				6.278	8.56E-06
	sden				4.492	0.0126
Q_{10}	*area*	67.60	8437.80	0.62	23.46	8.42E-12
	$I_{6,2}$				4.67	8.47E-12
	rain				6.91	0.009928
	evap				5.02	0.000111
	sden				3.15	0.04189
Q_{20}	*area*	62.20	18974.00	0.56	17.39	9.02E-10
	I6,2				4.41	0.000213
	rain				8.95	0.003489
	evap				3.99	0.00109
Q_{50}	*area*	56.20	45823.00	0.50	9.96	1.66E-09
	$I_{6,2}$				8.56	0.000309
	rain				12.12	0.000735
	evap				3.31	0.00326
Q_{100}	*area*	48.40	82994.00	0.44	17.32	1.53E-09
	$I_{6,2}$				11.53	0.000403
	rain				10.87	0.001332
	evap				2.46	0.028319

Model adequacy checking .

The GAM based prediction models (for the combined and clustering groups) are tested using a 10-fold cross validation (as noted in Chapter 4) as per the following criteria:

- Q_{pred}/ Q_{obs} ratio
- Plot of Q_{obs} and Q_{pred}
- Median relative error (RE)

GAM based models are ranked based on their relative performances in relation to these criteria. Figures 6.4, 6.5 and 6.6 represent the relationship between the observed and predicted flood quantiles. The observed flood quantiles at a given station are estimated by fitting a LP3 distribution to the annual maximum flood data. The predicted flood quantiles are obtained by the developed GAM models.

The scatter plot of the predicted and observed flood quantiles for the combined group (for 20 years of ARI) is shown in Figure 6.4. The plot generally presents a good agreement between the predicted and observed flood quantiles except for only a few stations. However, there are some overestimations and underestimations by the GAM model at lower discharges. Most of the catchments are within a narrow range of variability from the 45-degree line except for a few outliers, in particular for lower discharges. Ignoring those outliers, for most of the catchments, the scatter around the 45-degree line in Figure 6.4 is deemed to be reasonable. Overall, the GAM model shows better results for medium to higher discharges.

The Q_{obs} vs Q_{pred} scatter plots for the remaining ARIs (i.e. for 2, 5, 10, 50 and 100 years) can be seen in Appendix E (Figure E.1 to E.5). Overall, the results show a better prediction for ARIs of 2, 5, 10 and 20 years. Results for ARIs of 50 and 100 years (Figures B.35 and B.36, respectively) are quite similar, with little variations for higher discharges.

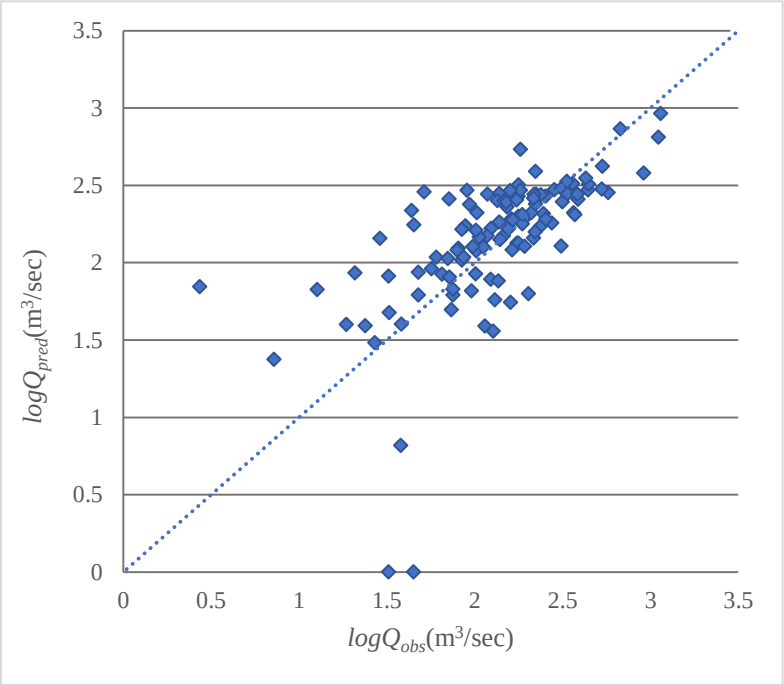

Figure 6.4 Comparison of observed and predicted flood quantiles for GAM model of combined group for Q_{20}

Figure 6.5 shows the boxplots of RE values for the GAM model of the combined group for the 6 ARIs. The median RE values match with the 0 – 0 line very well for ARI of 5 years, and reasonably well for the ARIs of 2 and 10 years. Except for 10 years ARI, there is a small to moderate underestimation by the GAM models. In terms of the RE band, ARI of 2 years and 10 years show almost similar spread, which are also smaller than the remaining ARIs. The lower to higher range of the RE spread for the remaining ARIs are in the order of 20, 5, 50 and 100 years, respectively. The RE bands for 50 and 100 years of ARIs are very similar, which indicates a similar level of prediction error for these ARIs by the GAM. These results show that in terms of RE, the best overall result (for the combined group) for the GAM model is achieved for 2 years ARI. Overall, the performances of the GAM models (as indicated by the RE bands) for the combined group do not show a large variation across the six ARIs.

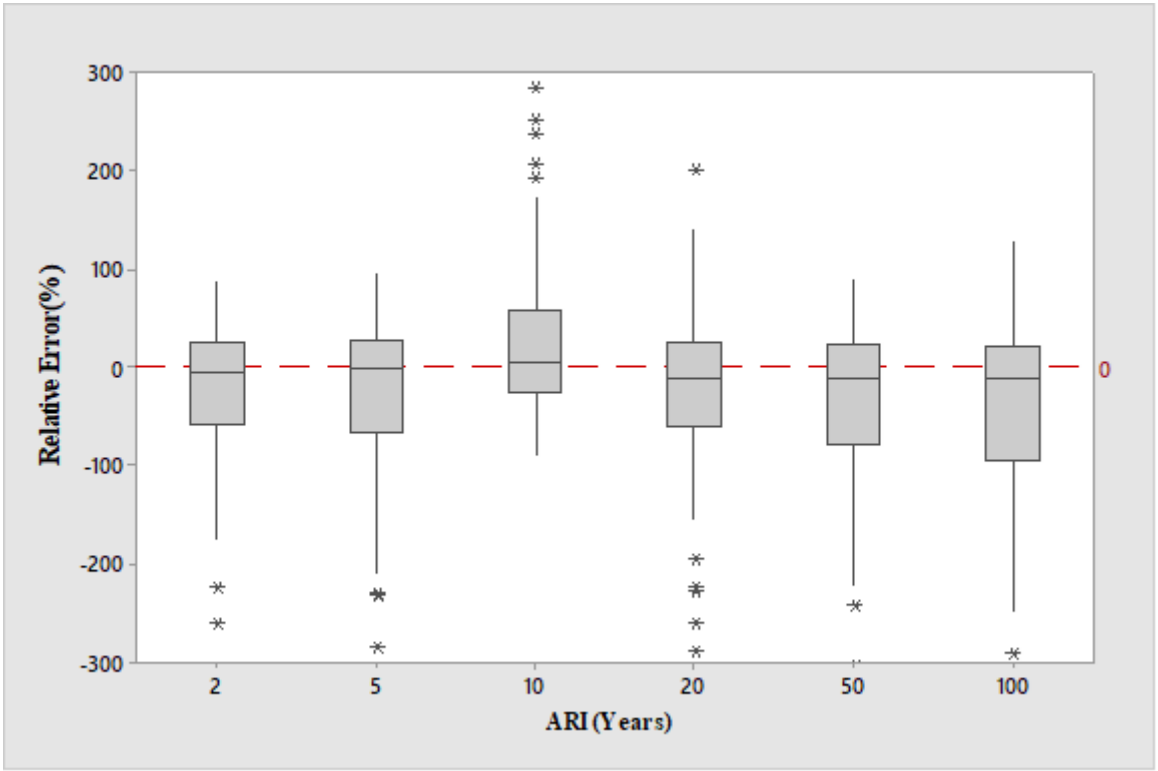

Figure 6.5 Boxplots of RE values for the GAM model of combined group

Figure 6.6 presents the boxplots of the Q_{pred}/Q_{obs} ratio values associated with the GAM models for the combined group for the six ARIs. It is found that the median Q_{pred}/Q_{obs} ratio values are located very close to 1 – 1 line, in particular for ARIs of 5 and 10 years, showing the best agreement for ARI of 10 years. However, for all the ARIs, the median Q_{pred}/Q_{obs}

ratio values are located within a short distance above the 1 – 1 line except for ARI of 100 years. For this ARI, there is a noticeable overestimation by the GAM model. These results indicate a slight to noticeable overestimation of the predicted flood quantiles for all the ARIs. In terms of the spread of the Q_{pred}/Q_{obs} ratio values, ARI of 2 years exhibits the lowest spread, whereas 10 and 20 years of ARI show similar spread. The spreads of the Q_{pred}/Q_{obs} ratio values are in the order of 5, 50 and 100 years ARIs. Furthermore, the spreads of the Q_{pred}/Q_{obs} ratio values for 50 and 100 years of ARIs are very similar, which are remarkably larger than 2, 5 and 10 years of ARIs.

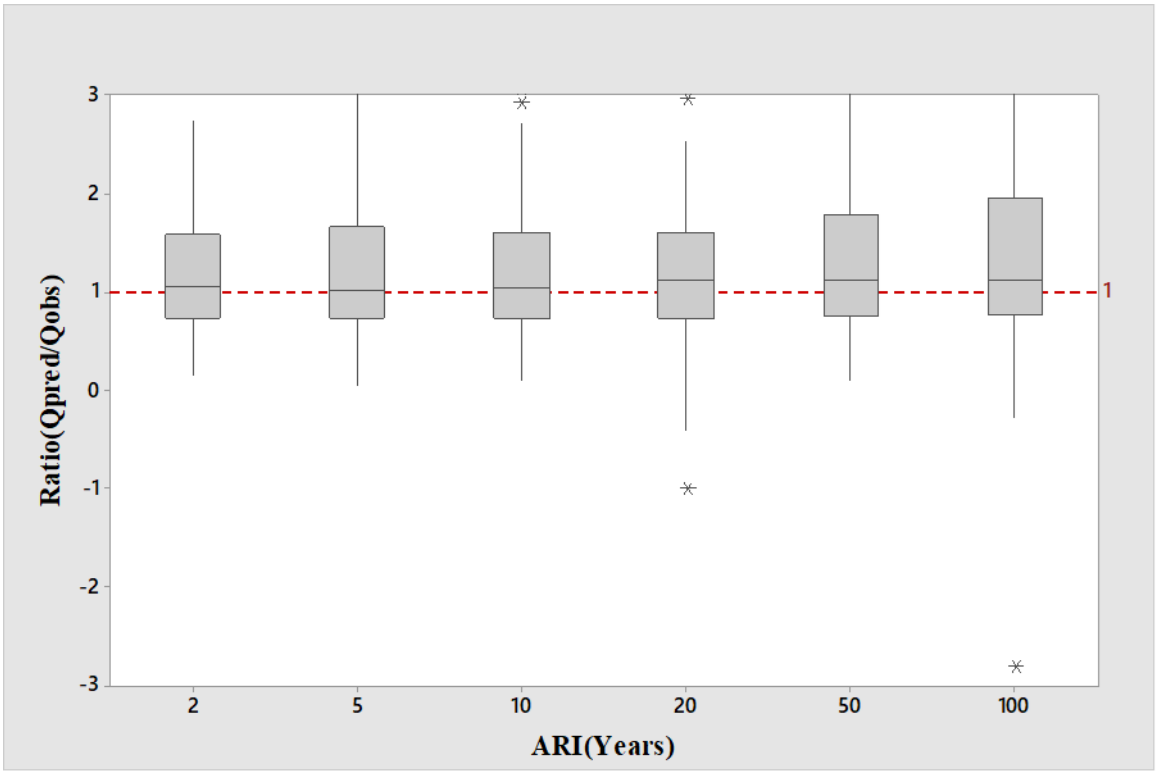

Figure 6.6 Boxplots of Q_{pred}/Q_{obs} ratio values for GAM model of combined group

6.3. GAM model performance for different clustering groups

6.3.1. Evaluation of GAM model performance (clustering group A1)

The model statistics for the developed GAM model for clustering group A1 is presented in Table 6.2. The R^2 values are ranged from 0.83 to 0.51, with a gradual decrease from Q_2 to Q_{100}. Smaller R^2 values are found for the higher ARIs indicating a higher variance of prediction for these ARIs. In particular, for 100 years ARI, the R^2 value is too low, i.e. only 0.512. This indicates that the GAM models are more accurate in predicting smaller ARI

floods, e.g. up to 20 years ARI. The GCV values of the GAM models vary from 271.84 to 75,772 for Q_2 to Q_{100} indicating associated higher cross validation errors for the higher ARIs.

Table 6.2 contains all the selected predictor variables for the individual models along with respective *p*-statistics. The most important predictor variable for this GAM model is found to be *area*, which is present in all the prediction equations for clustering group A1. The next most statistically significant independent variables are $I_{6,2}$, *evap* and *rain*. $I_{6,2}$ and *rain* are common for all the prediction equations except for Q_2. Overall, the prediction equations show consistency in selection of predictor variables except for Q_2. The developed prediction equations in the GAM are:

$$\ln(Q_2) = \alpha + \mathrm{s}(area) + \mathrm{s}(\mathrm{evap}) + \mathrm{s}(\mathrm{sden}) \quad \ldots(6.7)$$

$$\ln(Q_5) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(rain) + \mathrm{s}(evap) \quad \ldots(6.8)$$

$$\ln(Q_{10}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(rain) + \mathrm{s}(evap) \quad \ldots(6.9)$$

$$\ln(Q_{20}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(rain) + \mathrm{s}(evap) \quad \ldots(6.10)$$

$$\ln(Q_{50}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(rain) + \mathrm{s}(evap) \quad \ldots(6.11)$$

$$\ln(Q_{100}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(rain) \quad \ldots(6.12)$$

Table 6.2 Model statistics for GAM model of clustering group A1

Equation	Predictor variables	Deviance explained (%)	GCV	R^2	F value	*p* value
Q_2	*area*	86.5	271.84	0.83	25.29	<2e-16
	evap				15.65	<2e-16
	sden				2.53	0.0502
Q_5	*area*	82.5	1877.6	0.789	30.00	3.54E-15
	$I_{6,2}$				3.17	0.0419
	rain				5.16	0.0264
	evap				9.15	2.60E-09
Q_{10}	*area*	77.3	5746.4	0.731	25.93	1.49E-12
	I6,2				3.63	0.03097
	rain				9.10	0.00361
	evap				6.21	3.08E-06
Q_{20}	*area*	71.6	14447	0.666	21.40	1.84E-10
	$I_{6,2}$				4.45	0.031917
	rain				11.02	0.001452
	evap				4.24	0.000357
Q_{50}	*area*	63.6	40058	0.577	16.95	1.92E-08
	$I_{6,2}$				5.07	0.02676
	rain				11.78	0.00101
	evap				2.55	0.01834
Q_{100}	*area*	56.3	75,772	0.512	16.30	2.44E-07
	$I_{6,2}$				4.02	0.001298
	rain				16.03	0.000149

The scatter plot of Q_{obs} vs $Q_{pred,,}$ box plots of Q_{pred}/Q_{obs} ratio and RE values are presented below in order to check the adequacy of GAM model of clustering group A1. Figures 6.7, 6.8 and 6.9 illustrate the overall performance of this model.

The scatter plot of the predicted and the observed flood quantiles for the GAM model for clustering group A1 for 20 years ARI is shown in Figure 6.7. The plot illustrates a reasonable agreement between the predicted and observed flood quantiles. The plotted points scatter within a narrow range of variability around the 45-degree line for medium to large discharges. However, the plot shows noticeable scatter for lower discharges. There are also some outliers which are particularly found for lower discharges showing both overestimations and underestimations. Ignoring the outliers, for most of the catchments, the scatter around the 45-degree line in Figure 6.7 is deemed to be reasonable. However, there

are some outliers in this plot showing negative predictions by the GAM model for some of the discrete observed flood quantile values, which is due to computational uncertainties in the GAM model.

The rest of the plots can be seen in Appendix E (Figure E.6 to E.10). The results are very similar for ARIs of 2 and 5 years, which show a good scatter around the 45-degree line. The results show a noticeable range of variability around the 45-degree line, in particular for lower discharges. Overall, the GAM model for clustering group A1 shows better results for a medium range of ARIs.

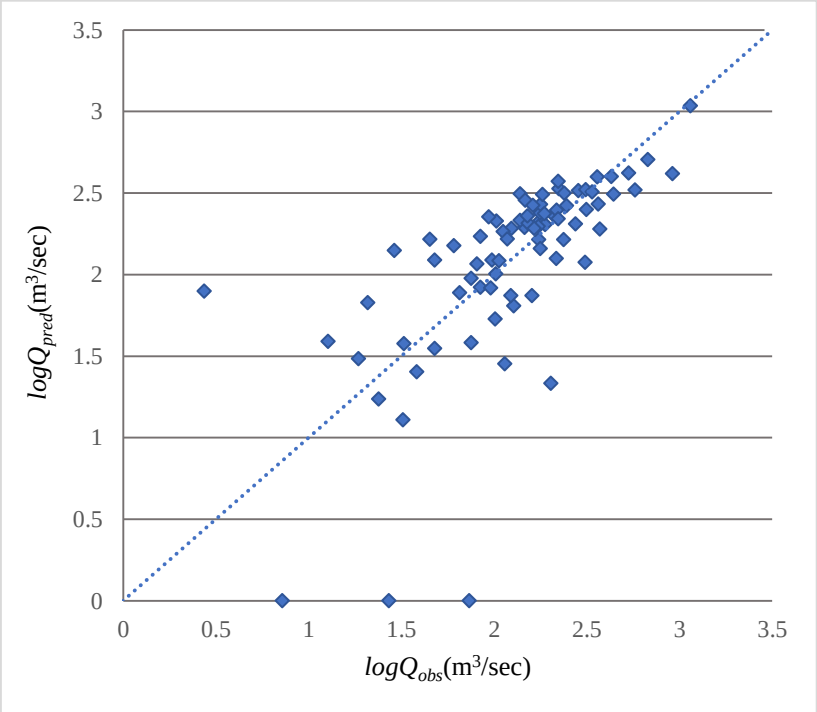

Figure 6.7 Comparison of observed and predicted flood quantiles for GAM for clustering group A1 for Q_{20}

Figure 6.8 shows the boxplots of RE values for the GAM model for clustering group A1. The median RE values match with the 0 – 0 line very well for ARIs of 2 and 5 years, and reasonably well for ARIs of 10, 20, 50 and 100 years. For ARIs of 10 to 100 years, a degree of underestimation is observed by the GAM model. In terms of the RE band, ARI of 2 years shows the lowest spread, which is slightly lower than the RE band for 5 years of ARI. The lower to higher range of RE spreads are for ARIs of 10, 20, 50 and 100 years, respectively. The RE band for 100 years ARI is about twice than those of ARIs of 2 and 5 years. These

results show that in terms of RE, overall the best result is achieved for 2 years ARI GAM model. As per RE band, it is found that the performance of GAM is comparatively better for lower ARIs. The uncertainties are relatively smaller for clustering group A1 among all the groups. The higher ARIs like 50 and 100 years show comparatively larger spread of RE, i.e. a higher uncertainty in flood estimates, which is quite common in RFFA (e.g., Haddad and Rahman, 2012; Rahman et al., 2011; Rahman et al., 2016).

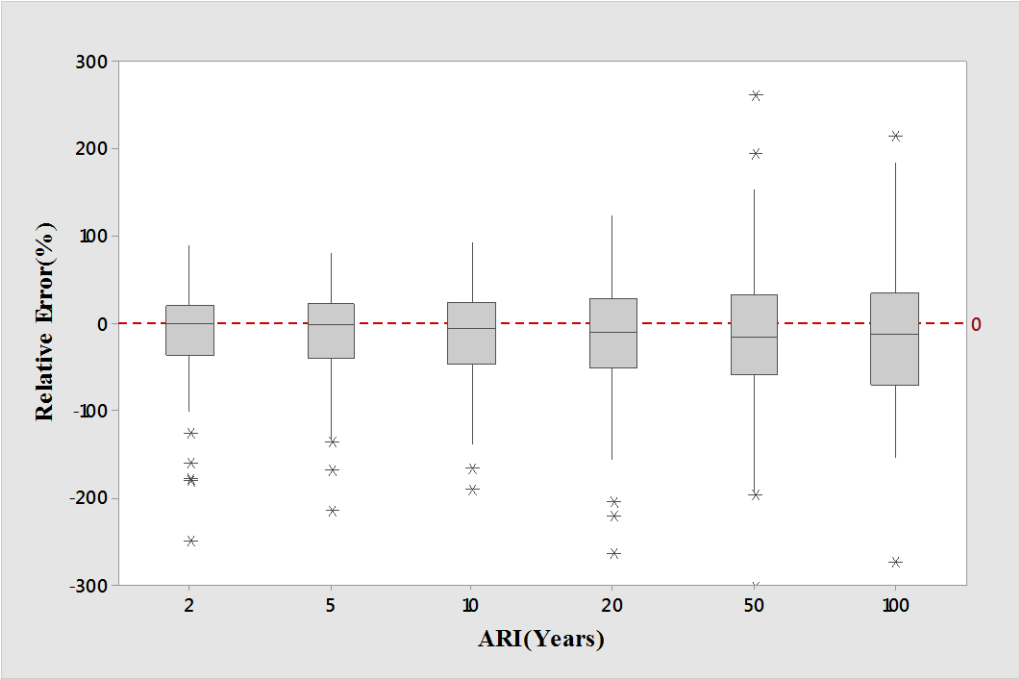

Figure 6.8 Boxplots of RE values for GAM for clustering group A1

Figure 6.9 presents the boxplots of the Q_{pred}/Q_{obs} ratio values of the GAM model for the clustering group A1 for the 6 different ARIs. It is found that the median Q_{pred}/Q_{obs} ratio values are located closer to 1 – 1 line, in particular for ARIs of 2 and 5 years, with the best agreement being for ARI of 2 years. However, for ARIs of 10, 20, 50 and 100 years, the median Q_{pred}/Q_{obs} ratio value is located within a short distance above the 1 – 1 line, indicating an overall overestimation. None of the values of the median Q_{pred}/Q_{obs} ratio are located below the 1 – 1 line, which indicates no overall underestimation for clustering group A1 in the GAM. These results indicate a minimum to reasonable overestimation for predicted flood quantiles by this GAM model for 10 to 100 years of ARIs. In terms of the spread of the Q_{pred}/Q_{obs} ratio values, ARI of 2 years exhibits the lowest spread followed by ARIs of 5, 10,

20, 50 and 100 years. Furthermore, the spreads of the Q_{pred}/Q_{obs} ratio values for 50 and 100 years are very similar, which are again remarkably larger than 2, 5 and 10 years.

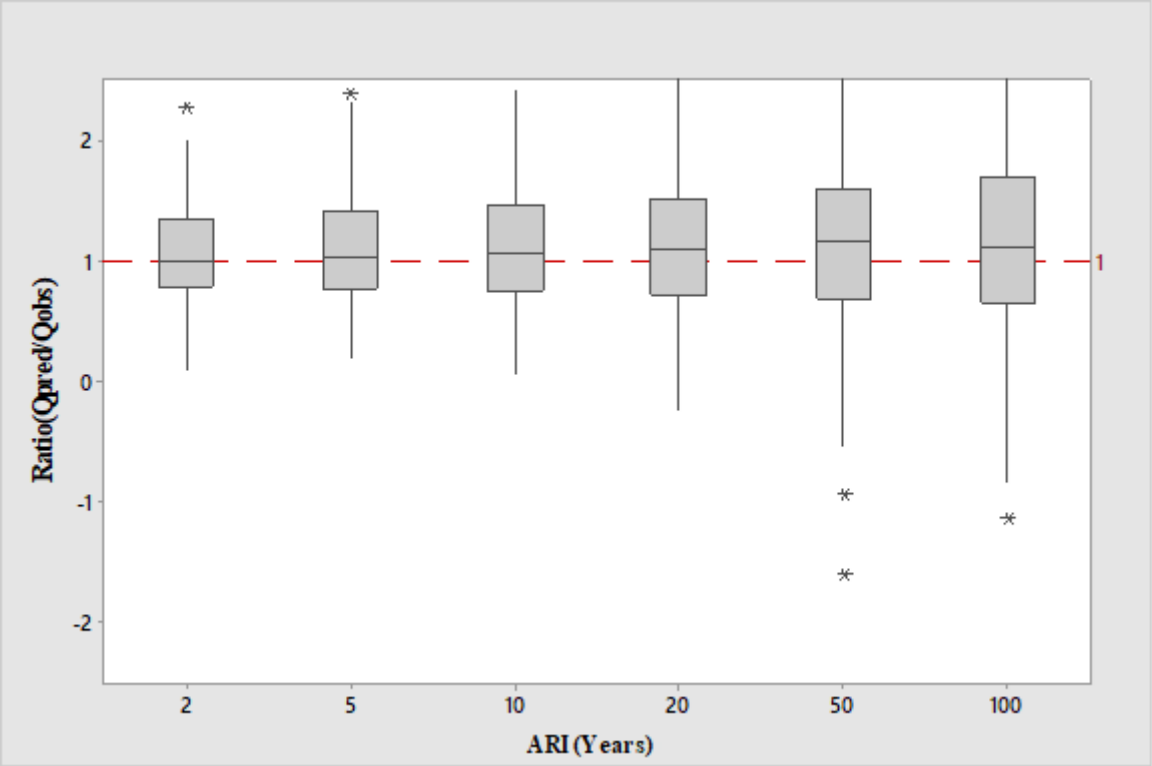

Figure 6.9 Boxplots Q_{pred}/Q_{obs} ratio value for GAM for clustering group A1

6.3.2. Evaluation of GAM model performance (clustering group A2)

The model statistics for the developed GAM model for clustering group A2 are presented in Table 6.3. The R^2 values decrease gradually with increasing ARIs, ranging 0.75 to 0.36 for Q_2 to Q_{100}. Smaller R^2 values are associated with higher ARIs indicating towards the associated larger variance of prediction. Overall, the R^2 values present a reasonable performance for lower ARIs (e.g. 2 and 5 years), but a poorer performance for higher ARIs.

The GCV values vary from 557.84 to 100,450 for Q_2 to Q_{100} indicating higher cross validation error for the higher ARI GAM models.

The final predictor variables in the model are selected based on the *p*-statistics. The criterion of selecting a predictor variable in the final model is $p \leq 0.10$. Table 6.3 contains all the selected predictor variables for the individual models along with the *p*-statistics. The most

important variables are $I_{6,2}$ and *evap*, which are common for all the ARIs. The second most statistically significant independent variable is *area* which is found in all the ARI models except for Q_{100}. For Q_2 and Q_5, *sden* is found statistically significant. Overall, Q_{10}, Q_{20} and Q_{50} models show a consistency in the selection of predictor variables (which are *area*, $I_{6,2}$ and *evap*).

The developed prediction equations for the GAM models in case of clustering group A2 are:

$$\ln(Q_2) = \alpha + \mathrm{s}(area) + s\left(I_{6,2}\right) + \mathrm{s}(\mathrm{evap}) + \mathrm{s}(\mathrm{sden}) \quad \ldots(6.13)$$

$$\ln(Q_5) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(\mathrm{evap}) + \mathrm{s}(\mathrm{sden}) \quad \ldots(6.14)$$

$$\ln(Q_{10}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(evap) \quad \ldots(6.15)$$

$$\ln(Q_{20}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(evap) \quad \ldots(6.16)$$

$$\ln(Q_{50}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(evap) \quad \ldots(6.17)$$

$$\ln(Q_{100}) = \alpha + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(evap) \quad \ldots(6.18)$$

Table 6.3 Model statistics for the GAM models of clustering group A2

Equation	Predictor variables	Deviance explained (%)	GCV	R^2	F value	p value
Q_2	*area*	81.6	557.43	0.752	5.28	2.99E-02
	$I_{6,2}$				7.36	0.00104
	evap				19.79	4.40E-07
	sden				3.11	0.04745
Q_5	*area*	75.8	4275.5	0.676	5.62	2.55E-02
	$I_{6,2}$				4.34	0.0145
	evap				13.31	8.54E-06
	sden				2.75	0.0679
Q_{10}	*area*	64.1	12348	0.554	3.55	7.01E-02
	$I_{6,2}$				5.06	0.007282
	evap				9.74	0.000105
Q_{20}	*area*	60.2	25633	0.506	3.30	8.01E-02
	$I_{6,2}$				4.20	0.016865
	evap				8.49	0.000297
Q_{50}	*area*	54.7	57731	0.437	2.93	9.81E-02
	$I_{6,2}$				3.22	0.04536
	evap				6.91	0.00119
Q_{100}	$I_{6,2}$	46.7	1.00E+05	0.36	3.77	0.02372
	evap				6.71	0.00155

Model adequacy

Figure 6.10 illustrates the scatter plot between predicted and observed floods for the GAM model of clustering group A2 (consisting of 35 catchments) for Q_{20}. The plot generally exhibits a good agreement between the predicted and observed flood quantiles; however, there are also a noticeable number of outliers showing both overestimation and underestimation, which are mostly found for lower discharges. This might happen due to poor prediction ability of this GAM model. The results show a good scatter around the 45-degree line for medium to large discharge values. Overall, the GAM model for A2 shows a poor performance considering the scatter plots of observed and predicted floods.

Scatter plots of the GAM models for the remaining ARIs for clustering group A2 can be seen in Appendix E (Figure E.11 to E.15). The results show a comparatively better scatter around the 45-degree slope line for 2 and 5 years ARIs. The results show a noticeable range of scatter around the 45-degree line, in particular for lower discharges.

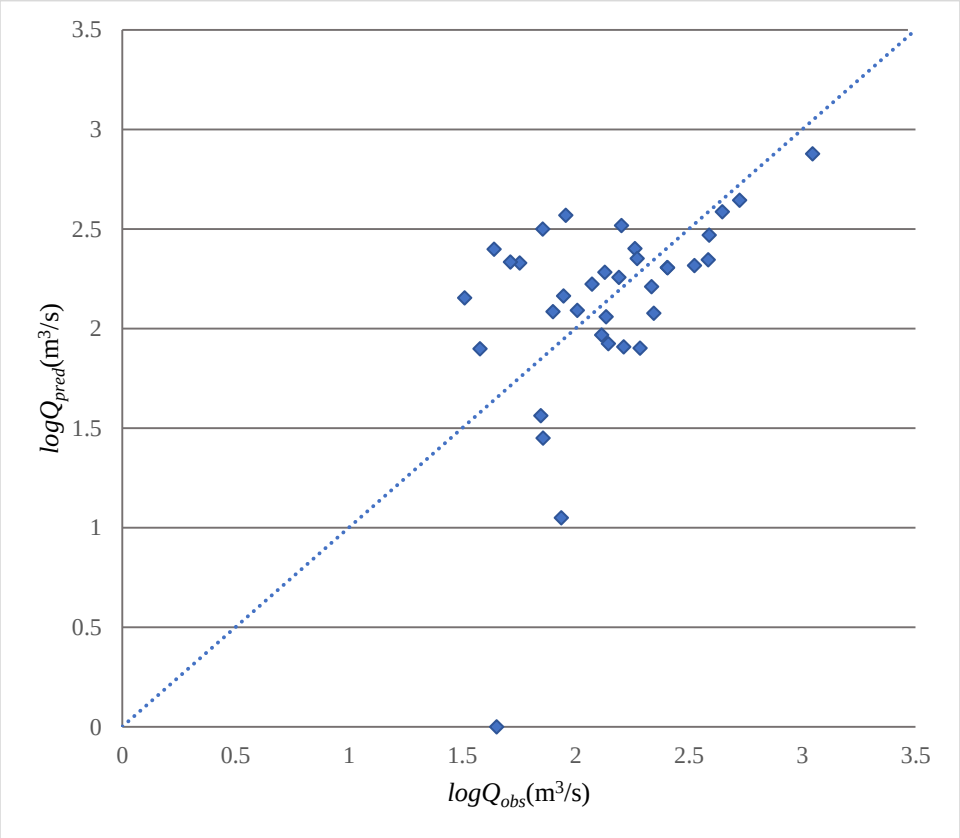

Figure 6.10 Comparison of observed and predicted flood quantiles for GAM for clustering group A2 for Q_{20}

Figure 6.11 shows the boxplots of RE values for the GAM model for clustering group A2. The median RE values show closest match with the 0 – 0 line for ARI of 5 years and slight to medium deviation from the 0 – 0 line is noticed for the remaining ARIs. Median RE values for ARIs of 10, 20 and 50 years show similar deviation around the 0-0 line. For ARIs of 10, 20 and 50 years, a noticeable overestimation is observed. For ARIs of 2 and 100 years, a noticeable underestimation is observed from these boxplots. In terms of the RE band, ARI of 2 years shows the lowest spread. The lower to higher spreads occur for ARIs of 5, 10, 20, 50 and 100 years, respectively. The RE band for 100 years ARI is the highest among all the ARIs. These results show that in terms of RE band, the overall best result is achieved for 2 years of ARI for the GAM model of clustering group A2. According to RE band, it is found that the performances of GAM models of clustering group A2 are relatively poor for the higher ARIs (i.e. 20, 50 to 100 years).

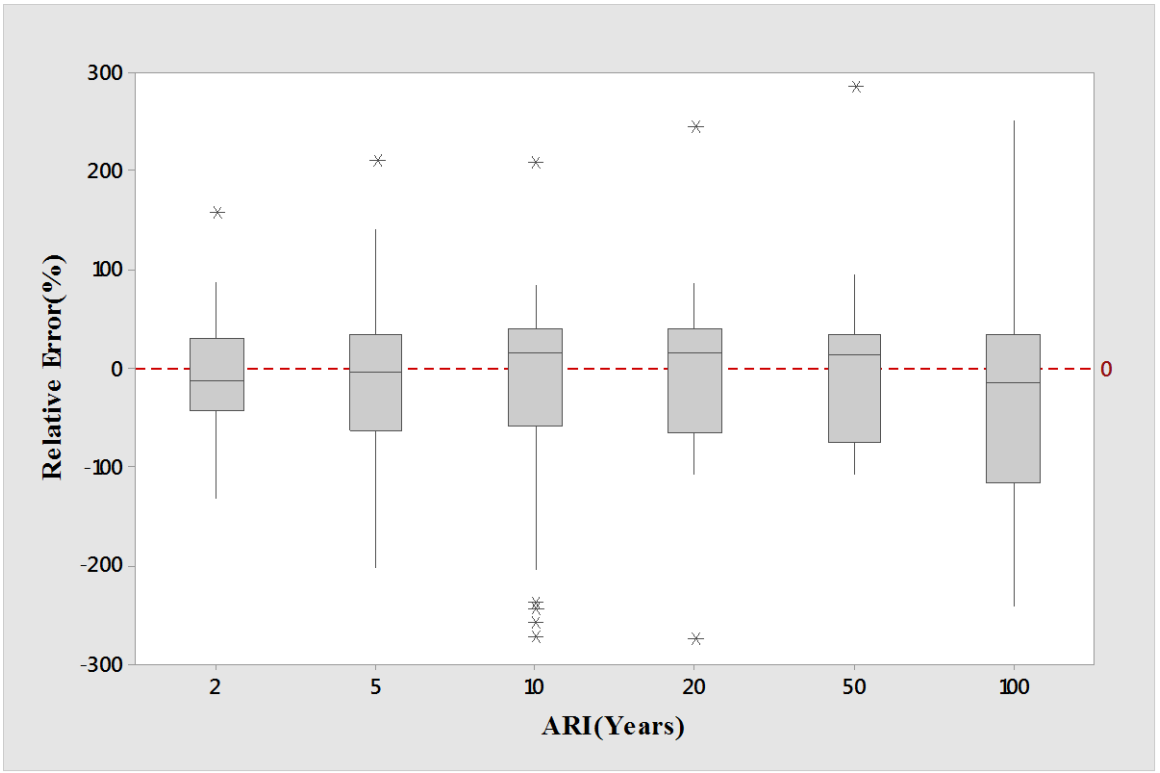

Figure 6.11 Boxplots of RE values for the GAM models for clustering group A2

Figure 6.12 presents the boxplots of the Q_{pred}/Q_{obs} ratio values for different ARIs for the GAM models of clustering group A2. It is found that the median Q_{pred}/Q_{obs} ratio value is located closer to 1 – 1 line for 5 years ARI, which shows the best agreement among all the ARIs. However, for ARIs of 10, 20 and 50 years, the median Q_{pred}/Q_{obs} ratio values are located a short distance below the 1 – 1 line, and for ARIs of 2, 5 and 100 years, the median Q_{pred}/Q_{obs} ratio values are located a short distance above the 1 – 1 line. These results indicate a noticeable underestimation for ARIs of 10, 20 and 50 years and slight to noticeable overestimations of the predicted flood quantiles for ARIs of 2, 5 and 100 years respectively. In terms of the spread of the Q_{pred}/Q_{obs} ratio values, ARI of 2 years exhibits the lowest spread. The ARIs of 5 and 10 years show a similar spread. The spread increases with ARIs for 20, 50 and 100 years. Furthermore, the spreads of the Q_{pred}/Q_{obs}ratio values for 20 and 50 are quite similar and 100 has the largest spread; all of these are remarkably larger than that of 2 year ARI.

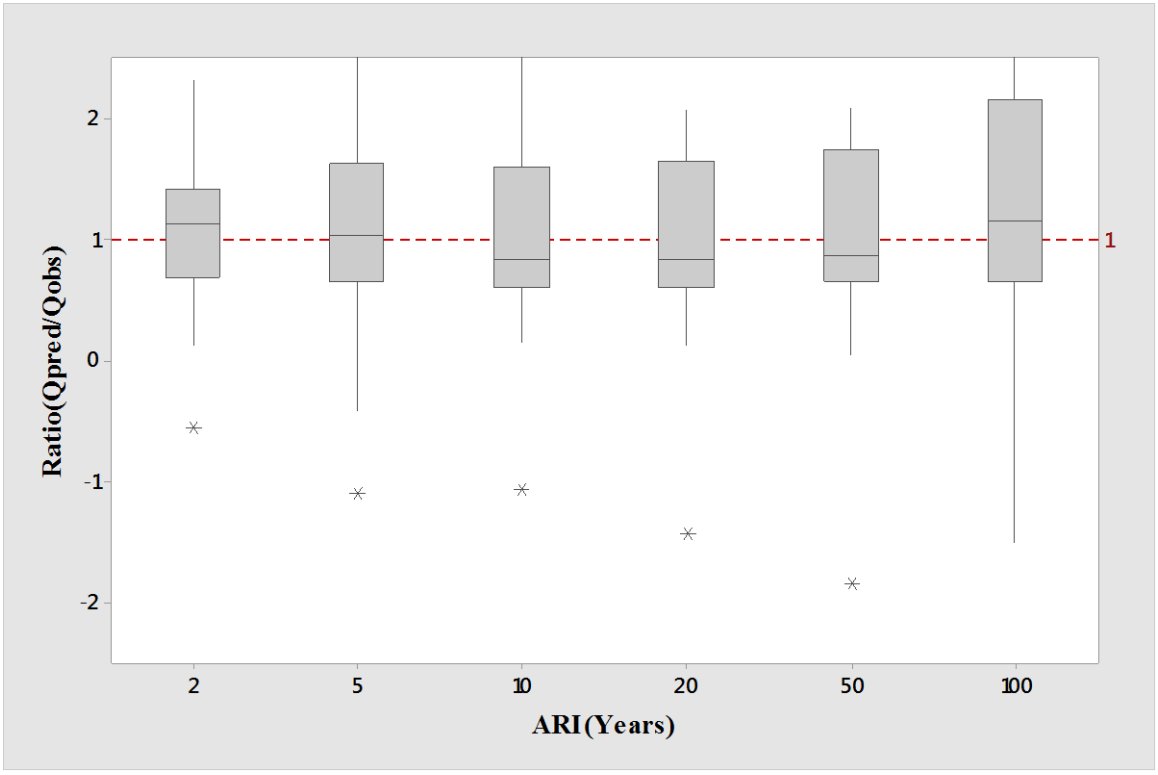

Figure 6.12 Boxplots of Q_{pred}/Q_{obs} ratio for GAM model of clustering group A2

6.3.3. Evaluation of GAM model performance (clustering group B1)

The model statistics for the developed GAM model for clustering group B1 are presented in Table 6.4. The R^2 values are ranged in a decreasing manner from 0.895 to 0.55 for Q_2 to Q_{100}. Relatively smaller R^2 values are found for higher ARIs (e.g., Q_{50} and Q_{100}) indicating a higher degree of error for higher ARI GAM models. Overall, the R^2 values indicate a reasonable performance for Q_2, Q_5, Q_{10} and Q_{20} prediction models.

The GCV values range 219.51 to 82,121.00 for Q_2 to Q_{100} indicating a higher cross validation error as ARI increases.

Table 6.3 contains all the selected predictor variables for individual models along with their respective *p*-statistics. The most important variables for estimating design floods by the GAM models are *area* and $I_{6,2}$ which are present for all the developed GAM models. The second most statistically significant predictor variable appears to be *rain*, which is found in all the prediction models except for Q_2. For Q_2, Q_5, Q_{10} and Q_{20}, *evap* is found statistically significant. Overall, the prediction equations show consistency in the selection of predictor variables for Q_{10}, Q_{20} and Q_{50}. The developed prediction equations by the GAM for group B2 are:

$$\ln(Q_2) = \alpha + \mathrm{s}(area) + s\left(I_{6,2}\right) + \mathrm{s}(\mathrm{SF}) + \mathrm{s}(evap) \quad \ldots(6.13)$$

$$\ln(Q_5) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(rain) + \mathrm{s}(evap) \quad \ldots(6.14)$$

$$\ln(Q_{10}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(rain) + \mathrm{s}(evap) \quad \ldots(6.15)$$

$$\ln(Q_{20}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(rain) + \mathrm{s}(evap) \quad \ldots(6.16)$$

$$\ln(Q_{50}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(rain) \quad \ldots(6.17)$$

$$\ln(Q_{100}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(rain) \quad \ldots(6.18)$$

Table 6.4 Model statistics for GAM model of clustering group B1

Equation	Predictor variables	Deviance explained (%)	GCV	R^2	F value	p value
Q_2	*area*	92.7	219.51	0.895	27.43	< 2e-16
	$I_{6,2}$				2.35	0.083
	SF				2.08	8.21E-02
	evap				12.74	9.64E-11
Q_5	*area*	87.1	1968.5	0.827	25.14	6.86E-12
	$I_{6,2}$				2.32	0.0632
	rain				2.86	2.77E-02
	evap				7.99	4.10E-07
Q_{10}	*area*	83.2	6196.3	0.775	20.48	1.20E-09
	$I_{6,2}$				2.49	4.40E-02
	rain				4.41	0.00341
	evap				4.83	0.000223
Q_{20}	*area*	76.8	15867	0.705	17.28	1.81E-08
	$I_{6,2}$				2.45	4.62E-02
	rain				12.23	0.000312
	evap				2.92	0.009818
Q_{50}	*area*	65.1	43304	0.598	15.18	3.35E-07
	$I_{6,2}$				5.34	0.000113
	rain				20.82	2.49E-05
Q_{100}	*area*	60.9	82121	0.551	13.15	1.35E-06
	$I_{6,2}$				4.41	0.000756
	rain				21.22	2.12E-05

Model Adequacy Checking

The scatter plot of the predicted and the observed flood quantiles for the GAM model of clustering group B1 for 20 years ARI is shown in Figure 6.13. The results show a good scatter with reasonable distance from the 45-degree line for medium to high range flood magnitudes. For lower discharges, a noticeable large scatter is found exhibiting both overestimations and underestimations.

The scatter plots of the GAM model for clustering group B1 for other ARIs can be seen in Appendix E (Figure E.16 to E.20). These plots generally present a good agreement between the predicted and observed flood quantiles. However, there are notable discrepancies

observed for lower discharges showing both overestimations and underestimations. Overall, the GAM model for clustering group B1 shows reasonable results for higher discharges.

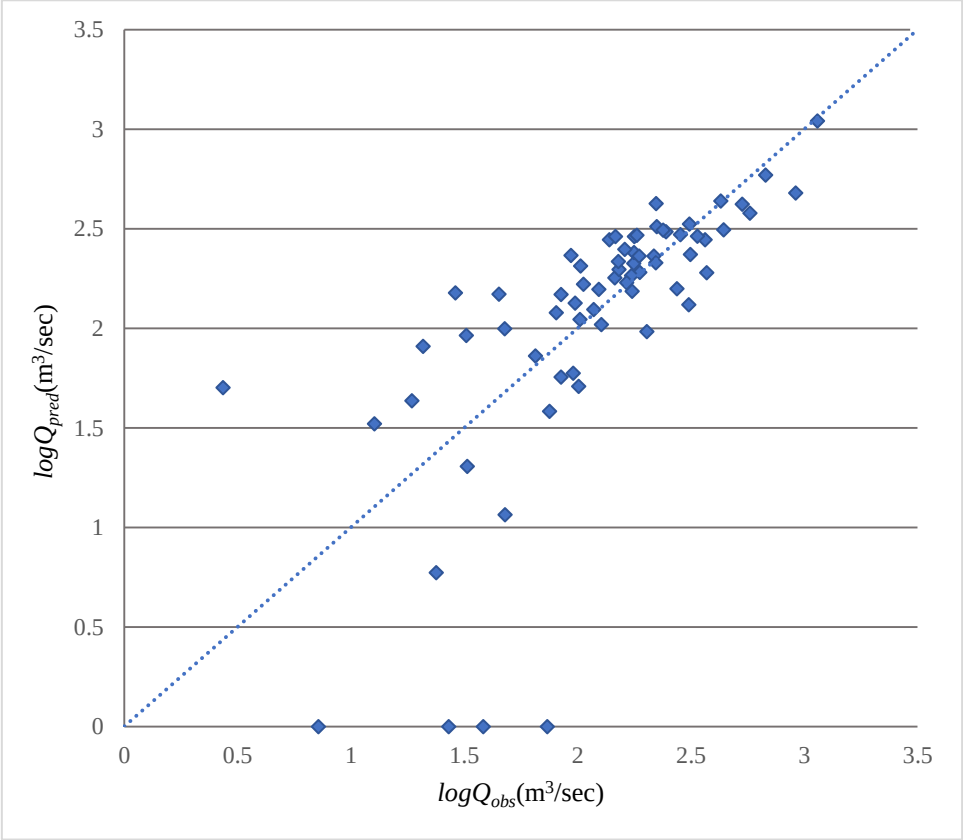

Figure 6.13 Comparison of observed and predicted flood quantiles for GAM model of clustering group B1 for Q_{20}

Figure 6.14 shows the boxplots of RE values for the GAM model for clustering group B1. A large number of outliers can be observed from the predictions. The median RE values match with the 0 – 0 line very well for ARI of 5 and 10 years, and reasonably well for ARIs of 2 and 20 years. For ARIs of 20, 50 and 100 years, slight to noticeable underestimations are provided. In terms of the RE band, ARI of 2 years shows the lowest spread among all the REs. The second lowest spread is found for 5 years ARI, which is more than twice than that of 2 years ARI. The lower to higher range of spreads are seen for ARIs of 5, 10, 20, 50 and 100 years, respectively. The RE band for 100 years ARI is the highest among all the ARIs. These results show that in terms of RE, the best result overall is achieved for 2 years ARI for the GAM model for clustering group B1. According to RE band, it is revealed that the

performance of this GAM model is relatively poorer for the higher ARIs (i.e. 50 to 100 years) due to larger uncertainty associated with the estimation of higher discharges.

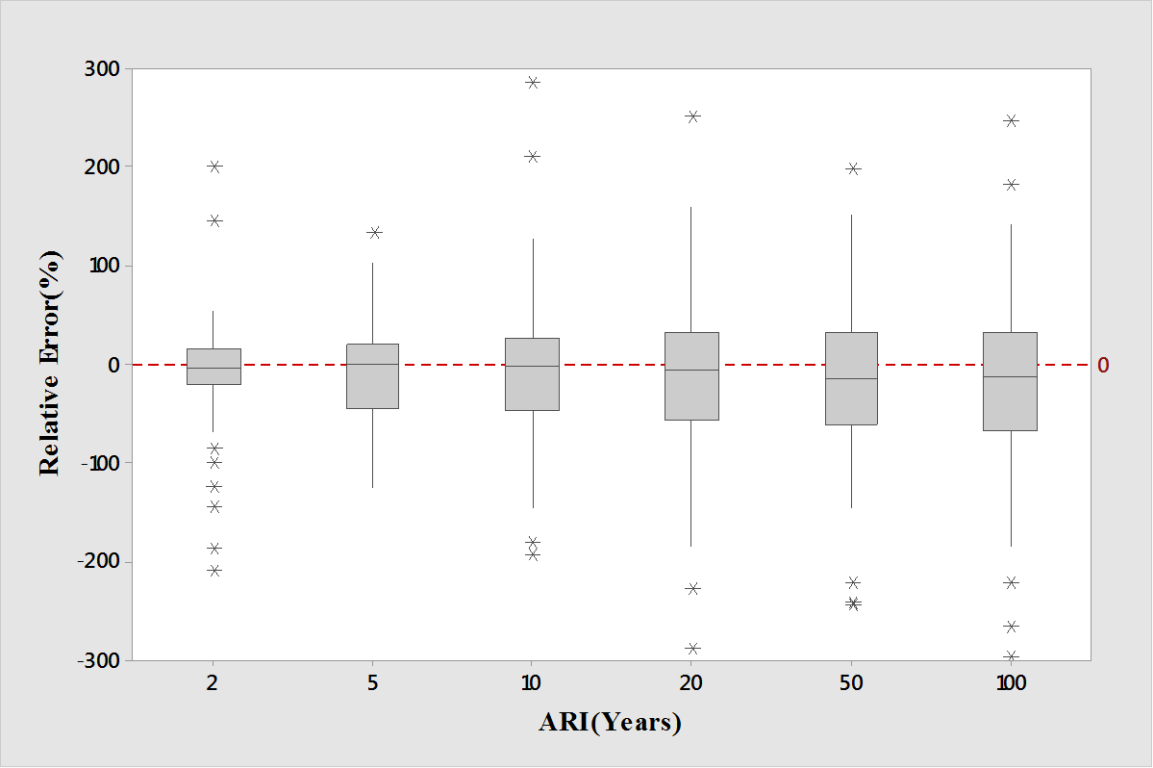

Figure 6.14 Boxplots of RE values for GAM for clustering group B1

Figure 6.15 presents the boxplots of the Q_{pred}/Q_{obs} ratio values for the GAM model for clustering group B1 for different ARIs. It is found that the median Q_{pred}/Q_{obs} ratio values are located closer to 1 – 1 line, in particular for ARIs of 5 and 10 years with best agreement is for ARI of 5 years, and reasonable agreement for ARIs of 2 and 20 years. These results indicate a good prediction by the GAM model for clustering group B1. The highest median Q_{pred}/Q_{obs} ratio value is found for ARI of 100 years. In terms of the spread of the Q_{pred}/Q_{obs} ratio values, ARI of 2 years exhibits the lowest spread followed by ARIs of 5, 10, 20, 50 and 100 years. Furthermore, the spreads of the Q_{pred}/Q_{obs} ratio values for 50 and 100 years are very similar, which are remarkably larger than 2, 5 and 10 years.

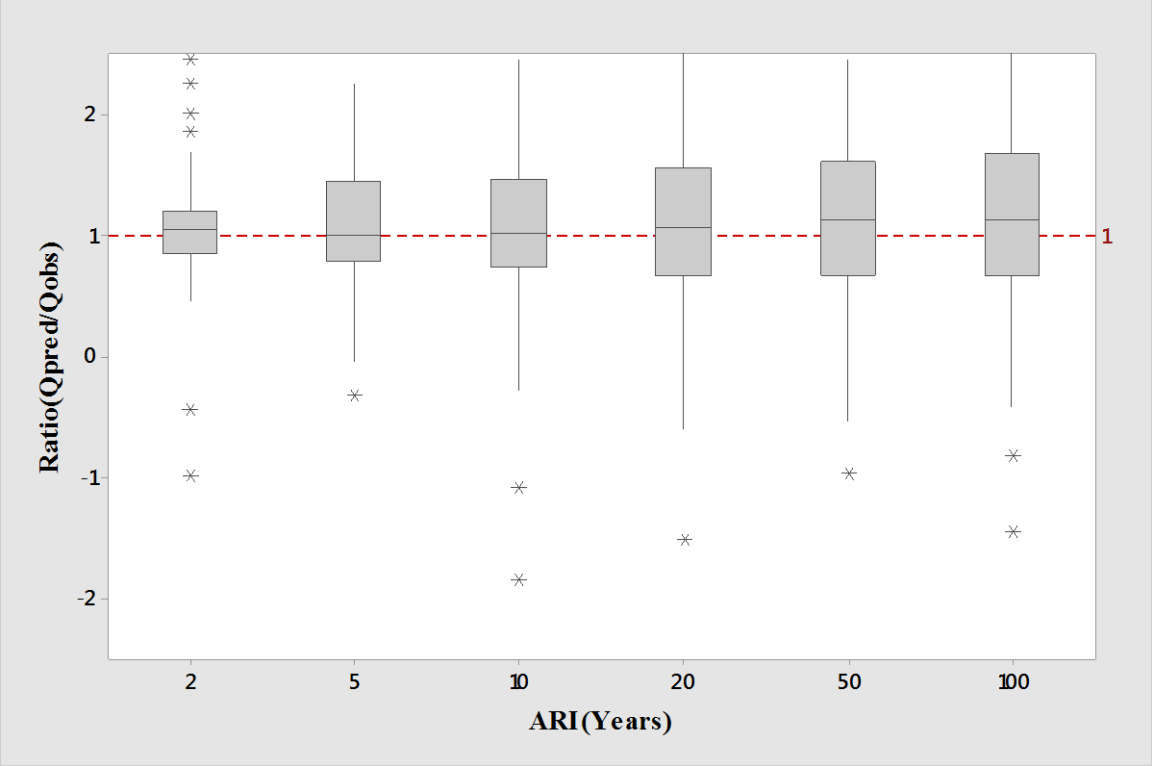

Figure 6.15 Boxplots of Q_{pred}/Q_{obs} **ratio values for the GAM for clustering group B1**

6.3.4. Evaluation of GAM model performance (clustering group B2)

The model statistics for the developed GAM model for clustering group B2 is summarised in Table 6.5. The R^2 values are ordered in a decreasing magnitude from 0.712 to 0.30 for Q_2 to Q_{100}. Smaller R^2 values are found for higher ARIs similar to other GAM models. The GCV values vary from 494.3 to 77,576 for Q_2 to Q_{100} indicating a higher cross validation error for the higher ARI GAM models.

Table 6.5 contains the selected predictor variables in the GAM models along with respective *p*-statistics. The most common variables for estimating design floods for these GAM models are *area* and *evap*. The next most statistically significant predictor variables are $I_{6,2}$ and *sden*. $I_{6,2}$ is found in prediction models of Q_2, Q_5 and Q_{20} and *sden* is present in prediction models of Q_2, Q_5 and Q_{10}. The developed prediction equations by the GAM for clustering group B2 are:

$$\ln(Q_2) = \alpha + \mathrm{s}(area) + s\left(I_{6,2}\right) + \mathrm{s}(evap) + \mathrm{s}(sden) \quad \ldots(6.19)$$

$$\ln(Q_5) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(evap) + \mathrm{s}(sden) \quad \ldots(6.20)$$

$$\ln(Q_{10}) = \alpha + \mathrm{s}(area) + \mathrm{s}(evap) + \mathrm{s}(sden) \quad \ldots(6.21)$$

$$\ln(Q_{20}) = \alpha + \mathrm{s}(area) + \mathrm{s}\left(I_{6,2}\right) + \mathrm{s}(evap) \quad \ldots(6.22)$$

$$\ln(Q_{50}) = \alpha + \mathrm{s}(area) + \mathrm{s}(evap) \quad \ldots(6.23)$$

$$\ln(Q_{100}) = \alpha + \mathrm{s}(area) + \mathrm{s}(evap) \quad \ldots(6.24)$$

Table 6.5 Model statistics for GAM model for clustering group B2

Equation	Predictor variables	Deviance explained (%)	GCV	R^2	F value	p value
Q_2	*area*	76.6	494.3	0.712	12.79	9.52E-04
	$I_{6,2}$				6.15	0.001994
	evap				21.97	1.68E-08
	sden				4.92	8.60E-03
Q_5	*area*	69.5	3528.2	0.626	13.98	5.90E-04
	$I_{6,2}$				3.19	0.04364
	evap				14.99	1.41E-06
	sden				3.67	0.02884
Q_{10}	*area*	56.4	9694.8	0.506	20.33	4.88E-05
	evap				7.77	0.000265
	sden				4.65	0.009152
Q_{20}	*area*	53.4	20616	0.456	9.57	3.58E-03
	$I_{6,2}$				3.46	3.51E-02
	evap				8.28	0.000217
Q_{50}	*area*	37.1	47250	0.322	12.00	1.20E-03
	evap				3.73	1.39E-02
Q_{100}	*area*	35.2	77576	0.3	11.82	1.29E-03
	evap				2.97	2.96E-02

The predicted and the observed flood quantiles for the GAM model for 20 years ARI for B2 is shown in Figure 6.16. The plot presents a reasonable agreement between the predicted and observed flood quantiles. Overall, the GAM based RFFA model shows a reasonable result for 20 years of ARI. The remaining GAM models for clustering group B2 can be seen in Appendix E (Figure E.20 to E.25). These plots generally present a good agreement between the predicted and observed flood quantiles for the lower ARIs (2, 5 and 10 years) except for a few outliers.

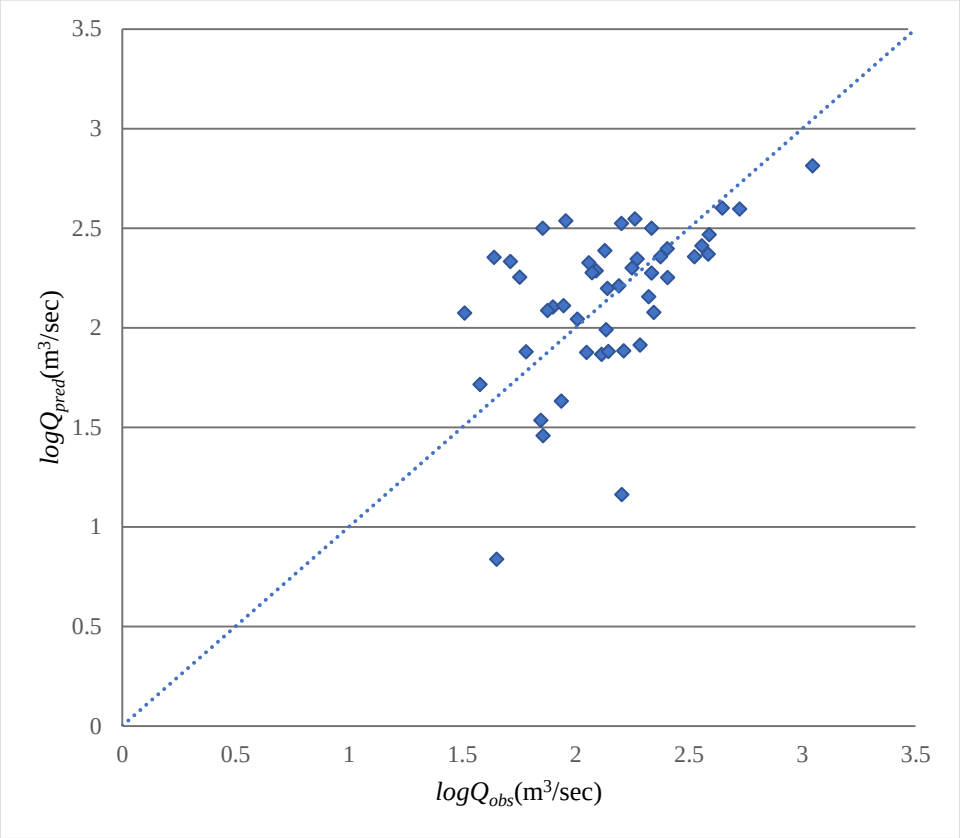

Figure 6.16 Comparison of observed and predicted flood quantiles for GAM model for clustering group B2 for Q_{20}

Figure 6.17 shows the boxplots of RE values for the GAM model for clustering group B2. The median RE values are relatively closer to the 0 – 0 line for ARIs of 20, 50 and 100 years. The boxplots show very few outliers, indicating a better GAM model in comparison to other groups. Median RE values do not scatter much from the 0-0 line, which indicate minimum underestimations and overestimations for this GAM model. In terms of the RE band, ARI of 2 years shows the lowest spread, which is slightly lower than RE band of 5 and 10 years of ARIs. The RE band of 5 and 10 years of ARIs show a similar result. The lower to higher spread levels are seen for ARIs of 2, 5, 10, 20, 50 and 100 years, respectively.

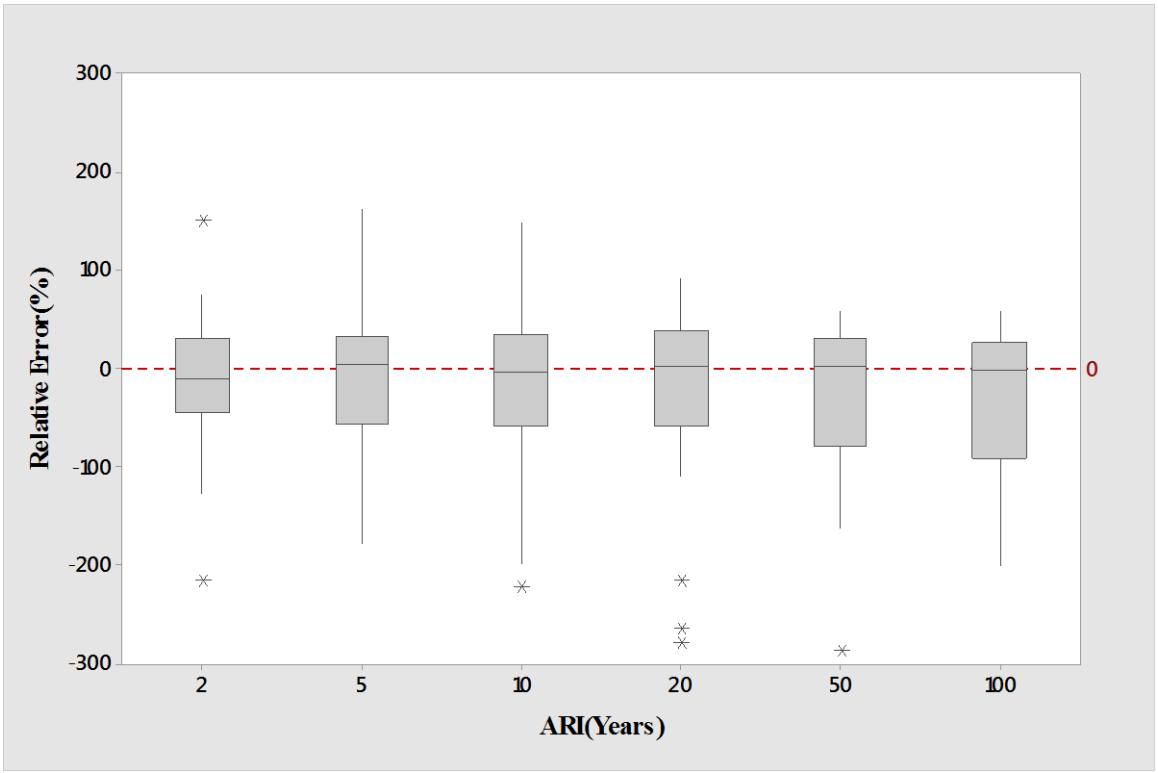

Figure 6.17 Boxplots of RE values for GAM for clustering group B2

Figure 6.18 presents the boxplots of the Q_{pred}/Q_{obs} ratio values for the GAM model for clustering group B2 for different ARIs. It is found that the median Q_{pred}/Q_{obs} ratio values are quite closer to 1 – 1 line for ARIs of 10, 20, 50 and 100 years, with the best agreement found for ARI of 20 years. However, for ARI of 2 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance above the 1 – 1 line and for ARI of 5 years, the median Q_{pred}/Q_{obs} ratio value is located a short distance below the 1 – 1 line. These results indicate noticeable overestimations and underestimations of the predicted flood quantiles by the GAM model for 2 years and 5 years ARI, respectively. In terms of the spread of the Q_{pred}/Q_{obs} ratio values, ARI of 2 years exhibits the lowest spread. The ARIs of 5, 10 and 20 shows almost similar spread, and 50 and 100 years ARIs with the largest spread. Furthermore, the spreads of the Q_{pred}/Q_{obs} ratio values for 50 and 100 years are very similar, which are again remarkably larger than 2, 5 and 10 years.

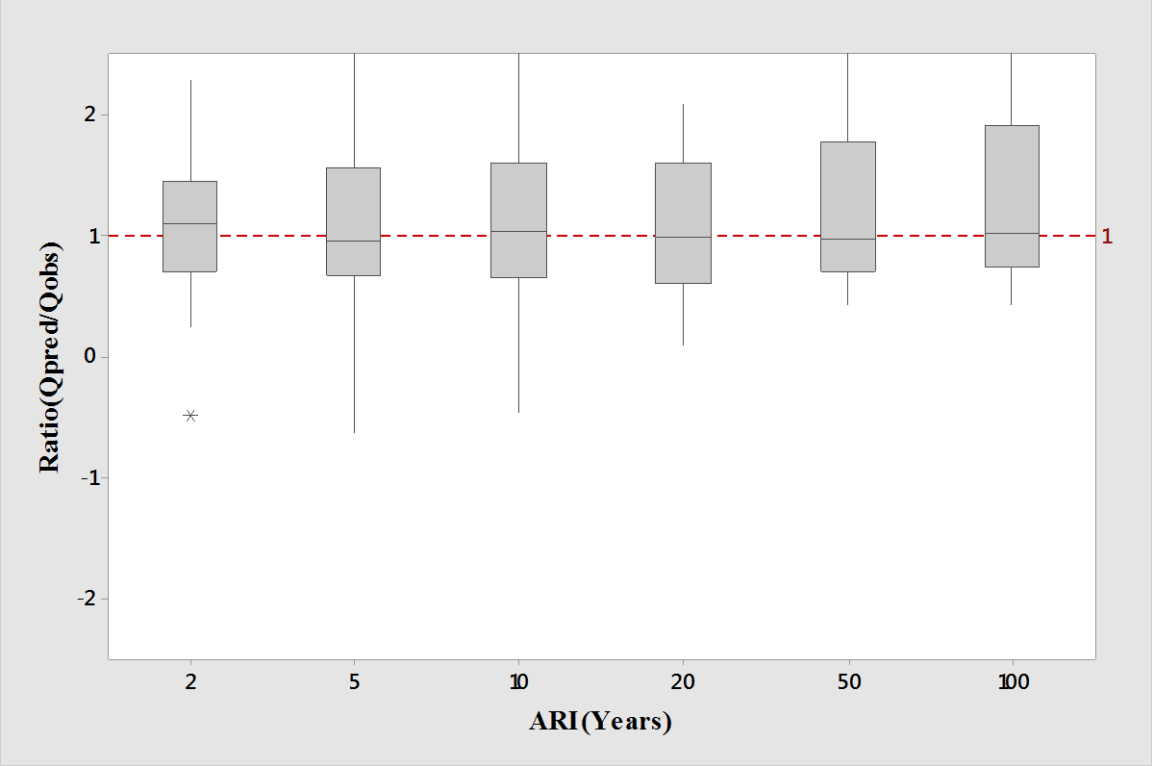

Figure 6.18 Boxplots of median Q_{pred}/Q_{obs} ratio for GAM for clustering group B2

.

6.4. Comparison of performances of the GAM models based on numerical measures

6.4.1. Median RE

Table 6.6 summarizes the median RE values of GAM models of the combined and clustering groups A1, A2, B1, B2. From the results of the combined group, median RE values range from 33.75 % to 49.09 %. The smallest and highest median RE values are found for 10 years and 100 years of ARIs, respectively. This model is ranked three with overall median RE value of 38.04 % (from Table 6.8).

For GAM models of clustering group A1, the ranges of median RE values are 22.52 % to 53.38 %. The median RE increases with the increasing ARIs except for 5 and 10 years of ARIs where median RE value of 10 years ARI is lower than that of 5 years ARI (31.96 % and 33.10 %, respectively). The overall median RE values for clustering group A1 is found to be 36.77 % which places it at rank 2 among the 5 GAM models (Table 6.8).

For GAM model of clustering group A2, the smallest to largest value of median RE is found to be 39.31 % and 49.59 %, which are for 2 years and 50 years of ARIs, respectively. The difference between smallest to highest value of median RE values for the GAM model of clustering group A2 is smaller compared to other GAM models. The median RE value for 2 years ARI is found to be 39.31%, which is the highest median RE for 2 years ARI among all the GAM models. The overall median RE values for A2 clustering group is found as 43.73 %, which places it at rank 5 among the 5 clustering groups of GAM model (Table 6.8).

In case of clustering group B1, median RE values range from 16.80 % to 45.9 %. This model shows a large median RE value for higher ARIs with almost similar results for 20 years, 50 years and 100 years of ARIs. The overall median RE values for clustering group B1 is found to be 35.10 % which places it at rank 1 among the 5 clustering groups of GAM model (Table 6.8).

GAM model of clustering group B2 shows lowest median RE value as 33.24 % and highest median RE value as 45.82 % which are for 2 years and 20 years of ARIs respectively. The overall median RE for clustering group B2 is found as 38.13 %, which places it at rank 4 among the 5 clustering groups of GAM model.

Table 6.6 Median RE between combined data and clustering groups for GAM

Flood quantile	Combined	A1	A2	B1	B2
Q_2	34.81	22.52	39.31	16.80	33.24
Q_5	33.88	33.10	41.46	28.92	41.11
Q_{10}	33.75	31.96	40.29	34.46	38.17
Q_{20}	34.05	39.53	42.35	42.47	45.82
Q_{50}	42.67	40.12	49.59	42.08	31.38
Q_{100}	49.09	53.38	49.37	45.90	39.04
Overall	38.04	36.77	43.73	35.10	38.13

6.4.2. Median Ratio

Table 6.7 summarises the median Q_{pred}/Q_{obs} ratio values of the five different GAM models. All the GAM models are ranked according to the overall median Q_{pred}/Q_{obs} ratio values in Table 6.8.

For the GAM model of the combined group, the median Q_{pred}/Q_{obs} ratio values range from 1.02 to 1.12. All the median Q_{pred}/Q_{obs} ratio values of the combined group is found slightly higher than 1, which hints to overall overestimation by the GAM models. The best result is obtained for 5 years of ARI which is 1.02. The median Q_{pred}/Q_{obs} ratios for ARIs of 20, 50 and 100 years is 1.12. In summary, the GAM model for the combined group shows a reasonable overall result with median Q_{pred}/Q_{obs} ratio of 1.08, which puts it at rank 4 among all the five GAM models.

For clustering group A1, the median Q_{pred}/Q_{obs} ratio values range from 1.01 to 1.16 which is for 2 years and 50 years of ARIs, respectively. All the median Q_{pred}/Q_{obs} ratio values indicate

towards an overestimation by the GAM. The highest median Q_{pred}/Q_{obs} ratio value is 1.16, which is found for the clustering group A1 for 50 years of ARI. The overall median Q_{pred}/Q_{obs} ratio value is found to be 1.08, ranking it 5 among the 5 GAM models of the clustering groups.

For clustering group A2, the flood quantiles seem to have reasonable performance with the lowest value of 0.86 and highest value of 1.14 for 20 years and 100 years of ARIs, respectively. This models show a similar range of underestimation for the 10, 20 and 50 years of ARIs with median Q_{pred}/Q_{obs} ratio value of 0.83, 0.84 and 0.86, respectively. The overall median Q_{pred}/Q_{obs} ratio value is found to be 0.97, which ranks it 2nd among the five GAM models.

For the GAM model of the clustering group B1, the median Q_{pred}/Q_{obs} ratio ranges from 1.00 to 1.14 for 5 and 100 years of ARIs, respectively. The predicted flood quantiles are overestimated by this GAM model with a median Q_{pred}/Q_{obs} ratio value of 1.07, which ranks it at position 3 among the 5 GAM models.

For the GAM model of clustering group B2, the predicted values are underestimated for 5, 20 and 50 years of ARIs with the median Q_{pred}/Q_{obs} ratio values of 0.95, 0.98 and 0.98, respectively. However, the highest median Q_{pred}/Q_{obs} ratio value is found as 1.10, which is for 2 years of ARI. Overall, median Q_{pred}/Q_{obs} ratio value is found to be 1.01; therefore, the GAM model for clustering group B2 ranks 1 among the 5 GAM models.

Table 6.7 Median Q_{pred}/Q_{obs} ratio comparison between groups for GAM

Flood quantile	Combined group	A1	A2	B1	B2
Q_2	1.07	1.01	1.13	1.05	1.10
Q_5	1.02	1.03	1.04	1.00	0.95
Q_{10}	1.04	1.06	0.83	1.02	1.04
Q_{20}	1.12	1.10	0.84	1.06	0.98
Q_{50}	1.12	1.16	0.86	1.14	0.98
Q_{100}	1.12	1.12	1.14	1.13	1.01
Overall	1.08	1.08	0.97	1.07	1.01

6.4.3. Ranking of GAM models

Table 6.8 presents a subjective ranking of the GAM models for the four clustering groups and combined group based on median RE and median Q_{pred}/Q_{obs} ratio values. None of the GAM models are found to be equally well with respect to all the assessment criteria, which makes it difficult to select the best performing GAM model.

Table 6.8 Comparing the overall performance of GAM models

Criteria	Rank 1	Rank 2	Rank 3	Rank 4	Rank5
Median RE (%)	B1	A1	Combined	B2	A2
Median Q_{pred}/Q_{obs} ratio	B2	A2	B1	Combined	A1

6.5. Overall performance comparison

The following sub-sections give an overall assessment of the model performances of five different log-log linear models and five different GAM models.

6.5.1. R^2

The R^2 values of the 10 different RFFA models are compared in Table 6.9. From this table, R^2 values from the GAM models are found higher than respective log-log linear model for smaller ARIs. It is also found that GAM models based on clustering groups give better results, e.g. models for smaller ARIs show better R^2 values. For example, the R^2 values of Q_2, Q_5 and Q_{10} for GAM models of the combined group is found to be 0.83, 0.73 and 0.70, respectively which are 10%, 8% and 4% higher than respective log-log linear models. On the other hand, GAM models show comparatively lower R^2 values than respective log-log linear models for higher ARIs of flood (e.g., 0.67, 0.58 and 0.51, which are 1%, 10% and 17% lower than respective log-log linear model). Also, the GAM models of clustering groups give better results for Q_2 with the highest value of 0.90.

Overall, the log-log linear models give better performance for higher ARIs (i.e., 20, 50 and 100 years) and GAM models show better performance for smaller ARIs (i.e., 2, 5 and 10 years).

Table 6.9 R^2 values of the GAM and log-log linear models for 10 cases

Flood quantile	Combined group		Group (A1)		Group (A2)		Group (B1)		Group (B2)	
	log-log linear model	GAM	log-log linear model	GAM	log-log linear model	GAM	log-log linear model	GAM	log-log linear model	GAM
Q_2	0.69	0.69	0.74	0.83	0.69	0.75	0.78	0.90	0.65	0.712
Q_5	0.67	0.66	0.72	0.79	0.55	0.676	0.74	0.83	0.57	0.626
Q_{10}	0.63	0.62	0.70	0.73	0.48	0.554	0.71	0.78	0.48	0.506
Q_{20}	0.61	0.56	0.68	0.67	0.43	0.506	0.69	0.71	0.42	0.456
Q_{50}	0.57	0.50	0.65	0.58	0.32	0.437	0.65	0.60	0.39	0.322
Q_{100}	0.53	0.44	0.62	0.51	0.27	0.36	0.62	0.55	0.32	0.30
Overall	0.62	0.58	0.69	0.69	0.46	0.55	0.70	0.73	0.47	0.49

6.5.2. Median RE

In Table 6.10, the median RE values are summarised for the log-log linear and GAM models for the one combined and four clustering groups. The median RE values are calculated considering the absolute relative error value of the test catchments. The highest RE is 59.94 %, which is found for log-log linear model for the clustering group A2 for 100 years of ARI, and the lowest RE is 16.8 %, which is found for the GAM model of group B1 data for 2 years ARI.

For the log-log linear models, median RE values range from 18.73 % to 59.94 %. The smallest and highest median RE values are found for the log-log linear models of the combined group for 2 years of ARI and clustering group A2 for 100 years ARI, respectively. From the overall median RE values for the log-log linear models, the smallest result is found from clustering group A1 with median RE of 31.11 %. The overall highest median RE value for the log-log linear model is found from clustering group A2 with the value of 42.40 %. The overall median RE values range from 31.11 % to 42.40 %, which indicate that the median RE does not differ much between different groups of the log-log linear models. Lowest values of RE are mostly found from 2 years of ARI for log-log linear model, which range from 18.73% to 30.33 % which are for the combined group and clustering group B1, respectively. The highest values of RE are found for 100 years ARI for the log-log linear models, which range from 37 % to 59.94 %, which are for clustering groups B1 and A2, respectively.

Table 6.10 Median RE values (%) for the GAM and log-log linear model based RFFA techniques for ten cases

Flood quantile	Combined group		Group (A1)		Group (A2)		Group (B1)		Group (B2)	
	log-log linear model	GAM	log-log linear model	GAM	log-log linear model	GAM	log-log linear model	GAM	log-log linear model	GAM
Q_2	18.73	34.81	29.56	22.52	23.10	39.31	30.33	16.80	25.82	33.24
Q_5	32.88	33.88	28.60	33.10	34.69	41.46	28.20	28.92	31.97	41.11
Q_{10}	19.36	33.75	27.47	31.96	40.54	40.29	27.37	34.46	33.05	38.17
Q_{20}	34.51	34.05	30.74	39.53	43.02	42.35	29.37	42.47	36.69	45.82
Q_{50}	40.41	42.67	33.25	40.12	53.10	49.59	37.42	42.08	39.29	31.38
Q_{100}	40.99	49.09	37.05	53.38	59.94	49.37	37.00	45.90	42.63	39.04
Overall	31.15	38.04	31.11	36.77	42.40	43.73	31.61	35.10	34.91	38.13

In case of GAM, median RE values range from 16.8 % to 53.38 %. The smallest and highest median RE values are found for 2 years of ARI for clustering group B1, and for 100 years of ARI for clustering group A1, respectively. With respect to the overall median RE, the smallest value is found for the clustering group B1 with median RE of 35.10 %. The overall highest median RE value is found for clustering group A2 (43.73 %). The overall median RE values range from 35.10 % to 43.73 % for the GAM models. Lower values of median RE are mostly found for 2 years of ARI for the GAM, which range from 16.80 % to 39.31 % (for the clustering groups B1 and A2). The highest values of RE are found for 100 years of ARI for the GAM, which ranges from 39.04 % (clustering group B2) to 53.38 % (clustering group

A1). It is observed that in most cases, the median RE values of GAM are greater than respective log-log linear models. For group A2, median RE values of the GAM models are lower than the log-log linear models for ARIs of 10, 20, 50 and 100 years. However considering overall performance of median RE, log-log linear model is found to have better accuracy than GAM.

Figure 6.19 presents the comparative performance of the log-log linear and GAM models with respect to median RE. It shows that, overall lower range of median RE values are observed for the log-log linear model for clustering group A1. However, overall, the highest range of median RE is observed for the log-log linear model for clustering group A2. Although, the highest range of median RE is found for the clustering group A2, the remaining groups of log-log linear model outperform the respective GAM models. Overall, log-log linear models show better results than the GAM models with respect to median RE.

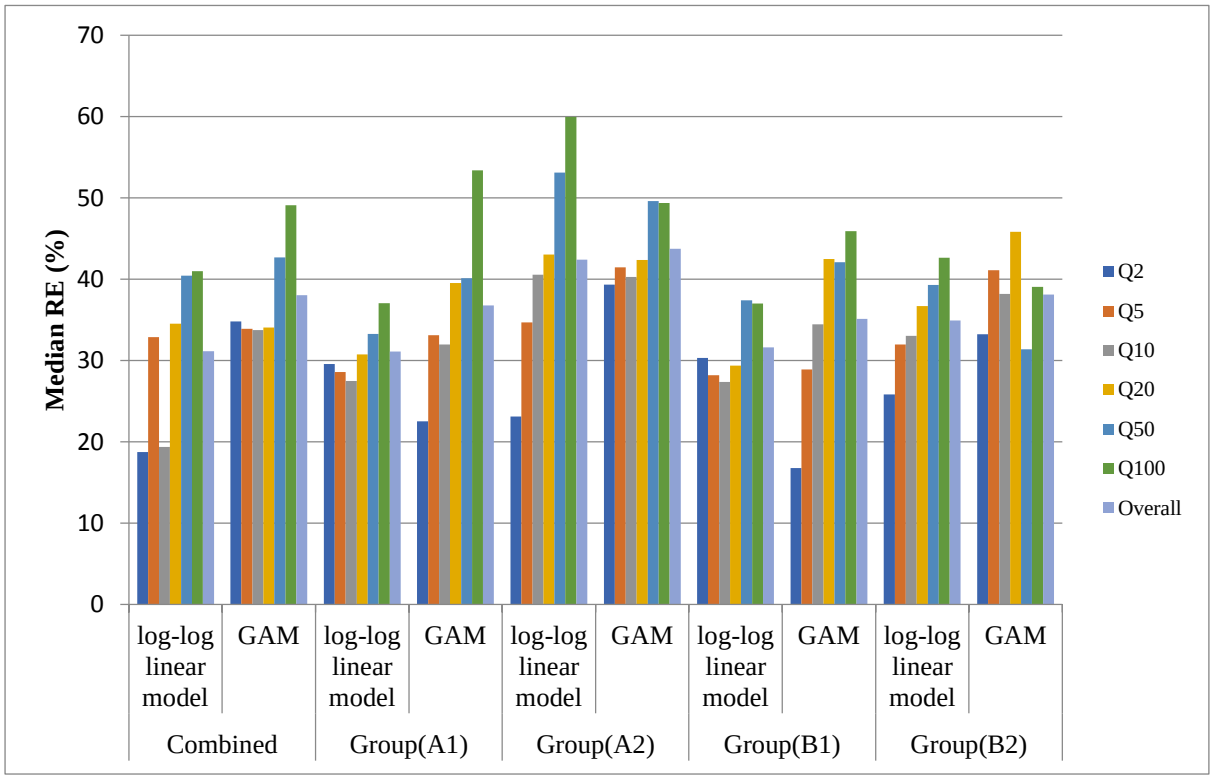

Figure 6.19 Plot of median RE values for different log-log linear and GAM models

6.5.3. Median Ratio (Q_{pred}/Q_{obs})

In Table 6.11, the median ratio (Q_{pred}/Q_{obs}) values are summarised for 5 log-log linear models and 5 GAM models. The median ratio values are important as these are considered to be an effective indicator of overestimation or underestimation (i.e. a measure of bias) of the prediction model. The highest Q_{pred}/Q_{obs} ratio is 1.16, which is found for the log-log linear model for clustering group A1 for ARI of 50 years, and the lowest median Q_{pred}/Q_{obs} ratio is 0.83, which is found for GAM model for clustering group A2 data of 10 years of ARI.

For log-log linear models, median Q_{pred}/Q_{obs} ratio values range from 0.90 to 1.09. The smallest and highest median ratio values are found for 100 years of ARI for the log-log linear model of the clustering group B2 and log-log linear model of the clustering group B1, respectively. The overall smallest median Q_{pred}/Q_{obs} ratio values for the log-log linear models are found as 0.96, which is for the clustering group B2 and the highest median Q_{pred}/Q_{obs} ratio value for log-log linear model is found for the clustering group B1, which is 1.01. The overall median ratio values range from 0.96 to 1.01, which indicate a very small percentage of difference between different groups of the log-log linear models. Most of the median Q_{pred}/Q_{obs} ratio values obtained from log-log linear model are in the range of 0.95 to 0.99, which indicate a slight underestimated prediction of flood quantiles. The best result is obtained for 20 and 5 years of ARIs for the combined group, with the median ratio value of 1.00. In summary, log-log linear model-based RFFA techniques show a very reasonable and consistent median Q_{pred}/Q_{obs} ratio value.

In case of GAM, median Q_{pred}/Q_{obs} ratio values range from 0.83 to 1.16. The smallest and highest median Q_{pred}/Q_{obs} ratio values are found for ARIs of 10 years for the clustering group A2 and 50 years of ARI for the clustering group A1, respectively. The overall smallest median Q_{pred}/Q_{obs} ratio value for GAM is found for clustering group A2 with median Q_{pred}/Q_{obs} ratio of 0.97. The overall highest median Q_{pred}/Q_{obs} ratio value is found for combined group with median ratio of 1.08. The overall median Q_{pred}/Q_{obs} ratio value ranges from 0.98 to 1.08, which indicates that GAM tends to make an overestimation. Moreover, the overall median Q_{pred}/Q_{obs} ratio values for the GAM models are higher compared with respective log-log linear models. Most of the median Q_{pred}/Q_{obs} ratio values are found above 1.00 for the GAM models, which indicates an overestimation. Lower values of median Q_{pred}/Q_{obs} ratio values for GAM are mostly found for the clustering group A2 that ranges from 0.83 to 1.14, which are comparatively lower than median Q_{pred}/Q_{obs} ratio values of the

log-log linear models of the clustering group A2. For clustering group A2, median Q_{pred}/Q_{obs} ratio values are lower for the GAM than the log-log linear models for higher ARIs i.e., for 10, 20 and 50 years. However, in the most cases, the median Q_{pred}/Q_{obs} ratio values of GAM are greater than the respective log-log linear models. Overall, median Q_{pred}/Q_{obs} ratio values indicate that the log-log linear models produce better predictions than GAM.

Figure 6.17 plots the median ratio values of the log-log linear and GAM based RFFA techniques for different ARIs considering all the ten groups. It shows that the log-log linear model maintains a better consistency with smaller levels of fluctuations in median ratio values than the GAM.

Table 6.11 Median Q_{pred}/Q_{obs} ratio values for the GAM and log-log linear model based RFFA techniques for 10 cases

Flood quantile	Combined		Group (A1)		Group (A2)		Group (B1)		Group (B2)	
	log-log linear model	GAM	log-log linear model	GAM	log-log linear model	GAM	log-log linear model	GAM	log-log linear model	GAM
Q_2	1.03	1.07	1.04	1.01	1.00	1.13	1.01	1.05	1.04	1.10
Q_5	1.00	1.02	0.95	1.03	0.99	1.04	0.98	1.00	1.03	0.95
Q_{10}	0.97	1.04	0.94	1.06	0.98	0.83	0.96	1.02	0.92	1.04
Q_{20}	1.00	1.12	0.97	1.10	1.01	0.84	1.01	1.06	0.94	0.98
Q_{50}	0.98	1.12	1.02	1.16	0.95	0.86	1.05	1.14	0.94	0.98
Q_{100}	0.94	1.12	1.02	1.12	0.95	1.14	1.09	1.13	0.90	1.01
Overall	0.99	1.08	0.99	1.08	0.98	0.97	1.01	1.07	0.96	1.01

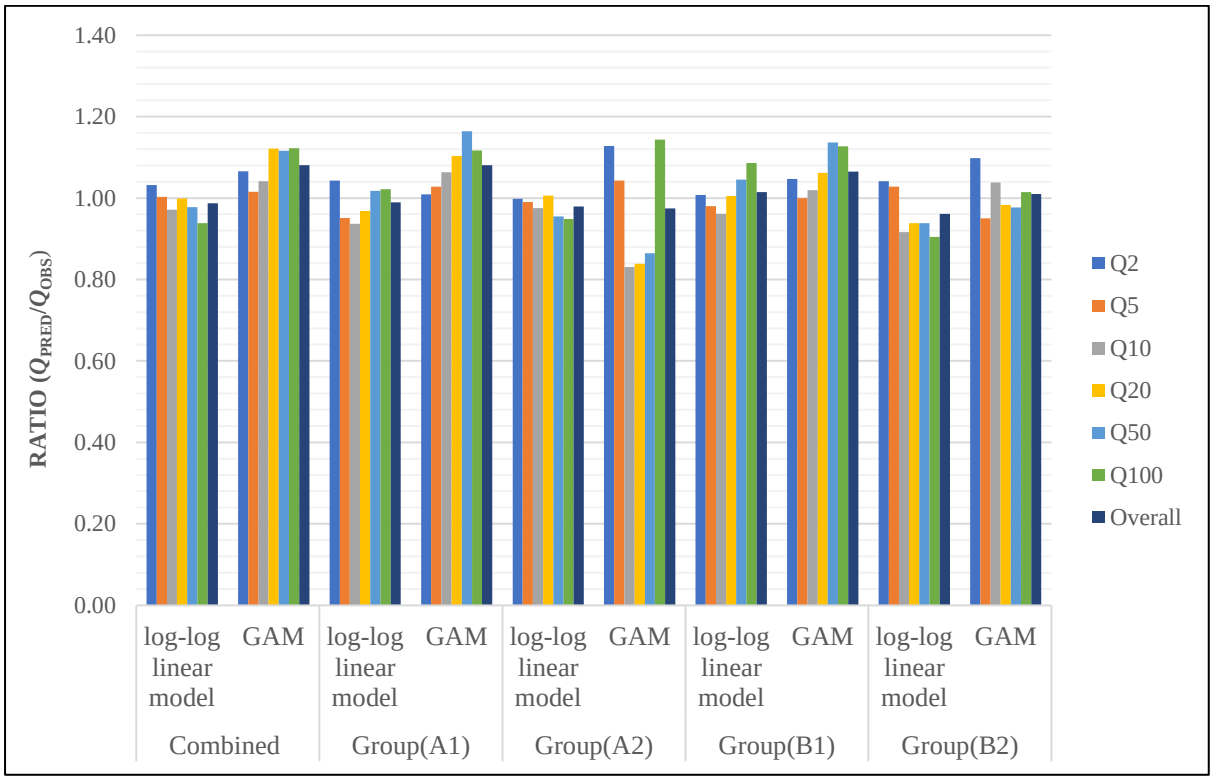

Figure 6.20 Plot of Median Q_{pred}/Q_{obs} Ratio values for the GAM and log-log linear model based RFFA model for multiple datasets

6.6. Comparison of this study with similar previous RFFA studies

Rahman et al. (2018) assessed the adequacy of the GAM models using 85 catchments from NSW. The R^2 values found for the GAM in the study of Rahman et al. (2017) exhibited better performance for 10 and 50 years of ARIs (they considered only these two ARIs) as compared to the current study for the Victorian catchments. Rahman et al. (2017) found R^2 values for the GAM models as 0.656 and 0.576 for 10 and 50 years ARIs, respectively. In this study, the R^2 values are found to be 0.62 and 0.50 for the GAM models for 10 and 50 years ARIs (for the combined data set), which are a little smaller than those of Rahman et al. (2018).

The median RE values in this study for the combined data set range from 18.73 % to 40.99 % for the log-log linear models and 33.88 % to 49.09 % for the GAM models. The ARR RFFE Model reported a median RE values in the range of 49 % to 59 %, which are much higher than those of this study (Rahman et al., 2016). It should be noted that a total 558 stations

from the east coast of Victoria, NSW and Queensland were developed to form Region 1 in ARR RFFE Model. The differences in RE values between this study and ARR RFFE Model are possibly due to different data sets, and the differences in the validation method. The ARR RFFE Model adopted a leave-one-out (LOO) validation approach, which is much more rigorous than the 10-fold cross validation technique adopted in this study.

It should be noted that the relative accuracy of the RFFE models in Australia is generally smaller than USA and European countries as Australian hydrology is more heterogeneous (Bloschl et al., 2013; Bates et al., 1998; Haddad and Rahman, 2012; Micevski et al., 2015).

6.7. Summary

In this chapter, five GAM based models from 5 groups of datasets are evaluated based on R^2, median RE and median Q_{pred}/Q_{obs} ratio values. It is found that there is no single GAM model which performs the best across all the six ARIs; however, clustering group A1 may be taken as the best performing group among all the five different GAM models.

Considering the R^2 values of both the GAM and log-log linear models, the log-log linear models from the combined group show overall higher values. However, for the clustering groups, the overall R^2 values are generally higher for the GAM models (i.e. for clustering groups A2, B1 and B2); for A1, both the models have the same R^2 value of 0.69. The GAM models with smaller ARIs (i.e., 2, 5 and 10 years) are found to outperform the log-log linear models in most cases. But for higher ARIs, log-log linear models perform better than the GAM models except for clustering group A2.

The overall median RE values are found to be quite similar or slightly higher for the GAM models considering all the five groups (i.e. one combined group and four clustering groups). For the combined group, the log-log linear model performs relatively better than the GAM model (i.e., RE of 31.15 % and 38.04 %, respectively). The median RE values of clustering group A2 for 20, 50 and 100 years ARIs are found to be relatively lower for the GAM models as compared with the log-log linear models.

The overall median Q_{pred}/Q_{obs} ratio values for the clustering groups A2 and B2 are found to be quite similar for the GAM and log-log linear models. In most cases, the overall median Q_{pred}/Q_{obs} ratio values from the GAM models are found to be slightly greater than 1, which

indicates towards the overestimation by this model. GAM models for clustering group A2 of 10, 20 and 50 years of ARIs give lower median Q_{pred}/Q_{obs} ratio values as compared to the log-log linear models. Moreover, the median Q_{pred}/Q_{obs} ratio values are found to be higher for most of the GAM models of 50 and 100 years of ARIs with an exception for clustering group A2. Overall, the median Q_{pred}/Q_{obs} ratio values indicate towards an overestimation tendency by the GAM models, in particular for higher ARIs.

CHAPTER 7

SUMMARY AND CONCLUSIONS

7.1. General

This thesis focuses on design flood estimation for ungauged catchments which is a common task in engineering planning and design. This thesis in particular, examines the applicability of a nonlinear technique in regional flood frequency analysis (RFFA) and Generalized Additive Models (GAM). It also compares the GAM based RFFA models with one of the most frequently adopted RFFA models: log-log linear regression method. In this regard, the development and testing of both the log-log linear and GAM based RFFA models are compared using a data set from Victoria, Australia. The selected dataset consisted of 114 small to medium sized catchments; this data was primarily compiled as a part of Australian Rainfall and Runoff (ARR) Project 5-Regional Flood Methods (Rahman et al., 2015; Rahman et al., 2016). A suite of statistical measures was used to assess the performances of the adopted RFFA models based on a 10-fold cross validation. This chapter presents a summary of the research works undertaken in this study, conclusions and recommendations for further studies to enhance the developed RFFA models.

7.2. Summary

7.2.1. Data selection

The State of Victoria in Australia has been selected for this study as it has the best flood data in Australia in terms of data quality, record length and geographical distributions of gauged catchments. A total of 114 small to medium sized gauged catchments are selected from Victoria. The data used for this study is obtained from Australian Rainfall Runoff Project 5 Regional Flood Methods. The geographical locations of the selected 114 catchments are presented in Figure 3.2. The selected catchments are mostly rural which are not subjected to any major regulation or land use changes during the period of streamflow data availability. The area of the selected catchments range from 3 to 997 km^2 (mean: 317.5 km^2 and median: 270.5 km^2). The annual maximum (AM) flood record lengths range from 26 years to 62 years (mean: 38 years and median: 39 years). At site flood quantiles for 6 different average recurrence intervals (ARIs) (i.e., 2, 5, 10, 20, 50 and 100 years) were estimated as a part of

ARR Project 5 (Rahman et al., 2015). These flood quantiles are used as target/dependent variables in the development of log-log linear and GAM based RFFA models. Data for eight catchment characteristics are selected as explanatory/predictor variables: *area, $I_{6,2}$, rain, evap*, SF, S1085, *sden* and *forest.* The summary of these catchment characteristics data are provided in Table 3.1.

7.2.2. Formation of regions

The data set of the selected 114 catchments are divided into five alternative groups: combined group (consisting of all the 114 catchments) and four clustering groups derived by cluster analysis on the selected predictor variables. Both hierarchical (based on Ward Manhattan method) and K-means clustering cluster analysis methods are adopted to form clustering groups.

7.2.3. Development of log-log linear model based RFFA technique

In order to develop the log-log linear regression model, both the dependent variables (i.e. flood quantiles) and independent variables (i.e. predictors) are log transformed. The prediction equations are developed using a backward stepwise procedure. The performances of the developed prediction equations are assessed based on three statistical measures/criteria: median Q_{pred}/Q_{obs} ratio, plot of Q_{obs} and Q_{pred}, absolute median relative error (RE). It is found that no individual model performs equally well across all the six ARIs with respect to all of the adopted criteria. Among all the developed log-log linear models, the one formed based on the clustering group A1 (consisting of 79 catchments) demonstrate the best performance.

7.2.4. Development of GAM based RFFA technique

For development of GAM, thin plate regression splines are adopted as they provide fast computation, and do not require a selection of knot locations and have optimality in approximating smoothness (Wood 2003, 2006). Backward stepwise procedure is utilized to select the most significant predictor variables. The predictor variables that are generally found to be statistically significant in the GAM models are: *area*, $I_{6,2}$, *rain* and *evap*. The statistical significance of each predictor variable is measured using the *p*-statistics. The performances of the developed models are assessed using three statistical criteria as mentioned earlier.

7.2.5. Comparison of log-log and GAM based RFFA models

Based on the R^2 values of both the GAM and log-log linear models, overall, the log-log linear models for the combined group show higher values. However, for the clustering groups, the overall R^2 values are generally higher for the GAM models (i.e. for clustering groups A2, B1 and B2); for A1, both the models have same R^2 value of 0.69. The GAM models with smaller ARIs (i.e., 2, 5 and 10 years) are found to outperform the log-log linear models in most of the cases; however, for the higher ARIs, log-log linear models perform better than the GAM models considering R^2 values except for clustering group A2.

The overall median RE values are found to be quite similar or slightly higher for the GAM models considering all the five groups. For the combined group, both the log-log linear and GAM models perform very similarly with respect to median RE (i.e., 31.15 % and 38.04 %, respectively). Although in most cases the median RE values are almost similar or slightly higher for GAM models, there is an exception for clustering group A2. The median RE values of clustering group A2 for 20, 50 and 100 years ARIs are found to be relatively lower for the GAM models as compared with the log-log linear models (which are 42.35 %, 49.59 % and 49.37 %, respectively). The median Q_{pred}/Q_{obs} ratio values for the clustering groups A2 and B2 are found to be quite similar for the GAM and log-log linear models. In the most cases, the overall median Q_{pred}/Q_{obs} ratio values from the GAM models are found to be slightly larger than 1.00, which indicates towards the overestimation of design floods by the GAM model, in particular for the higher ARIs. The RE values are found to be in the lowest range for clustering group A1, which are in the range of 29 % to 37 %, and 23 % to 59 %, respectively for the log-log linear and GAM based RFFA models.

7.3. Conclusions

This study develops and compares log-log linear and GAM based RFFA models for Victoria, Australia using data from 114 small to medium sized gauged catchments. The following conclusions can be drawn from this study:

- GAM can deal with non-linearity in RFFA better than the widely used log-log linear models, in particular for the smaller return periods (e.g. 2 to 10 years).
- It is found that none of the RFFA models examined in this study perform equally well across all the six ARIs with respect to all the adopted statistical measures/criteria.

- Based on overall average values of R^2, median RE and median Q_{pred}/Q_{obs} ratio, it is found that log-log linear models from clustering group A1 outperform the respective GAM models. However, for smaller ARIs (i.e., 2, 5, and 10 years), GAM based RFFA models perform almost similar or better than the log-log linear models. This is as expected, since for smaller floods (i.e. for smaller ARIs), catchments generally tend to behave more non-linearly, i.e. a higher loss values. For higher ARIs (e.g. 50 and 100 years), catchments behave more linearly, hence log-log linear regression models are expected to perform better, which is confirmed by this study.

- There are predictor variables, which were previously found (e.g. Haddad et al., 2012; Pilgrim et al., 1987; Rahman et al., 2015; Rahman et al., 2016) to be insignificant in RFFA, but are found statistically significant for the GAM models developed here. For example, *evap* is found statistically significant for most of the GAM models as opposed to previous RFFA studies in Australia.

- Overall, cluster analysis has not delivered superior groups in RFFA except for one case. Among all the five groups, the median RE values are found to be the lowest (29 % to 37 %) for the log-log linear models based on the clustering group A1 (consisting of 79 catchments); however, the other clustering groups perform poorly.

- It is found that *area*, $I_{6,2}$ and *rain* are the most significant predictor variables for the log-log linear models. For the GAM models, the most important predictor variables are *area*, $I_{6,2}$, *rain* and *evap*.

- Finally, it can be recommended that the users should apply the developed log-log linear models for estimating higher ARI design floods (20, 50 and 100 years ARI) and GAM model for smaller ARIs (2, 5 and 10 years) for Victoria.

7.4. Limitations of the study

The study has used only 114 catchments in Victoria, it would have been much better to include some more bigger sized catchments. Also, it would have been appropriate to use vary smaller sized catchments less than 1 km2 since RFFE model is widely used for very smaller catchments. In reality, there is no/little recorded streamflow data available for these smaller catchments, which is a major limitation for all the RFFE studies conducted in Australia including this study. Another limitation of the study is that, we have used only a limited set of

catchment characteristics (8 characteristics only) variables in RFFE model development, it would have been better to include other relevant catchment characteristics such as soil characteristics, stream order, base flow index and aridity index.

7.5. Recommendations for further research

The following studies are recommended for further enhancement of the RFFA models developed in this study:

- Develop and test both the log-log linear and GAM based RFFA models, using a greater number of predictor variables by extracting these data from GIS.
- Repeat the study for other Australian states to explore the viability of the GAM based RFFA modelling in Australia.
- Compare the log-log linear and GAM based RFFA models with Generalised Least Squares Regression (GLSR) based RFFA models, which are currently the recommended methods in Australian Rainfall and Rainfall – the national guide.
- Compare leave-one-out and 10-fold cross validation techniques for future RFFA studies using GAM.
- Assess the impacts of climate change on RFFA methods using GAM, as this can deal with the non-linearity in the rainfall-runoff-climate change issues more explicitly than the linear methods.

It is expected that the findings of this study and recommended future studies can provide enough scientific basis to replace the currently recommended RFFA techniques in the ARR, to enhance the overall accuracy and reliability of regional flood estimates in Australia, which currently sits in the range of 30 % to 60%.

REFERENCES

Abrahart, R. J., Heppenstall, A. J., & See, L. M. (2007). Timing error correction procedure applied to neural network rainfall-runoff modelling. *Hydrological Sciences Journal, 52*(3), 414-431. doi:10.1623/hysj.52.3.414

Acreman, M., & Sinclair, C. (1986). Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. *Journal of Hydrology, 84*(3-4), 365-380.

Adams, C. A. (1984). *Regional Flood Estimation for Ungauged Rural Catchments in Victoria*.

Asquith, W. H., Herrmann, G. R., & Cleveland, T. G. (2013). Generalized Additive Regression Models of Discharge and Mean Velocity Associated with Direct-Runoff Conditions in Texas: Utility of the U.S. Geological Survey Discharge Measurement Database. *Journal of Hydrologic Engineering, 18*(10), 1331-1348. doi:10.1061/(ASCE)HE.1943-5584.0000635

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. doi:10.1007/s00477-013-0771-5

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. doi:10.1007/s11069-015-1625-x

Ball, G. H., & Hall, D. J. (1965). *ISODATA, a novel method of data analysis and pattern classification*. Stanford research inst Menlo Park CA.

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

Bayentin, L., El Adlouni, S., Ouarda, T. B., Gosselin, P., Doyon, B., & Chebana, F. (2010). Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989-2006 in Quebec, Canada. *International journal of health geographics, 9*(1), 5.

Bertaccini, P., Dukic, V., & Ignaccolo, R. (2012). Modeling the short-term effect of traffic and meteorology on air pollution in turin with generalized additive models. *Advances in Meteorology, 2012*.

Burn, D. H. (1990). Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research, 26*(10), 2257-2265. doi:10.1029/WR026i010p02257

Burn, D. H., & Boorman, D. B. (1993). Estimation of hydrological parameters at ungauged catchments. *Journal of Hydrology, 143*(3-4), 429-454.

Caballero, W. L., & Rahman, A. (2014). Development of regionalised joint probability approach to flood estimation: a case study for Eastern New South Wales, Australia. *Hydrological Processes, 28*(13), 4001-4010.

Carbone, D., & Hanson, J. (2013). The worst floods in Australian history: Australian Geographic. Retrieved from http://www.australiangeographic.com.au/journal/the-worst-floods-in-australian-history.html

Chebana., F., Charron., C., Ouarda., T. B., & Martel., B. (2014). Regional frequency analysis at ungauged sites with the generalized additive model. *Journal of Hydrometeorology, 15*(6), 2418-2428.

Chow, V., Maidment, D., & Mays, L. (1988). Applied Hydrology. *McGraw-Hill Series in Water Resources and Environmental Engineering*.

Clifford, S., Choy, S. L., Hussein, T., Mengersen, K., & Morawska, L. (2011). Using the generalised additive model to model the particle number count of ultrafine particles. *Atmospheric Environment, 45*(32), 5934-5945.

Cohn, T. A., Delong, L. L., Gilroy, E. J., Hirsch, R. M., & Wells, D. K. (1989). Estimating constituent loads. *Water Resources Research, 25*(5), 937-942.

Dalrymple, T. (1960). *Flood-frequency analyses, manual of hydrology: Part 3*.

De Michele, C., & Rooso, R. (2002). A multi-level approach to flood frequency regionalisation. *Hydrology and Earth System Sciences, 6*(2), 185-194.

Di Prinzio, M., Castellarin, A., & Toth, E. (2011). Data-driven catchment classification: application to the pub problem. *Hydrology and Earth System Sciences, 15*(6), 1921.

Feldman, A. D. (1979). *Flood Hydrograph and Peak Flow Frequency Analysis*.

Flavell, D. J., Belstead, B. S., Chivers, B., & Walker, M. C. (1983). Runoff routing model parameters for catchments in Western Australia. In *Hydrology and Water Resources Symposium 1983: Preprints of Papers* (p. 22). Institution of Engineers, Australia.

Galiano, S. G. G., Gimenez, P. O., & Giraldo-Osorio, J. D. (2015). Assessing Nonstationary Spatial Patterns of Extreme Droughts from Long-Term High-Resolution Observational Dataset on a Semiarid Basin (Spain). *Water, 7*(10), 5458-5473.

Gentle, N., Kierce, S., & Nitz, A. (2001). Economic costs of natural disasters in Australia. *Australian Journal of Emergency Management, The, 16*(2), 38.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*(2), 215-223.

Guan, B. T., Hsu, H.-W., Wey, T.-H., & Tsao, L.-S. (2009). Modeling monthly mean temperatures for the mountain regions of Taiwan by generalized additive models. *agricultural and forest meteorology, 149*(2), 281-290.

Guénoche, A., Hansen, P., & Jaumard, B. (1991). Efficient algorithms for divisive hierarchical clustering with the diameter criterion. *Journal of classification, 8*(1), 5-30.

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*, 142-161.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association, 82*(398), 371-386.

Holder, R. (1985). *Multiple regression in hydrology*: Institute of hydrology.

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional Frequency Analysis An Approach Based on L-Moments*: Cambridge University Press.

Huang, Z., & Ng, M. K. (2003). A note on k-modes clustering. *Journal of Classification*, *20*(2), 257-261.

Hughes, J. M. R., & James, B. (1989). A hydrological regionalization of streams in Victoria, Australia, with implications for stream ecology. *Marine and Freshwater Research*, *40*(3), 303-326.

Ishak, E., Rahman, A., Westra, S., Sharma, A. and Kuczera, G. (2013). Evaluating the Non-stationarity of Australian Annual Maximum Floods. Journal of Hydrology, 494, 134-145.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data.

James, W., & Robinson, M. (1986). *Continuous deterministic urban runoff modelling.* Paper presented at the Urban Drainage Modeling, Proc., International Symp. on Comparison of Urban Drainage Models with Real Catchments Data, UDM'86.

Kalkstein, L. S., & Corrigan, P. (1986). A synoptic climatological approach for geographical analysis: assessment of sulfur dioxide concentrations. *Annals of the Association of American Geographers*, *76*(3), 381-395.

Kauermann, G., & Opsomer, J. D. (2003). Local likelihood estimation in generalized additive models. *Scandinavian Journal of Statistics*, *30*(2), 317-337.

Koch, R. W., & Smillie, G. M. (1986). Bias in hydrologic prediction using log-transformed regression models. *JAWRA Journal of the American Water Resources Association*, *22*(5), 717-723.

Leitte, A. M., Petrescu, C., Franck, U., Richter, M., Suciu, O., Ionovici, R., . . . Schlink, U. (2009). Respiratory health, effects of ambient air pollution and its modification by air humidity in Drobeta-Turnu Severin, Romania. *Science of the Total Environment*, *407*(13), 4004-4011.

Lumb, A. M., & James, L. D. (1976). Runoff files for flood hydrograph simulation. *Journal of the Hydraulics Division*, *102*(ASCE# 12499).

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

McCuen, R. H., Leahy, R. B., & Johnson, P. A. (1990). Problems with logarithmic transformations in regression. *ASCE, J. Hydraul. Engng, 116 (3)* pp. 414-428.

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.

Miller, D. M. (1984). Reducing transformation bias in curve fitting. *The American Statistician*, *38*(2), 124-126.

Morlini, I. (2006). On multicollinearity and concurvity in some nonlinear multivariate models. *Statistical Methods and Applications*, *15*(1), 3-26.

Morton, R., & Henderson, B. L. (2008). Estimation of nonlinear trends in water quality: an improved approach using generalized additive models. *Water Resources Research*, *44*(7).

Mosley, M.P. (1981): Delimitation of New Zealand hydrological regions. *Journal of Hydrology*., 49: 173-192.

Mulvaney, T. (1851). On the use of self-registering rain and flood gauges in making observations of the relations of rainfall and flood discharges in a given catchment. *Proceedings of the institution of Civil Engineers of Ireland, 4*(2), 18-33.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal, 26*(4), 354-359.

Nathan, R., & McMahon, T. (1990). Identification of homogeneous regions for the purposes of regionalisation. *Journal of Hydrology, 121*(1-4), 217-238.

Nelder, J. A., & Baker, R. J. (1972). *Generalized linear models*: Wiley Online Library.

Newton, D. W., & Herrin, J. C. (1982). *Assessment of commonly used methods of estimating flood frequency.* Washington DC,.

Ouali, D., Chebana, F., & Ouarda, T. B. M. J. (2016). Quantile Regression in Regional Frequency Analysis: A Better Exploitation of the Available Information. *Journal of Hydrometeorology, 17*(6), 1869-1883. doi:10.1175/jhm-d-15-0187.1

Ouarda, T. B., Charron, C., Marpu, P. R., & Chebana, F. (2016). The Generalized Additive Model for the Assessment of the Direct, Diffuse, and Global Solar Irradiances Using SEVIRI Images, With Application to the UAE. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9*(4), 1553-1566.

Palmen, L. B., & Weeks, W. D. (2011). Regional flood frequency for Queensland using the quantile regression technique. *Australian Journal of Water Resources, 15*(1), 47-58.

Pandey, G., & Nguyen, V.-T.-V. (1999). A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology, 225*(1-2), 92-101.

Phien, H. N., Huong, B. K., & Loi, P. D. (1990). Daily flow forecasting with regression analysis. *Water S. A., 16*(3), 179-184.

Pilgrim, D. H. (1982). Assessment of derived rural and urban runoff coefficients. *Institution of Engineers, Australia, Civil Engineering Transactions, 24*(3), 235-241.

Pilgrim, D. H., & McDermott, G. E. (1982). Design floods for small rural catchments in eastern new south wales. *Institution of Engineers, Australia, Civil Engineering Transactions, 24*(3), 226-234.

Pilgrim, D., & Canterford, R. (1987). *Australian rainfall and runoff*: Institution of Engineers, Australia.

Pilgrim, D. H., & Cordery, I. (1993). Flood runoff. *Handbook of hydrology, 9*, 1-42.

Pirozzi, J., Ashraf, M., Rahman, A., & Haddad, K. (2009). *Design flood estimation for ungauged catchments in Eastern NSW: evaluation of the probabilistic rational method.* Paper presented at the H2009: 32nd Hydrology and Water Resources Symposium, Newcastle: Adapting to Change.

Queensland Reconstruction Authority. (2011). Rebuilding a stronger, more resilient Queensland. Brisbane: Queensland Government.

Rahman, A. (1997). *Flood Estimation for ungauged catchments: A regional approach using flood and catchment characteristics* (Doctoral dissertation, PhD thesis, Department of Civil Engineering, Monash University.

Rahman, A., & Hollerbach, D. (2003). Study of Runoff Coefficients Associated with the probabilistic rational method for flood estimation in South-East Australia. Paper presented at the 28th International Hydrology and Water Resources Symposium: About Water; Symposium Proceedings.

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

Rahman, A., Haddad, K., Caballero, W., & Weinmarm, P. E. (2008). Progress on the enhancement of the Probabilistic Rational Method for Victoria in Australia. *31 Hydrology and Water Resources Symp.*940-951.

Rahman., A., Haddad., K., Zaman., M., Kuczera., G., & Weinmann., P. E. (2011). Design flood estimation in ungauged catchments: A comparison between the probabilistic rational method and quantile regression technique for NSW. *Australian Journal of Water Resources, 14*(2), 127-140.

Rahman, A., Haddad, K., Haque, M., Kuczera, G., & Weinmann, P. E. (2015). *Australian rainfall and runoff project 5: regional flood methods: stage 3 report* (No. P5/S3, p. 025). technical report.

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

Rahman, A. (2017). Social hydrology. Chapter 155, pp. 155-1 to 155-10, In: Handbook of Applied Hydrology, edited by Singh V P, McGraw-Hill.

Rahman., A., Charron., C., Ouarda., T. B., & Chebana., F. (2018). Development of regional flood frequency analysis techniques using generalized additive models for Australia. *Stochastic Environmental Research and Risk Assessment, 32*(1), 123-139.

Ralambondrainy, H. (1995). A conceptual version of the K-means algorithm. *Pattern Recognition Letters, 16*(11), 1147-1157.

Rao, A. R., & Srinivas, V. (2006). Regionalization of catchments by hybrid-cluster analysis. *Journal of Hydrology, 318*(1-4), 37-56.

Reis, D. S., Jr., Stedinger, J. R., & Martins, E. S. (2005). Bayesian GLS regression with application to LP3 regional skew estimation. *Water Resources Research, 41*(1).

Riggs, H. C. (1973). *Regional analyses of streamflow characteristics*: US Government Printing Office.

Roald, L. A. (1989). Application of regional flood frequency analysis to basins in northwest Europe. In *V: Roald, L.(ur.), Nordseth, K.(ur.), Anker Hassel, K.(ur.). V: FRIENDS in Hydrology: proceedings of an international conference at Bolkesjø, Norway*.

Savaresi, S. M., Boley, D. L., Bittanti, S., & Gazzaniga, G. (2002, April). Cluster selection in divisive clustering algorithms. In *Proceedings of the 2002 SIAM International Conference on Data Mining* (pp. 299-314). Society for Industrial and Applied Mathematics.

Schindeler, S., Muscatello, D., Ferson, M., Rogers, K., Grant, P., & Churches, T. (2009). Evaluation of alternative respiratory syndromes for specific syndromic surveillance of influenza and respiratory syncytial virus: a time series analysis. *BMC Infectious Diseases, 9*(190).

Shortridge, J., Guikema, S., & Zaitchik, B. (2015). Empirical streamflow simulation for water resource management in data-scarce seasonal catchments. *Hydrology & Earth System Sciences Discussions, 12*(10).

Shu, C., & Ouarda, T. B. M. J. (2007). Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. *Water Resources Research, 43*(7). doi:10.1029/2006WR005142

Smith, J. A. (1989). Regional flood frequency analysis using extreme order statistics of the annual peak record. *Water Resources Research, 25*(2), 311-317. doi:10.1029/WR025i002p00311

Stedinger, J. R. (1983). Estimating a regional flood frequency distribution. *Water Resources Research, 19*(2), 503-510. doi:10.1029/WR019i002p00503

Stedinger, J. R., & Tasker, G. D. (1985). Regional Hydrologic Analysis: 1. Ordinary, Weighted, and Generalized Least Squares Compared. *Water Resources Research, 21*(9), 1421-1432. doi:10.1029/WR021i009p01421

Srinivas, V., Rao, A. R., Tripathi, S., & Govindaraju, R. S. (2007). *Regional Flood Frequency Analysis using Two-level Clustering Approach.* Paper presented at the World Environmental and Water Resources Congress 2007: Restoring Our Natural Habitat.

Tasker, G. D. (1980). Hydrologic regression with weighted least squares. *Water Resources Research, 16*(6), 1107-1113. doi:10.1029/WR016i006p01107

Tasker, G.D. (1982b): Comparing methods of hydrological regionalisation. Water Resources Bulletin., 18(6): 965-970.

Tasker, G. D., & Stedinger, J. R. (1989). An operational GLS model for hydrologic regression. *Journal of Hydrology, 111*(1-4), 361-375. doi:10.1016/0022-1694(89)90268-0

Thomas, D. M., & Benson, M. A. (1970). *Generalization of streamflow characteristics from drainage-basin characteristics*. Washington, DC: US Government Printing Office.

Tisseuil, C., Vrac, M., Lek, S., & Wade, A. J. (2010). Statistical downscaling of river flows. *Journal of Hydrology, 385*(1-4), 279-291. doi:10.1016/j.jhydrol.2010.02.030

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2).

Vieira, V., Webster, T., Weinberg, J., & Aschengrau, A. (2009). Spatial analysis of bladder, kidney, and pancreatic cancer on upper Cape Cod: an application of generalized additive models to case-control data. *Environmental Health, 8*(1), 3.

Wang, Y., Li, J., Feng, P., & Hu, R. (2015). A time-dependent drought index for non-stationary precipitation series. *Water Resources Management, 29*(15), 5631-5647.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-244.

Wazneh, H., Chebana, F., & Ouarda, T. B. (2013). Depth-based regional index-flood model. *Water Resources Research, 49*(12), 7957-7972.

Wen, L., Rogers, K., Ling, J., & Saintilan, N. (2011). The impacts of river regulation and water diversion on the hydrological drought characteristics in the Lower Murrumbidgee River, Australia. *Journal of Hydrology, 405*(3-4), 382-391. doi:10.1016/j.jhydrol.2011.05.037

Willmott, C. J., & Vernon, M. T. (1980). Solar climates of the conterminous United States: A preliminary investigation. *Solar Energy*, *24*(3), 295-303.

Winkler, J. (1985). *Regionalization of the diurnal distribution of summertime heavy precipitation.* Paper presented at the Preprints, Sixth Conference of Hydrometeorology, American Meteorological Society.

Wood, S. N., & Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling, 157*(2-3), 157-177.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(1), 95-114.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*: Chapman and Hall/CRC.

APPENDIX A

List of Study Catchments

Table A. 1 Study Catchments of Combined group

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
221207	Errinundra	Errinundra	158	40	1971 - 2010	-37.45	148.91
221209	Weeragua	Cann(East Branch	154	39	1973 - 2011	-37.37	149.2
221210	The Gorge	Genoa	837	40	1972 - 2011	-37.43	149.53
221211	Combienbar	Combienbar	179	38	1974 - 2011	-37.44	148.98
221212	Princes HWY	Bemm	725	37	1975 - 2011	-37.61	148.9
222202	Sardine Ck	Brodribb	650	47	1965 - 2011	-37.51	148.55
222206	Buchan	Buchan	822	38	1974 - 2011	-37.5	148.18
222210	Deddick (Caseys)	Deddick	857	42	1970 - 2011	-37.09	148.43
222213	Suggan Buggan	Suggan Buggan	357	41	1971 - 2011	-36.95	148.33
222217	Jacksons Crossing	Rodger	447	36	1976 - 2011	-37.41	148.36
223202	Swifts Ck	Tambo	943	38	1974 - 2011	-37.26	147.72
223204	Deptford	Nicholson	287	38	1974 - 2011	-37.6	147.7
224213	Lower Dargo Rd	Dargo	676	39	1973 - 2011	-37.5	147.27
224214	Tabberabbera	Wentworth	443	38	1974 - 2011	-37.5	147.39
225213	Beardmore	Aberfeldy	311	33	1973 - 2005	-37.85	146.43
225218	Briagalong	Freestone Ck	309	41	1971 - 2011	-37.81	147.09
225219	Glencairn	Macalister	570	45	1967 - 2011	-37.52	146.57
225223	Gillio Rd	Valencia Ck	195	41	1971 - 2011	-37.73	146.98
225224	The Channel	Avon	554	40	1972 - 2011	-37.8	146.88
226204	Willow Grove	Latrobe	580	41	1971 - 2011	-38.09	146.16
226209	Darnum	Moe	214	40	1972 - 2011	-38.21	146
226222	Near Noojee (U/S Ada R Jun	Latrobe	62	41	1971 - 2011	-37.88	145.89
226226	Tanjil Junction	Tanjil	289	52	1960 - 2011	-38.01	146.2

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
226402	Trafalgar East	Moe Drain	622	37	1975 - 2011	-38.18	146.21
227200	Yarram	Tarra	25	47	1965 - 2011	-38.46	146.69
227205	Calignee South	Merriman Ck	36	37	1975 - 2011	-38.36	146.65
227210	Carrajung Lower	Bruthen Ck	18	39	1973 - 2011	-38.4	146.74
227211	Toora	Agnes	67	38	1974 - 2011	-38.64	146.37
227213	Jack	Jack	34	42	1970 - 2011	-38.53	146.53
227219	Loch	Bass	52	39	1973 - 2011	-38.38	145.56
227225	Fischers	Tarra	16	40	1973 - 2012	-38.47	146.56
227226	Dumbalk North	Tarwineast Branc	127	42	1970 - 2011	-38.5	146.16
227231	Glen Forbes South	Bass	233	37	1974 - 2010	-38.47	145.51
227236	D/S Foster Ck Jun	Powlett	228	33	1979 - 2011	-38.56	145.71
228217	Pakenham	Toomuc Ck	41	29	1974 - 2002	-38.07	145.46
229218	Watsons Ck	Watsons Ck	36	26	1974 - 1999	-37.67	145.26
230204	Riddells Ck	Riddells Ck	79	38	1974 - 2011	-37.47	144.67
230205	Bulla (D/S of Emu Ck Jun)	Deep Ck	865	38	1974 - 2011	-37.63	144.8
230211	Clarkefield	Emu Ck	93	36	1975 - 2010	-37.47	144.75
230213	Mount Macedon	Turritable Ck	15	38	1975- 2012	-37.42	144.58
231213	Sardine Ck- O'Brien Cro	Lerderderg Ck	153	53	1959 - 2011	-37.5	144.36
231231	Melton South	Toolern Ck	95	32	1979 - 2010	-37.91	144.58
232213	U/S of Bungal Dam	Lal Lal Ck	157	33	1977 - 2009	-37.66	144.03
233214	Forrest (above Tunnel)	Barwoneast Branc	17	34	1978 - 2011	-38.53	143.73
234200	Pitfield	Woady Yaloak	324	40	1972 - 2011	-37.81	143.59
235202	Upper Gellibrand	Gellibrand	53	37	1975 - 2011	-37.56	143.64
235203	Curdie	Curdies	790	37	1975 - 2011	-38.45	142.96
235204	Beech Forest	Little Aire Ck	11	36	1976 - 2011	-38.66	143.53

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
235205	Wyelangta	Arkins Ck West B	3	34	1978 - 2011	-38.65	143.44
235227	Bunkers Hill	Gellibrand	311	38	1974 - 2011	-38.53	143.48
235233	Apollo Bay-Paradise	Barhameast Branc	43	35	1977 - 2011	-38.76	143.62
235234	Gellibrand	Love Ck	75	33	1979 - 2011	-38.49	143.57
236205	Woodford	Merri	899	38	1974 - 2011	-38.32	142.48
236212	Cudgee	Brucknell Ck	570	37	1975 - 2011	-38.35	142.65
237207	Heathmere	Surry	310	37	1975 - 2011	-38.25	141.66
238207	Jimmy Ck	Wannon	40	38	1974 - 2011	-37.37	142.5
238219	Morgiana	Grange Burn	997	39	1973 - 2011	-37.71	141.83
401208	Berringama	Cudgewa Ck	350	47	1965 - 2011	-36.21	147.68
401209	Omeo	Livingstone Ck	243	27	1968 - 1994	-37.11	147.57
401210	below Granite Flat	Snowy Ck	407	44	1968 - 2011	-36.57	147.41
401212	Upper Nariel	Nariel Ck	252	58	1954 - 2011	-36.45	147.83
401216	Jokers Ck	Big	356	60	1952 - 2011	-36.95	141.47
401217	Gibbo Park	Gibbo	389	41	1971 - 2011	-36.75	147.71
401220	McCallums	Tallangatta Ck	464	36	1976 - 2011	-36.21	147.5
402203	Mongans Br	Kiewa	552	42	1970 - 2011	-36.6	147.1
402204	Osbornes Flat	Yackandandah Ck	255	45	1967 - 2011	-36.31	146.9
402206	Running Ck	Running Ck	126	37	1975 - 2011	-36.54	147.05
402217	Myrtleford Rd Br	Flaggy Ck	24	41	1970 - 2010	-36.39	146.88
403205	Bright	Ovens Rivers	495	41	1971 - 2011	-36.73	146.95
403209	Wangaratta North	Reedy Ck	368	39	1973 - 2011	-36.33	146.34
403213	Greta South	Fifteen Mile Ck	229	39	1973 - 2011	-36.62	146.24
403221	Woolshed	Reedy Ck	214	37	1975 - 2011	-36.31	146.6
403222	Abbeyard	Buffalo	425	39	1973 - 2011	-36.91	146.7

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
403233	Harris Lane	Buckland	435	40	1972 - 2011	-36.72	146.88
404207	Kelfeera	Holland Ck	451	37	1975 - 2011	-36.61	146.06
405205	Murrindindi above Colwells	Murrindindi	108	37	1975 - 2011	-37.41	145.56
405209	Taggerty	Acheron	619	39	1973 - 2011	-37.32	145.71
405212	Tallarook	Sunday Ck	337	37	1975 - 2011	-37.1	145.05
405214	Tonga Br	Delatite	368	55	1957 - 2011	-37.15	146.13
405215	Glen Esk	Howqua	368	38	1974 - 2011	-37.23	146.21
405217	Devlins Br	Yea	360	37	1975 - 2011	-37.38	145.48
405218	Gerrang Br	Jamieson	368	53	1959 - 2011	-37.29	146.19
405226	Moorilim	Pranjip Ck	787	38	1974 - 2011	-36.62	145.31
405227	Jamieson	Big Ck	619	42	1970 - 2011	-37.37	146.06
405229	Wanalta	Wanalta Ck	108	43	1969 - 2011	-36.64	144.87
405230	Colbinabbin	Cornella Ck	259	39	1973 - 2011	-36.61	144.8
405231	Flowerdale	King Parrot Ck	181	38	1974 - 2011	-37.35	145.29
405237	Euroa Township	Seven Creeks	332	39	1973 - 2011	-36.76	145.58
405240	Ash Br	Sugarloaf Ck	609	39	1973 - 2011	-37.06	145.05
405241	Rubicon	Rubicon	129	39	1973 - 2011	-37.29	145.83
405245	Mansfield	Ford Ck	115	42	1970 - 2011	-37.04	146.05
405248	Graytown	Major Ck	282	41	1971 - 2011	-36.86	144.91
405251	Ancona	Brankeet Ck	121	39	1973 - 2011	-36.97	145.78
405264	D/S of Frenchman Ck Jun	Big	333	37	1975 - 2011	-37.52	146.08
405274	Yarck	Home Ck	187	35	1977 - 2011	-37.11	145.6
406213	Redesdale	Campaspe	629	37	1975 - 2011	-37.02	144.54
406214	Longlea	Axe Ck	234	40	1972 - 2011	-36.78	144.43

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
406216	Sedgewick	Axe Ck	34	37	1975 - 2011	-36.9	144.36
406224	Runnymede	Mount Pleasant C	248	37	1975 - 2011	-36.55	144.64
406226	Derrinal	Mount Ida Ck	174	34	1978 - 2011	-36.88	144.65
407214	Clunes	Creswick Ck	308	37	1975 - 2011	-37.3	143.79
407217	Vaughan atD/S Fryers Ck	Loddon	299	44	1968 - 2011	-37.16	144.21
407220	Norwood	Bet Bet Ck	347	38	1973 - 2010	-37	143.64
407221	Yandoit	Jim Crow Ck	166	39	1973 - 2011	-37.21	144.1
407222	Clunes	Tullaroop Ck	632	39	1973 - 2011	-37.23	143.83
407230	Strathlea	Joyces Ck	153	39	1973 - 2011	-37.17	143.96
407246	Marong	Bullock Ck	184	39	1973 - 2011	-36.73	144.13
407253	Minto	Piccaninny Ck	668	39	1973 - 2011	-36.45	144.47
415207	Eversley	Wimmera	304	37	1975 - 2011	-37.19	143.19
415217	Grampians Rd Br	Fyans Ck	34	38	1973 - 2010	-37.26	142.53
415220	Wimmera HWY	Avon	596	37	1974 - 2010	-36.64	142.98
415226	Carrs Plains	Richardson	130	31	1971 - 2001	-36.75	142.79
415237	Stawell	Concongella Ck	239	35	1977 - 2011	-37.02	142.82
415238	Navarre	Wattle Ck	141	36	1976 - 2011	-36.9	143.1

Table A. 2 Study Catchments of Clustering group A1

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
221207	Errinundra	Errinundra	158	40	1971 - 2010	-37.45	148.91
221209	Weeragua	Cann(East Branch	154	39	1973 - 2011	-37.37	149.20
221210	The Gorge	Genoa	837	40	1972 - 2011	-37.43	149.53
221211	Combienbar	Combienbar	179	38	1974 - 2011	-37.44	148.98

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
221212	Princes HWY	Bemm	725	37	1975 - 2011	-37.61	148.90
222202	Sardine Ck	Brodribb	650	47	1965 - 2011	-37.51	148.55
222206	Buchan	Buchan	822	38	1974 - 2011	-37.50	148.18
222210	Deddick (Caseys)	Deddick	857	42	1970 - 2011	-37.09	148.43
222213	Suggan Buggan	Suggan Buggan	357	41	1971 - 2011	-36.95	148.33
222217	Jacksons Crossing	Rodger	447	36	1976 - 2011	-37.41	148.36
223202	Swifts Ck	Tambo	943	38	1974 - 2011	-37.26	147.72
223204	Deptford	Nicholson	287	38	1974 - 2011	-37.60	147.70
224213	Lower Dargo Rd	Dargo	676	39	1973 - 2011	-37.50	147.27
224214	Tabberabbera	Wentworth	443	38	1974 - 2011	-37.50	147.39
225213	Beardmore	Aberfeldy	311	33	1973 - 2005	-37.85	146.43
225218	Briagalong	Freestone Ck	309	41	1971 - 2011	-37.81	147.09
225219	Glencairn	Macalister	570	45	1967 - 2011	-37.52	146.57
225223	Gillio Rd	Valencia Ck	195	41	1971 - 2011	-37.73	146.98
225224	The Channel	Avon	554	40	1972 - 2011	-37.80	146.88
226204	Willow Grove	Latrobe	580	41	1971 - 2011	-38.09	146.16
226222	Near Noojee (U/S Ada R Jun	Latrobe	62	41	1971 - 2011	-37.88	145.89
226226	Tanjil Junction	Tanjil	289	52	1960 - 2011	-38.01	146.20
227200	Yarram	Tarra	25	47	1965 - 2011	-38.46	146.69
227205	Calignee South	Merriman Ck	36	37	1975 - 2011	-38.36	146.65
227210	Carrajung Lower	Bruthen Ck	18	39	1973 - 2011	-38.40	146.74
227211	Toora	Agnes	67	38	1974 - 2011	-38.64	146.37
227213	Jack	Jack	34	42	1970 - 2011	-38.53	146.53
227225	Fischers	Tarra	16	40	1973 - 2012	-38.47	146.56
227226	Dumbalk North	Tarwineast Branc	127	42	1970 - 2011	-38.50	146.16

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
228217	Pakenham	Toomuc Ck	41	29	1974 - 2002	-38.07	145.46
229218	Watsons Ck	Watsons Ck	36	26	1974 - 1999	-37.67	145.26
230204	Riddells Ck	Riddells Ck	79	38	1974 - 2011	-37.47	144.67
230211	Clarkefield	Emu Ck	93	36	1975 - 2010	-37.47	144.75
230213	Mount Macedon	Turritable Ck	15	38	1975- 2012	-37.42	144.58
231213	Sardine Ck-O'Brien Cro	Lerderderg Ck	153	53	1959 - 2011	-37.50	144.36
233214	Forrest (above Tunnel)	Barwoneast Branc	17	34	1978 - 2011	-38.53	143.73
234200	Pitfield	Woady Yaloak	324	40	1972 - 2011	-37.81	143.59
235202	Upper Gellibrand	Gellibrand	53	37	1975 - 2011	-37.56	143.64
235204	Beech Forest	Little Aire Ck	11	36	1976 - 2011	-38.66	143.53
235205	Wyelangta	Arkins Ck West B	3	34	1978 - 2011	-38.65	143.44
235227	Bunkers Hill	Gellibrand	311	38	1974 - 2011	-38.53	143.48
235233	Apollo Bay-Paradise	Barhameast Branc	43	35	1977 - 2011	-38.76	143.62
235234	Gellibrand	Love Ck	75	33	1979 - 2011	-38.49	143.57
238207	Jimmy Ck	Wannon	40	38	1974 - 2011	-37.37	142.50
401208	Berringama	Cudgewa Ck	350	47	1965 - 2011	-36.21	147.68
401209	Omeo	Livingstone Ck	243	27	1968 - 1994	-37.11	147.57
401210	below Granite Flat	Snowy Ck	407	44	1968 - 2011	-36.57	147.41
401212	Upper Nariel	Nariel Ck	252	58	1954 - 2011	-36.45	147.83
401216	Jokers Ck	Big	356	60	1952 - 2011	-36.95	141.47
401217	Gibbo Park	Gibbo	389	41	1971 - 2011	-36.75	147.71
401220	McCallums	Tallangatta Ck	464	36	1976 - 2011	-36.21	147.50
402203	Mongans Br	Kiewa	552	42	1970 - 2011	-36.60	147.10
402204	Osbornes Flat	Yackandandah Ck	255	45	1967 - 2011	-36.31	146.90
402206	Running Ck	Running Ck	126	37	1975 - 2011	-36.54	147.05

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
402217	Myrtleford Rd Br	Flaggy Ck	24	41	1970 - 2010	-36.39	146.88
403205	Bright	Ovens Rivers	495	41	1971 - 2011	-36.73	146.95
403213	Greta South	Fifteen Mile Ck	229	39	1973 - 2011	-36.62	146.24
403222	Abbeyard	Buffalo	425	39	1973 - 2011	-36.91	146.70
403233	Harris Lane	Buckland	435	40	1972 - 2011	-36.72	146.88
404207	Kelfeera	Holland Ck	451	37	1975 - 2011	-36.61	146.06
405205	Murrindindi above Colwells	Murrindindi	108	37	1975 - 2011	-37.41	145.56
405209	Taggerty	Acheron	619	39	1973 - 2011	-37.32	145.71
405212	Tallarook	Sunday Ck	337	37	1975 - 2011	-37.10	145.05
405214	Tonga Br	Delatite	368	55	1957 - 2011	-37.15	146.13
405215	Glen Esk	Howqua	368	38	1974 - 2011	-37.23	146.21
405217	Devlins Br	Yea	360	37	1975 - 2011	-37.38	145.48
405218	Gerrang Br	Jamieson	368	53	1959 - 2011	-37.29	146.19
405227	Jamieson	Big Ck	619	42	1970 - 2011	-37.37	146.06
405231	Flowerdale	King Parrot Ck	181	38	1974 - 2011	-37.35	145.29
405237	Euroa Township	Seven Creeks	332	39	1973 - 2011	-36.76	145.58
405241	Rubicon	Rubicon	129	39	1973 - 2011	-37.29	145.83
405245	Mansfield	Ford Ck	115	42	1970 - 2011	-37.04	146.05
405251	Ancona	Brankeet Ck	121	39	1973 - 2011	-36.97	145.78
405264	D/S of Frenchman Ck Jun	Big	333	37	1975 - 2011	-37.52	146.08
405274	Yarck	Home Ck	187	35	1977 - 2011	-37.11	145.60
407217	Vaughan atD/S Fryers Ck	Loddon	299	44	1968 - 2011	-37.16	144.21
407221	Yandoit	Jim Crow Ck	166	39	1973 - 2011	-37.21	144.10
415217	Grampians Rd Br	Fyans Ck	34	38	1973 - 2010	-37.26	142.53

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
415238	Navarre	Wattle Ck	141	36	1976 - 2011	-36.90	143.10

Table A. 3 Study Catchments of Clustering group A2

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
226209	Darnum	Moe	214	40	1972 - 2011	-38.21	146.00
226402	Trafalgar East	Moe Drain	622	37	1975 - 2011	-38.18	146.21
227219	Loch	Bass	52	39	1973 - 2011	-38.38	145.56
227231	Glen Forbes South	Bass	233	37	1974 - 2010	-38.47	145.51
227236	D/S Foster Ck Jun	Powlett	228	33	1979 - 2011	-38.56	145.71
230205	Bulla (D/S of Emu Ck Jun)	Deep Ck	865	38	1974 - 2011	-37.63	144.80
231231	Melton South	Toolern Ck	95	32	1979 - 2010	-37.91	144.58
232213	U/S of Bungal Dam	Lal Lal Ck	157	33	1977 - 2009	-37.66	144.03
235203	Curdie	Curdies	790	37	1975 - 2011	-38.45	142.96
236205	Woodford	Merri	899	38	1974 - 2011	-38.32	142.48
236212	Cudgee	Brucknell Ck	570	37	1975 - 2011	-38.35	142.65
237207	Heathmere	Surry	310	37	1975 - 2011	-38.25	141.66
238219	Morgiana	Grange Burn	997	39	1973 - 2011	-37.71	141.83
403209	Wangaratta North	Reedy Ck	368	39	1973 - 2011	-36.33	146.34
403221	Woolshed	Reedy Ck	214	37	1975 - 2011	-36.31	146.60
405226	Moorilim	Pranjip Ck	787	38	1974 - 2011	-36.62	145.31
405229	Wanalta	Wanalta Ck	108	43	1969 - 2011	-36.64	144.87
405230	Colbinabbin	Cornella Ck	259	39	1973 - 2011	-36.61	144.80
405240	Ash Br	Sugarloaf Ck	609	39	1973 - 2011	-37.06	145.05
405248	Graytown	Major Ck	282	41	1971 - 2011	-36.86	144.91
406213	Redesdale	Campaspe	629	37	1975 - 2011	-37.02	144.54
406214	Longlea	Axe Ck	234	40	1972 - 2011	-36.78	144.43
406216	Sedgewick	Axe Ck	34	37	1975 - 2011	-36.90	144.36
406224	Runnymede	Mount Pleasant C	248	37	1975 - 2011	-36.55	144.64

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
406226	Derrinal	Mount Ida Ck	174	34	1978 - 2011	-36.88	144.65
407214	Clunes	Creswick Ck	308	37	1975 - 2011	-37.30	143.79
407220	Norwood	Bet Bet Ck	347	38	1973 - 2010	-37.00	143.64
407222	Clunes	Tullaroop Ck	632	39	1973 - 2011	-37.23	143.83
407230	Strathlea	Joyces Ck	153	39	1973 - 2011	-37.17	143.96
407246	Marong	Bullock Ck	184	39	1973 - 2011	-36.73	144.13
407253	Minto	Piccaninny Ck	668	39	1973 - 2011	-36.45	144.47
415207	Eversley	Wimmera	304	37	1975 - 2011	-37.19	143.19
415220	Wimmera HWY	Avon	596	37	1974 - 2010	-36.64	142.98
415226	Carrs Plains	Richardson	130	31	1971 - 2001	-36.75	142.79
415237	Stawell	Concongella Ck	239	35	1977 - 2011	-37.02	142.82

Table A. 4 Study Catchments of Clustering group B1

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
221207	Errinundra	Errinundra	158	40	1971 - 2010	-37.45	148.91
221209	Weeragua	Cann(East Branch	154	39	1973 - 2011	-37.37	149.2
221210	The Gorge	Genoa	837	40	1972 - 2011	-37.43	149.53
221211	Combienbar	Combienbar	179	38	1974 - 2011	-37.44	148.98
221212	Princes HWY	Bemm	725	37	1975 - 2011	-37.61	148.9
222202	Sardine Ck	Brodribb	650	47	1965 - 2011	-37.51	148.55
222210	Deddick (Caseys)	Deddick	857	42	1970 - 2011	-37.09	148.43
222213	Suggan Buggan	Suggan Buggan	357	41	1971 - 2011	-36.95	148.33
222217	Jacksons Crossing	Rodger	447	36	1976 - 2011	-37.41	148.36
223202	Swifts Ck	Tambo	943	38	1974 - 2011	-37.26	147.72
223204	Deptford	Nicholson	287	38	1974 - 2011	-37.6	147.7
224213	Lower Dargo Rd	Dargo	676	39	1973 - 2011	-37.5	147.27
224214	Tabberabbera	Wentworth	443	38	1974 - 2011	-37.5	147.39
225213	Beardmore	Aberfeldy	311	33	1973 - 2005	-37.85	146.43
225218	Briagalong	Freestone Ck	309	41	1971 - 2011	-37.81	147.09
225219	Glencairn	Macalister	570	45	1967 - 2011	-37.52	146.57
225223	Gillio Rd	Valencia Ck	195	41	1971 - 2011	-37.73	146.98
225224	The Channel	Avon	554	40	1972 - 2011	-37.8	146.88
226204	Willow Grove	Latrobe	580	41	1971 - 2011	-38.09	146.16
226222	Near Noojee (U/S Ada R Jun	Latrobe	62	41	1971 - 2011	-37.88	145.89
226226	Tanjil Junction	Tanjil	289	52	1960 - 2011	-38.01	146.2
227200	Yarram	Tarra	25	47	1965 - 2011	-38.46	146.69
227205	Calignee South	Merriman Ck	36	37	1975 - 2011	-38.36	146.65
227210	Carrajung Lower	Bruthen Ck	18	39	1973 - 2011	-38.4	146.74

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
227211	Toora	Agnes	67	38	1974 - 2011	-38.64	146.37
227213	Jack	Jack	34	42	1970 - 2011	-38.53	146.53
227225	Fischers	Tarra	16	40	1973 - 2012	-38.47	146.56
228217	Pakenham	Toomuc Ck	41	29	1974 - 2002	-38.07	145.46
229218	Watsons Ck	Watsons Ck	36	26	1974 - 1999	-37.67	145.26
230213	Mount Macedon	Turritable Ck	15	38	1975- 2012	-37.42	144.58
231213	Sardine Ck-O'Brien Cro	Lerderderg Ck	153	53	1959 - 2011	-37.5	144.36
233214	Forrest (above Tunnel)	Barwoneast Branc	17	34	1978 - 2011	-38.53	143.73
235202	Upper Gellibrand	Gellibrand	53	37	1975 - 2011	-37.56	143.64
235204	Beech Forest	Little Aire Ck	11	36	1976 - 2011	-38.66	143.53
235205	Wyelangta	Arkins Ck West B	3	34	1978 - 2011	-38.65	143.44
235227	Bunkers Hill	Gellibrand	311	38	1974 - 2011	-38.53	143.48
235233	Apollo Bay-Paradise	Barhameast Branc	43	35	1977 - 2011	-38.76	143.62
235234	Gellibrand	Love Ck	75	33	1979 - 2011	-38.49	143.57
238207	Jimmy Ck	Wannon	40	38	1974 - 2011	-37.37	142.5
401208	Berringama	Cudgewa Ck	350	47	1965 - 2011	-36.21	147.68
401209	Omeo	Livingstone Ck	243	27	1968 - 1994	-37.11	147.57
401210	below Granite Flat	Snowy Ck	407	44	1968 - 2011	-36.57	147.41
401212	Upper Nariel	Nariel Ck	252	58	1954 - 2011	-36.45	147.83
401217	Gibbo Park	Gibbo	389	41	1971 - 2011	-36.75	147.71
401220	McCallums	Tallangatta Ck	464	36	1976 - 2011	-36.21	147.5
402203	Mongans Br	Kiewa	552	42	1970 - 2011	-36.6	147.1
402204	Osbornes Flat	Yackandandah Ck	255	45	1967 - 2011	-36.31	146.9
402206	Running Ck	Running Ck	126	37	1975 - 2011	-36.54	147.05
402217	Myrtleford Rd Br	Flaggy Ck	24	41	1970 - 2010	-36.39	146.88

Station ID	Station Name	River Name	Catchment Area (km^2)	Record Length (years)	Period of Record	Lat	Lon
403205	Bright	Ovens Rivers	495	41	1971 - 2011	-36.73	146.95
403213	Greta South	Fifteen Mile Ck	229	39	1973 - 2011	-36.62	146.24
403222	Abbeyard	Buffalo	425	39	1973 - 2011	-36.91	146.7
403233	Harris Lane	Buckland	435	40	1972 - 2011	-36.72	146.88
404207	Kelfeera	Holland Ck	451	37	1975 - 2011	-36.61	146.06
405205	Murrindindi above Colwells	Murrindindi	108	37	1975 - 2011	-37.41	145.56
405209	Taggerty	Acheron	619	39	1973 - 2011	-37.32	145.71
405214	Tonga Br	Delatite	368	55	1957 - 2011	-37.15	146.13
405215	Glen Esk	Howqua	368	38	1974 - 2011	-37.23	146.21
405217	Devlins Br	Yea	360	37	1975 - 2011	-37.38	145.48
405218	Gerrang Br	Jamieson	368	53	1959 - 2011	-37.29	146.19
405227	Jamieson	Big Ck	619	42	1970 - 2011	-37.37	146.06
405231	Flowerdale	King Parrot Ck	181	38	1974 - 2011	-37.35	145.29
405237	Euroa Township	Seven Creeks	332	39	1973 - 2011	-36.76	145.58
405241	Rubicon	Rubicon	129	39	1973 - 2011	-37.29	145.83
405251	Ancona	Brankeet Ck	121	39	1973 - 2011	-36.97	145.78
405264	D/S of Frenchman Ck Jun	Big	333	37	1975 - 2011	-37.52	146.08
415217	Grampians Rd Br	Fyans Ck	34	38	1973 - 2010	-37.26	142.53

Table A. 5 Study Catchments of Clustering group B2

Station ID	Station Name	River Name	Catchment area (km^2)	Record Length (years)	Period of Record	Lat	Lon
222206	Buchan	Buchan	822	38	1974 - 2011	-37.5	148.18
226209	Darnum	Moe	214	40	1972 - 2011	-38.21	146
226402	Trafalgar East	Moe Drain	622	37	1975 - 2011	-38.18	146.21
227219	Loch	Bass	52	39	1973 - 2011	-38.38	145.56
227226	Dumbalk North	Tarwineast Branc	127	42	1970 - 2011	-38.5	146.16
227231	Glen Forbes South	Bass	233	37	1974 - 2010	-38.47	145.51
227236	D/S Foster Ck Jun	Powlett	228	33	1979 - 2011	-38.56	145.71
230204	Riddells Ck	Riddells Ck	79	38	1974 - 2011	-37.47	144.67
230205	Bulla (D/S of Emu Ck Jun)	Deep Ck	865	38	1974 - 2011	-37.63	144.8
230211	Clarkefield	Emu Ck	93	36	1975 - 2010	-37.47	144.75
231231	Melton South	Toolern Ck	95	32	1979 - 2010	-37.91	144.58
232213	U/S of Bungal Dam	Lal Lal Ck	157	33	1977 - 2009	-37.66	144.03
234200	Pitfield	Woady Yaloak	324	40	1972 - 2011	-37.81	143.59
235203	Curdie	Curdies	790	37	1975 - 2011	-38.45	142.96
236205	Woodford	Merri	899	38	1974 - 2011	-38.32	142.48
236212	Cudgee	Brucknell Ck	570	37	1975 - 2011	-38.35	142.65
237207	Heathmere	Surry	310	37	1975 - 2011	-38.25	141.66
238219	Morgiana	Grange Burn	997	39	1973 - 2011	-37.71	141.83
401216	Jokers Ck	Big	356	60	1952 - 2011	-36.95	141.47
403209	Wangaratta North	Reedy Ck	368	39	1973 - 2011	-36.33	146.34
403221	Woolshed	Reedy Ck	214	37	1975 - 2011	-36.31	146.6
405212	Tallarook	Sunday Ck	337	37	1975 - 2011	-37.1	145.05
405226	Moorilim	Pranjip Ck	787	38	1974 -	-36.62	145.31

Station ID	Station Name	River Name	Catchment area (km^2)	Record Length (years)	Period of Record	Lat	Lon
					2011		
405229	Wanalta	Wanalta Ck	108	43	1969 - 2011	-36.64	144.87
405230	Colbinabbin	Cornella Ck	259	39	1973 - 2011	-36.61	144.8
405240	Ash Br	Sugarloaf Ck	609	39	1973 - 2011	-37.06	145.05
405245	Mansfield	Ford Ck	115	42	1970 - 2011	-37.04	146.05
405248	Graytown	Major Ck	282	41	1971 - 2011	-36.86	144.91
405274	Yarck	Home Ck	187	35	1977 - 2011	-37.11	145.6
406213	Redesdale	Campaspe	629	37	1975 - 2011	-37.02	144.54
406214	Longlea	Axe Ck	234	40	1972 - 2011	-36.78	144.43
406216	Sedgewick	Axe Ck	34	37	1975 - 2011	-36.9	144.36
406224	Runnymede	Mount Pleasant C	248	37	1975 - 2011	-36.55	144.64
406226	Derrinal	Mount Ida Ck	174	34	1978 - 2011	-36.88	144.65
407214	Clunes	Creswick Ck	308	37	1975 - 2011	-37.3	143.79
407217	Vaughan atD/S Fryers Ck	Loddon	299	44	1968 - 2011	-37.16	144.21
407220	Norwood	Bet Bet Ck	347	38	1973 - 2010	-37	143.64
407221	Yandoit	Jim Crow Ck	166	39	1973 - 2011	-37.21	144.1
407222	Clunes	Tullaroop Ck	632	39	1973 - 2011	-37.23	143.83
407230	Strathlea	Joyces Ck	153	39	1973 - 2011	-37.17	143.96
407246	Marong	Bullock Ck	184	39	1973 - 2011	-36.73	144.13
407253	Minto	Piccaninny Ck	668	39	1973 - 2011	-36.45	144.47
415207	Eversley	Wimmera	304	37	1975 - 2011	-37.19	143.19
415220	Wimmera HWY	Avon	596	37	1974 - 2010	-36.64	142.98
415226	Carrs Plains	Richardson	130	31	1971 - 2001	-36.75	142.79
415237	Stawell	Concongella Ck	239	35	1977 - 2011	-37.02	142.82
415238	Navarre	Wattle Ck	141	36	1976 - 2011	-36.9	143.1

APPENDIX B

Additional results from log-log linear model

Q_5 model

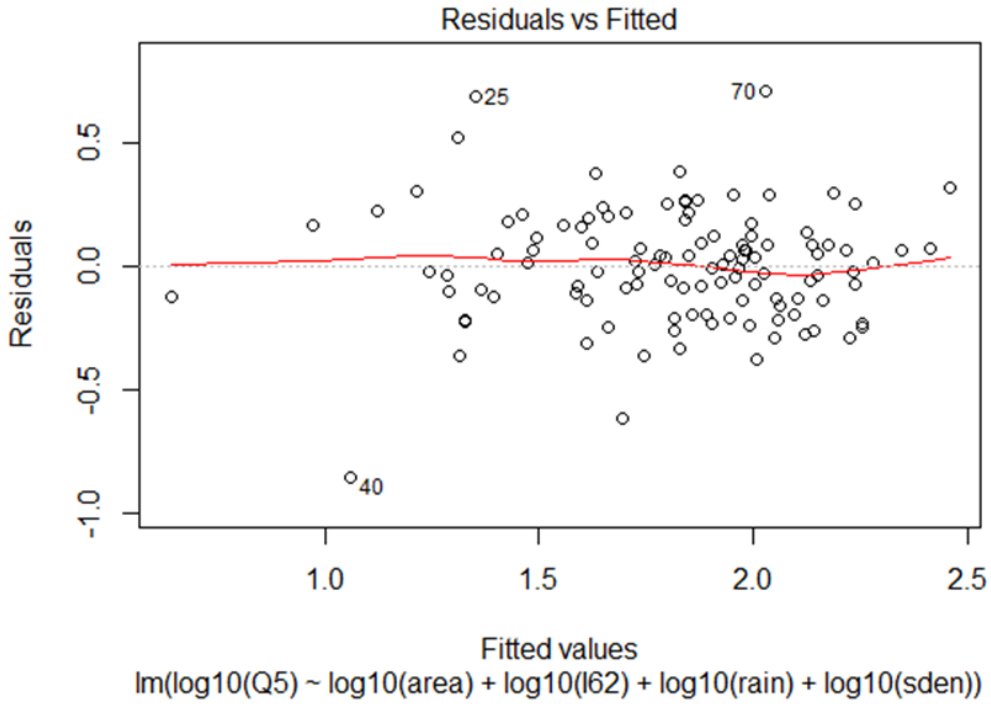

Figure B.1 Standardised residual vs fitted predicted value for the log-log linear model for combined group for Q_5

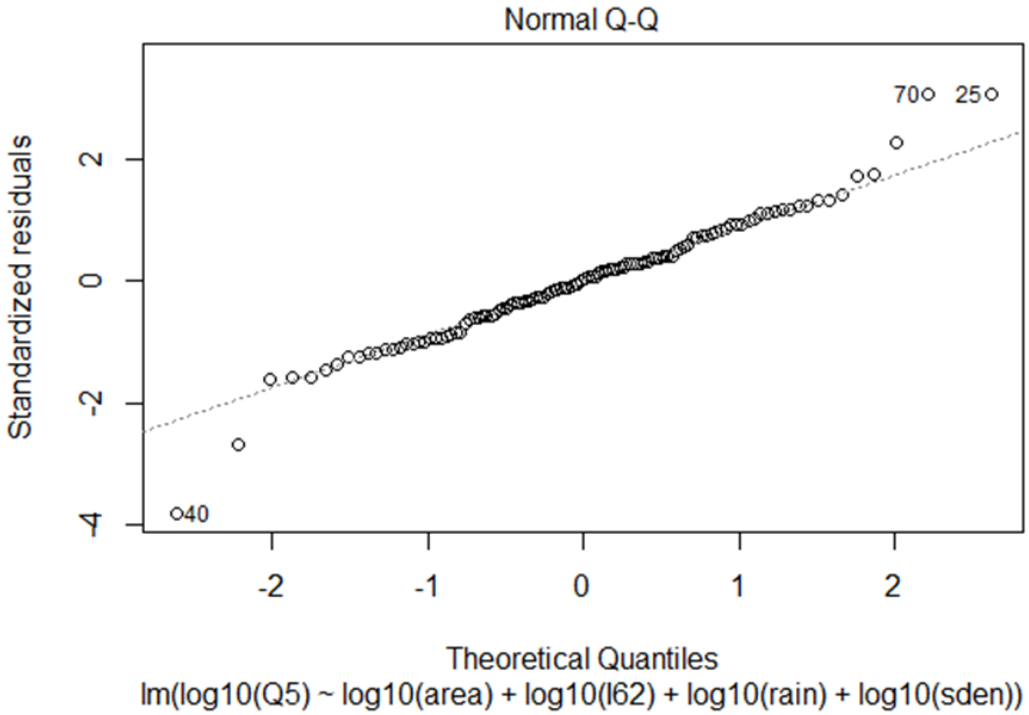

Figure B.2 Normal Q-Q plot for the standardised residuals for the log-log linear model for combined group for Q_5

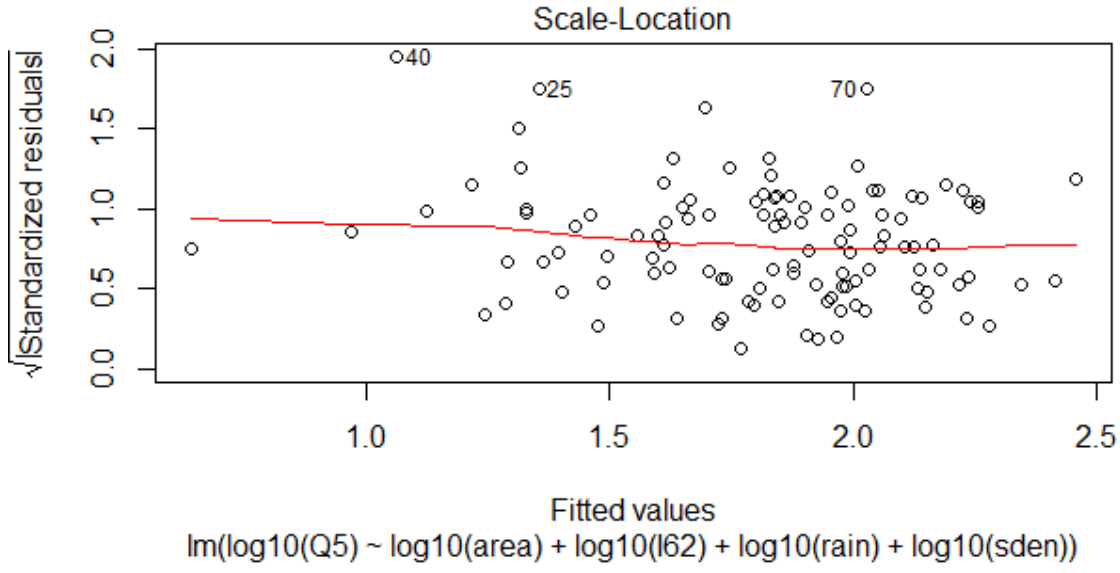

Figure B.3 Scale-location plot between predicted values and standardised residuals for the log-log linear model for combined group for Q_5

Q_{10} model

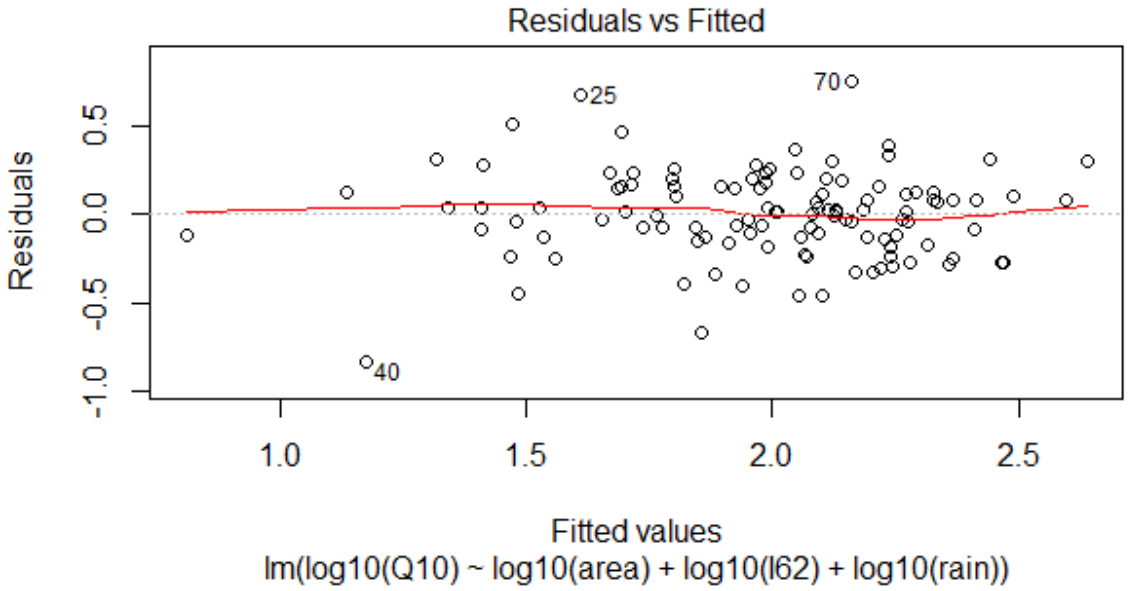

Figure B.4 Standardised residual vs fitted predicted value for the log-log linear model for combined group for Q_{10}

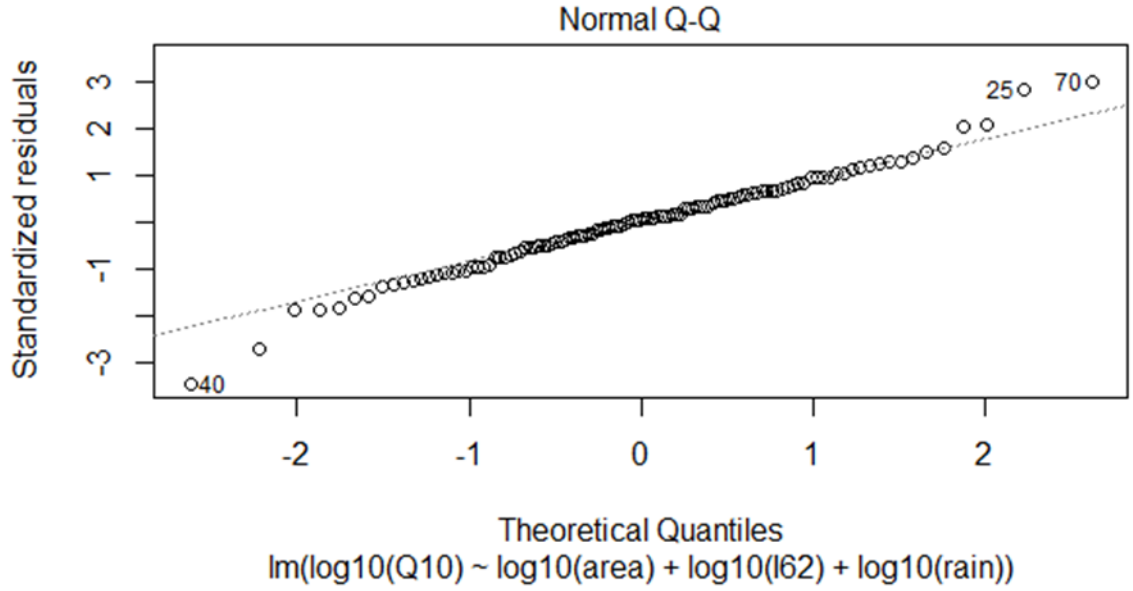

Figure B.5 Normal Q-Q plot for the standardised residuals for for the log-log linear model for combined group for Q_{10}

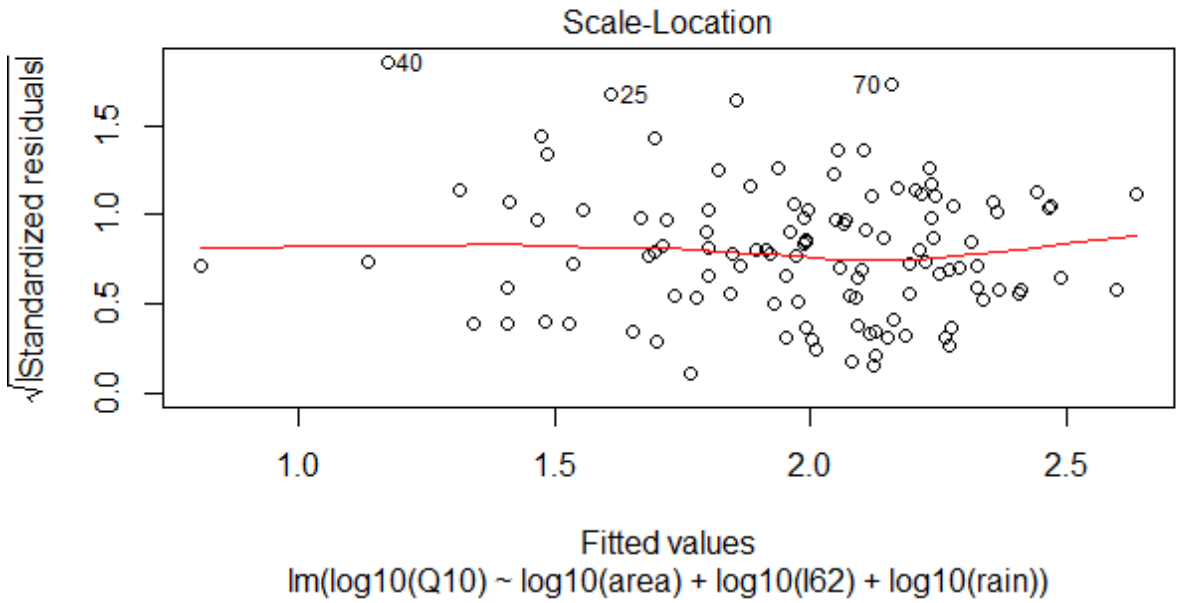

Figure B.6 Scale-location plot between predicted values and standardised residuals for the log-log linear model for combined group for Q_{10}

Q_{20} model diagnostics

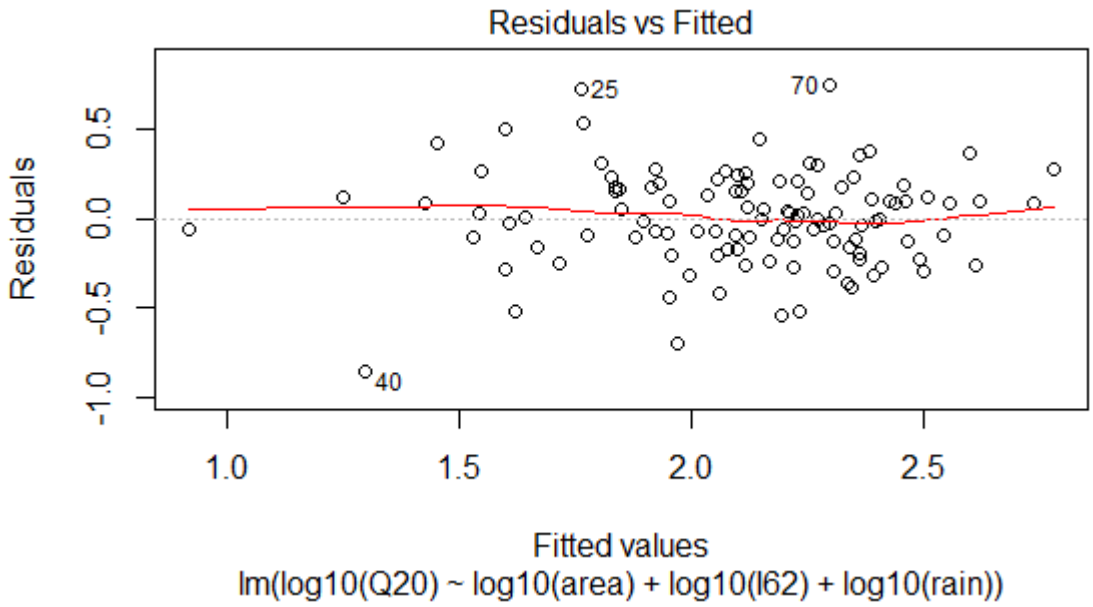

Figure B.7 Standardised residual vs fitted predicted value for the log-log linear model for combined group of Q_{20}

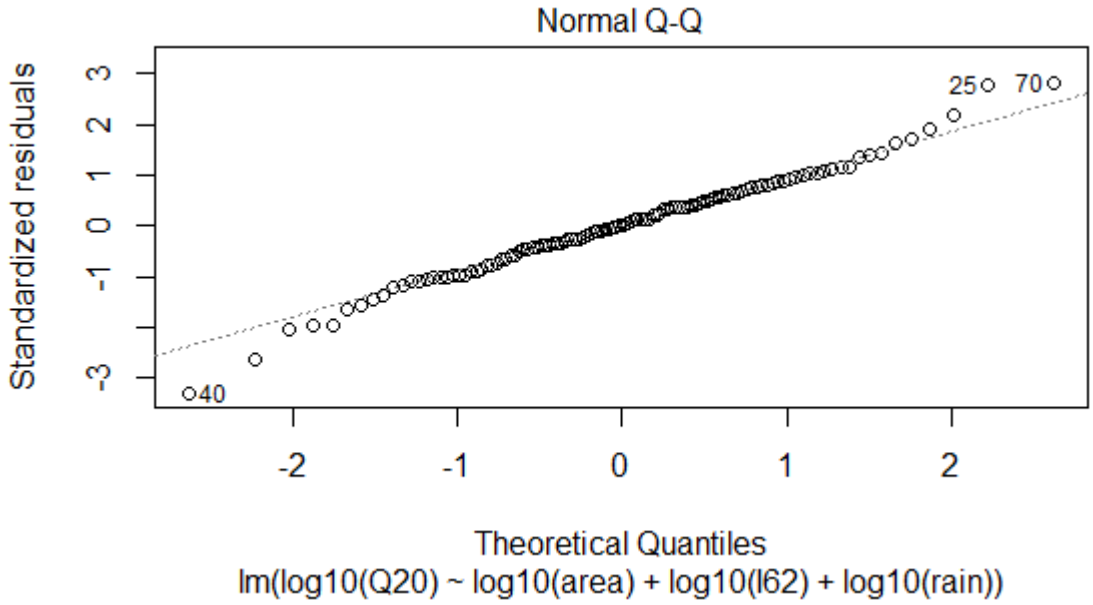

Figure B.8 Normal Q-Q plot for the standardised residuals for the log-log linear model for combined group of Q_{20}

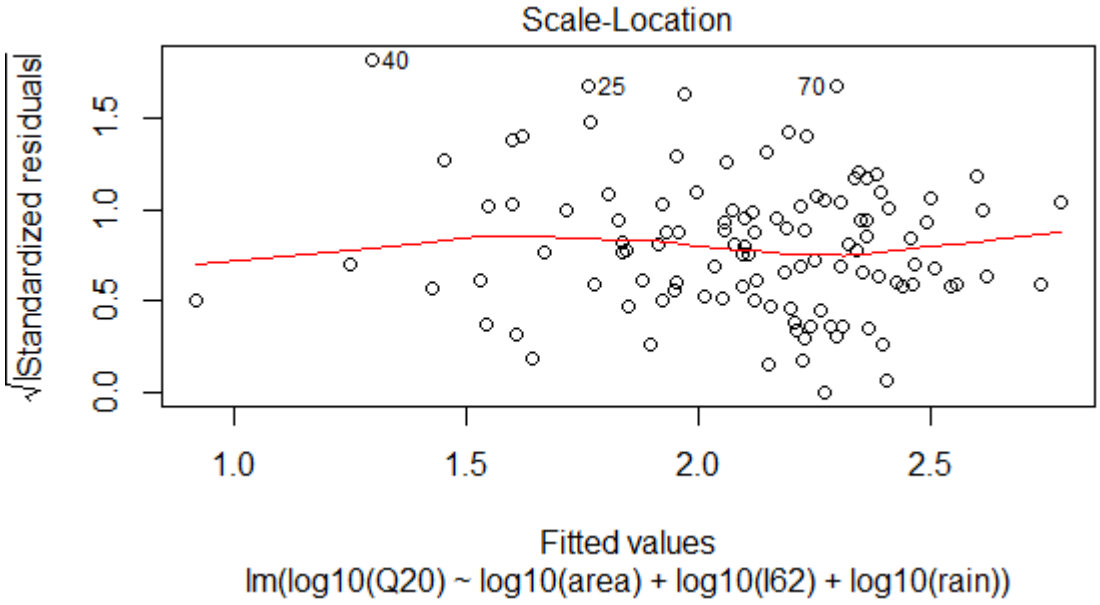

Figure B.9 Scale-location plot between predicted values and standardised residuals for the log-log linear model for combined group for Q_{20}

Q_{50} model diagnostics

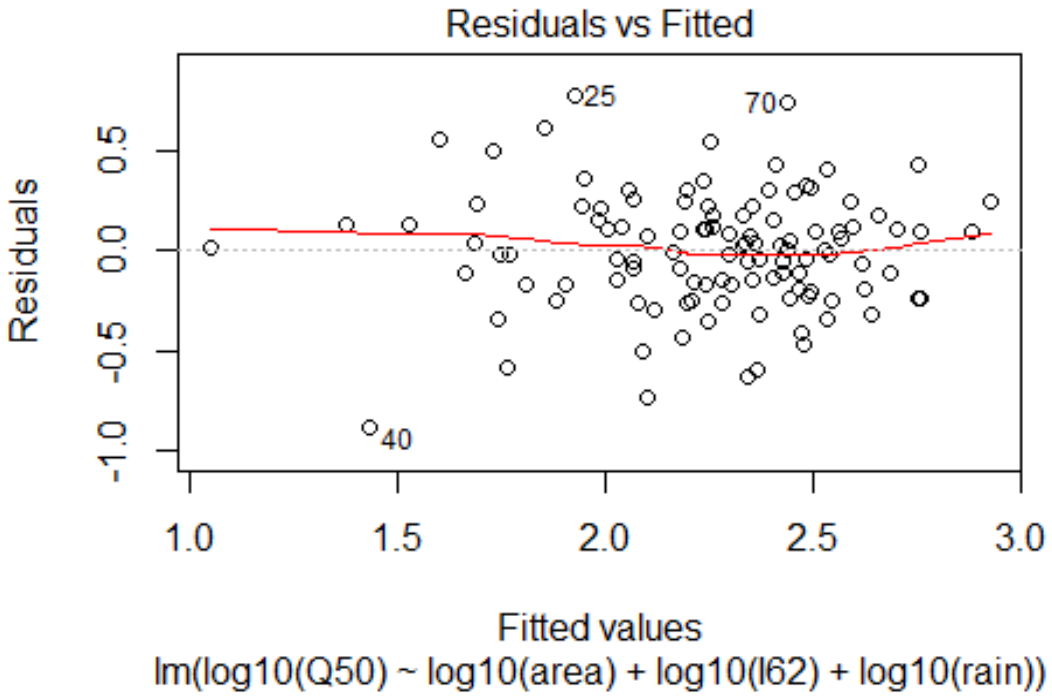

Figure B.10 Standardised residual vs fitted predicted value for the log-log linear model for combined group for Q_{50}

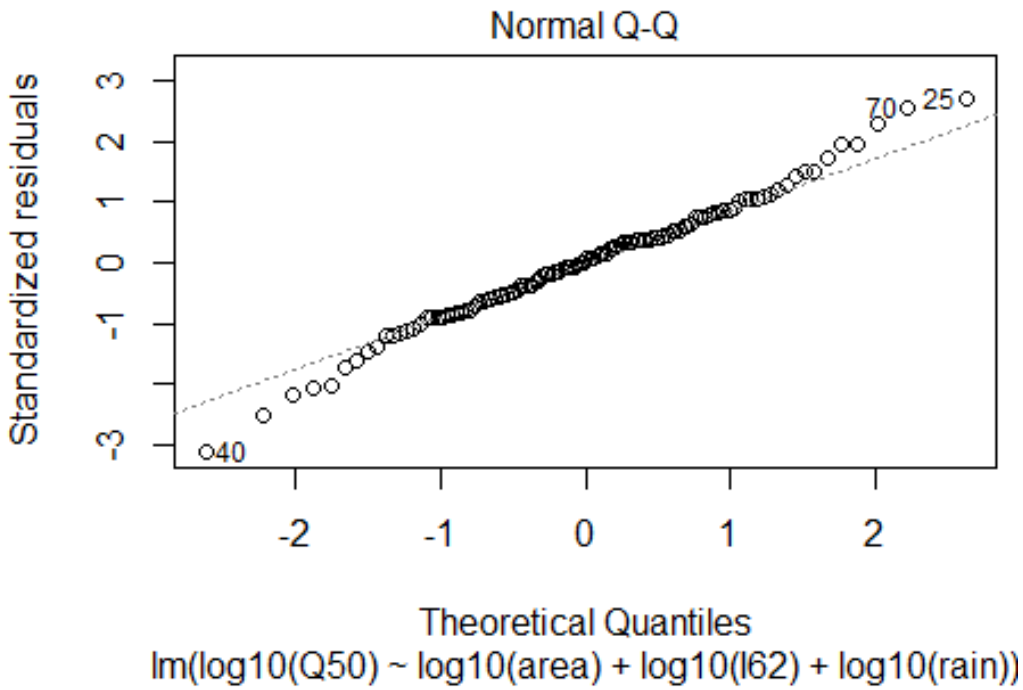

Figure B.11 Normal Q-Q plot for the standardised residuals for the log-log linear model for combined group for Q_{50}

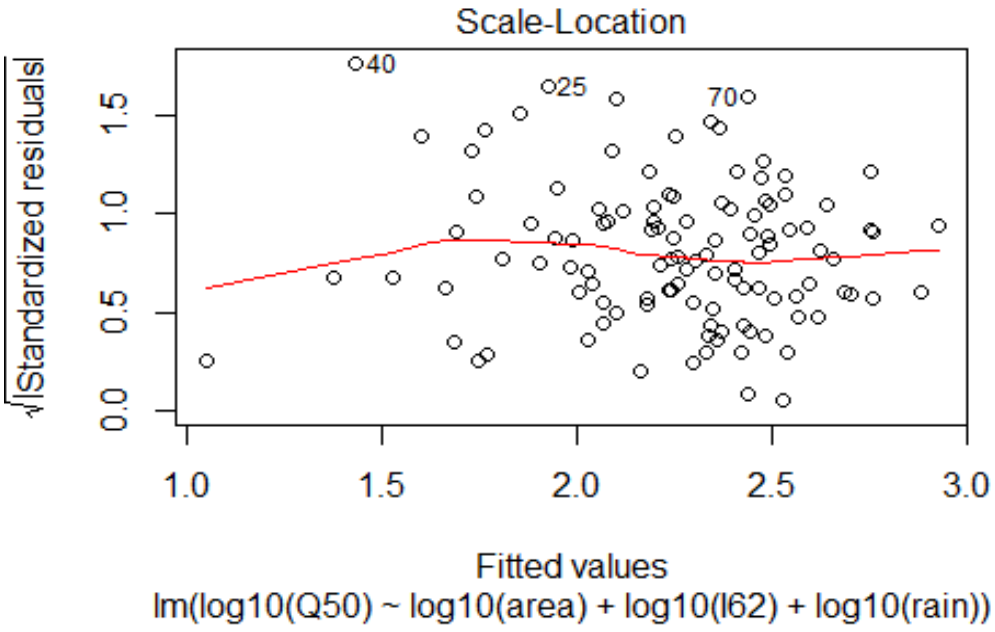

Figure B. 12 Scale-location plot between predicted values and standardised residuals for the log-log linear model for combined group for Q_{50}

Q_{100} model diagnostics

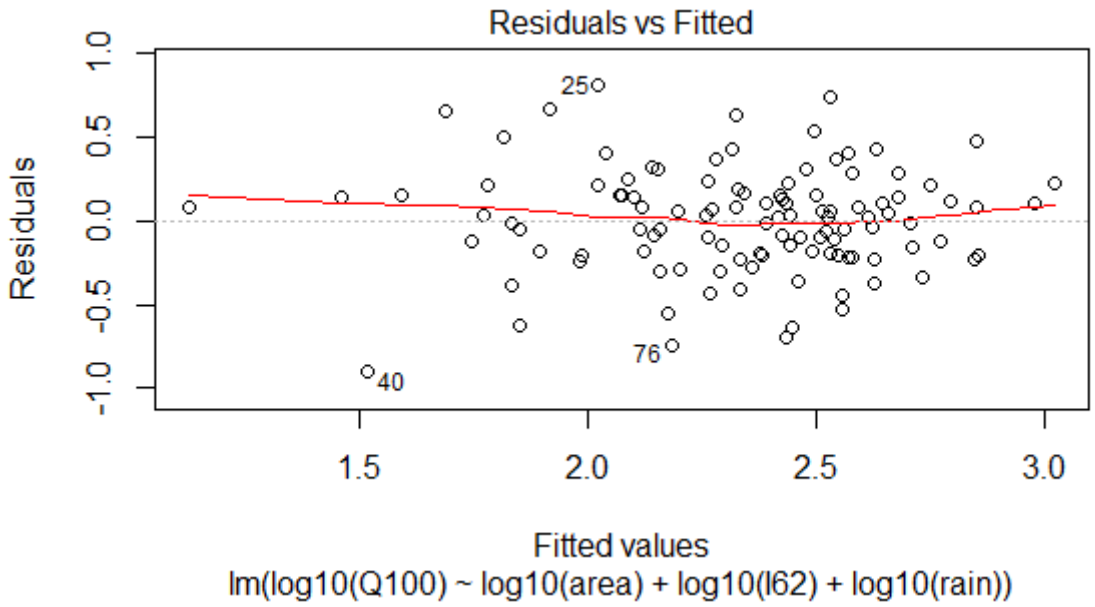

Figure B.13 Standardised residual vs fitted predicted value for the log-log linear model for combined group for Q_{100}

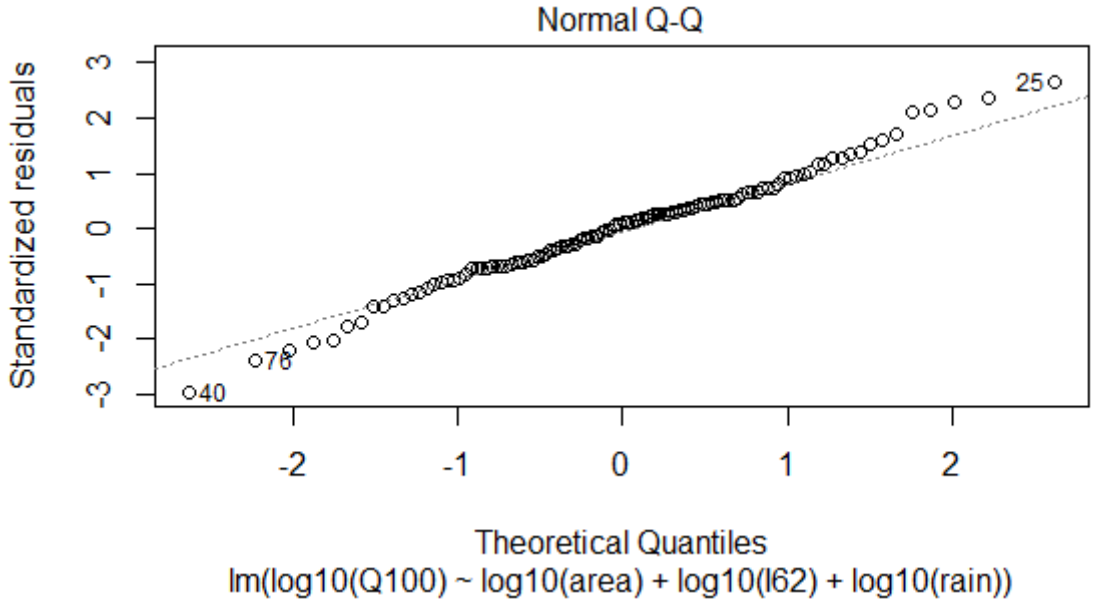

Figure B.14 Normal Q-Q plot for the standardised residuals for the log-log linear model for combined group for Q_{100}

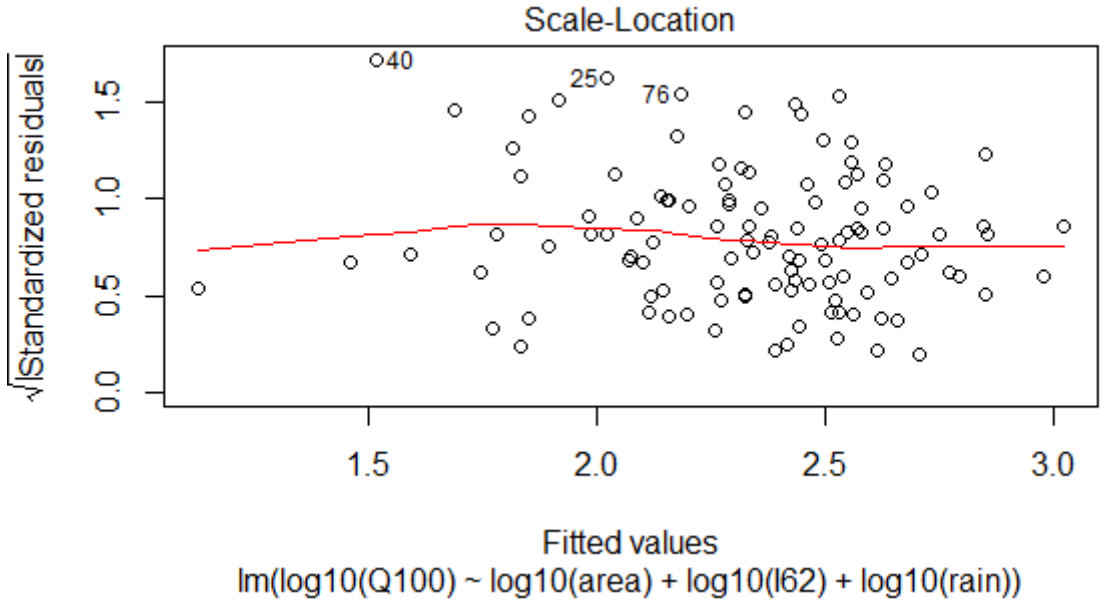

Figure B.15 Scale-location plot between predicted values and standardised residuals for the log-log linear model for combined group for Q_{100}

APPENDIX C

Additional results of log-log linear models (scatter plot of Q_{obs} vs Q_{pred})

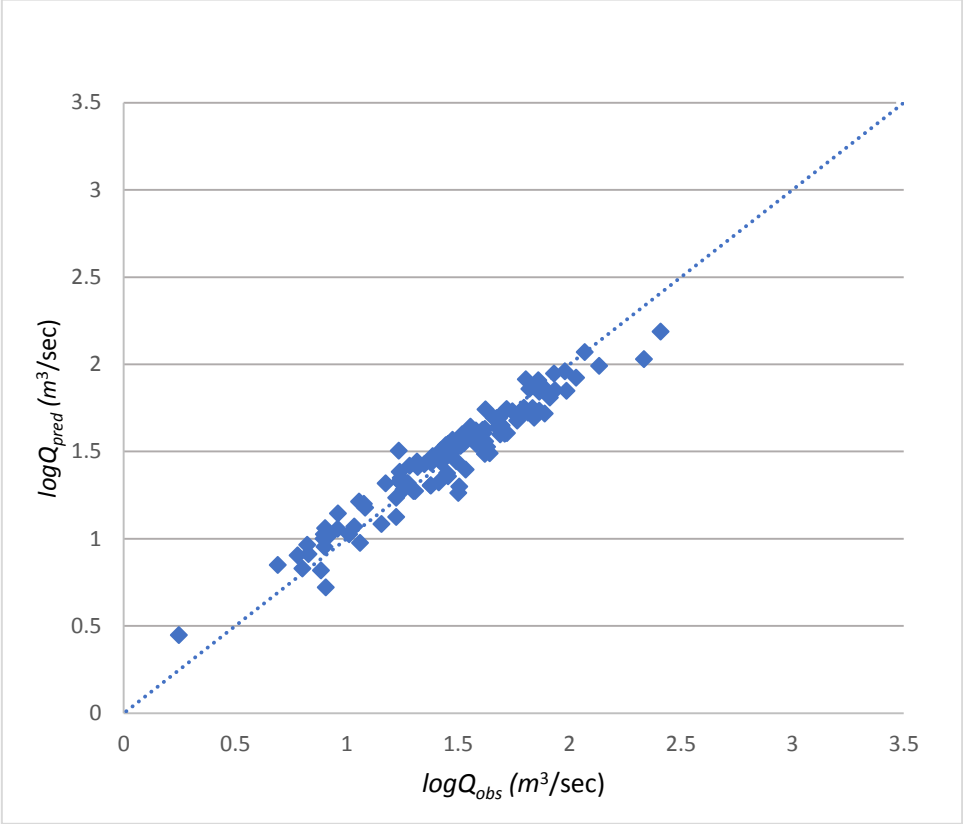

Figure C.1 Comparison of observed and predicted flood quantiles for log-log linear model of combined group for Q_2

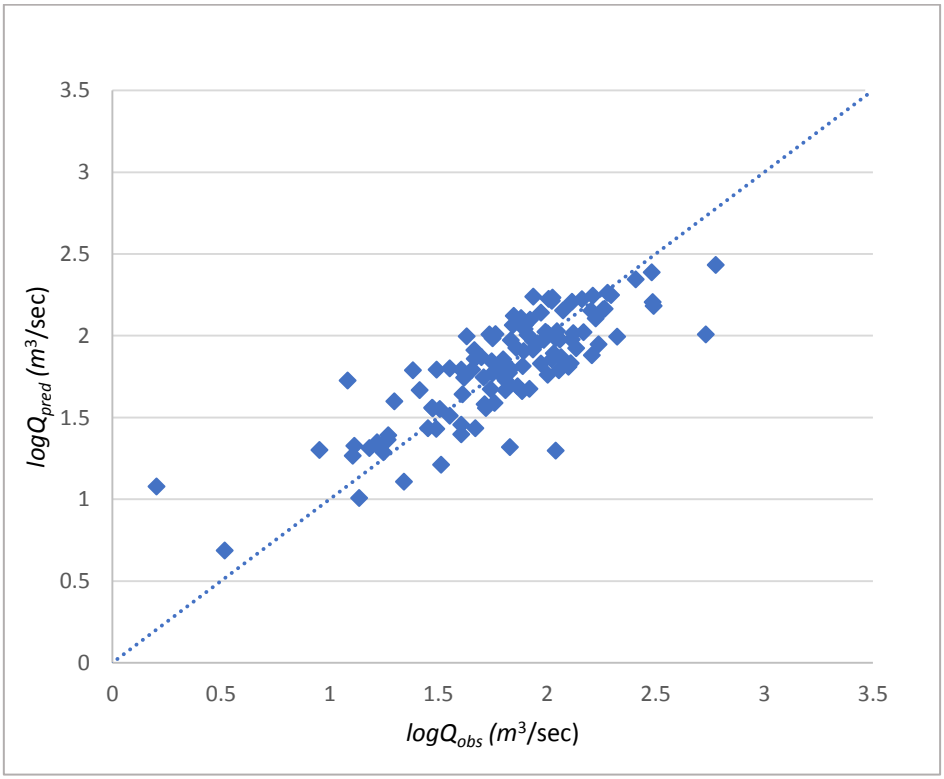

Figure C.2 Comparison of observed and predicted flood quantiles for log-log linear model of combined group for Q_5

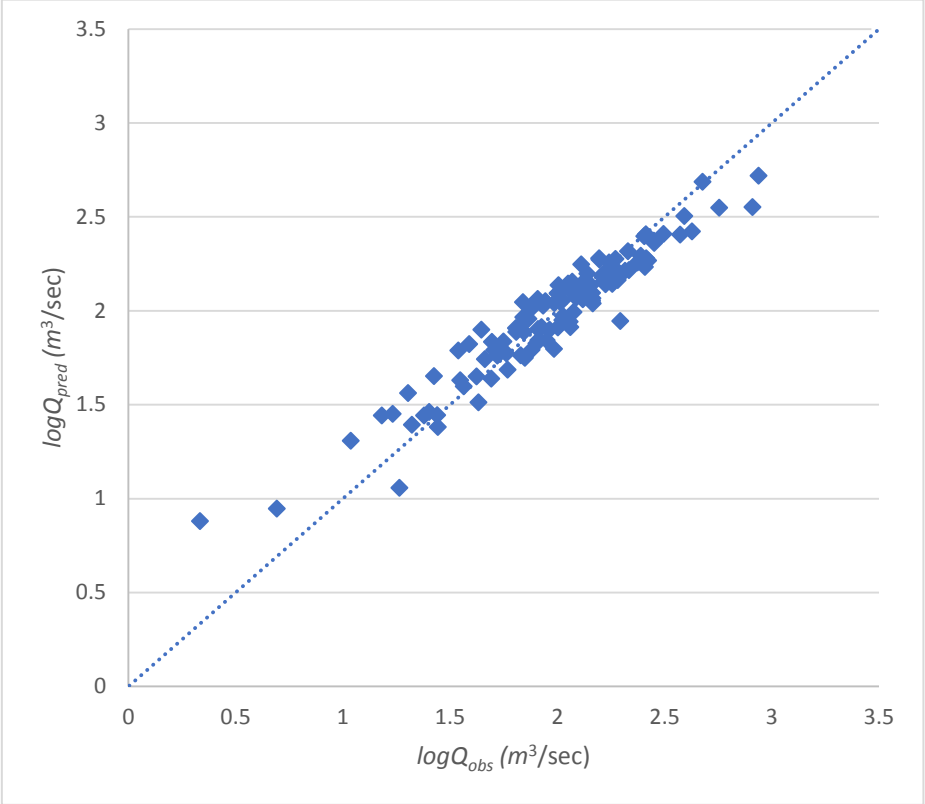

Figure C.3 Comparison of observed and predicted flood quantiles for for log-log linear model of combined group for Q_{10}

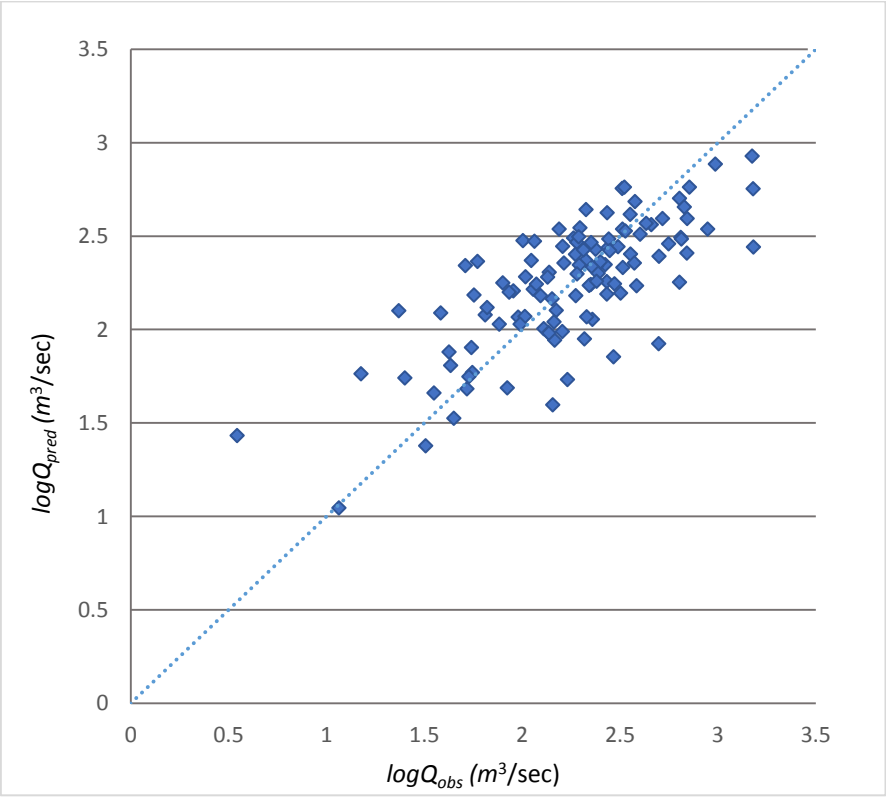

Figure C. 4 Comparison of observed and predicted flood quantiles for for log-log linear model of combined group for Q_{50}

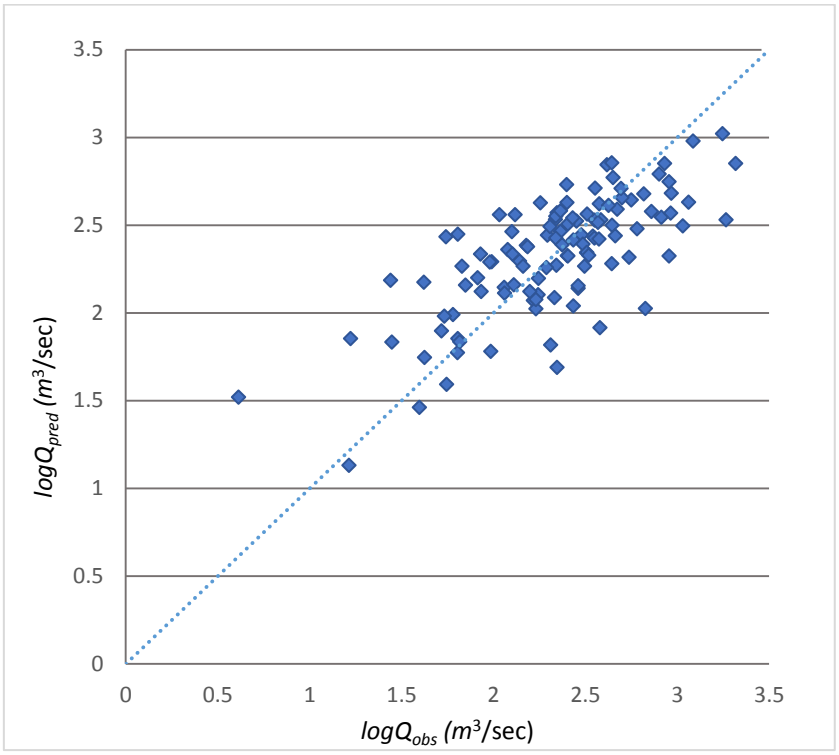

Figure C.5 Comparison of observed and predicted flood quantiles for for log-log linear model of combined group for Q_{100}

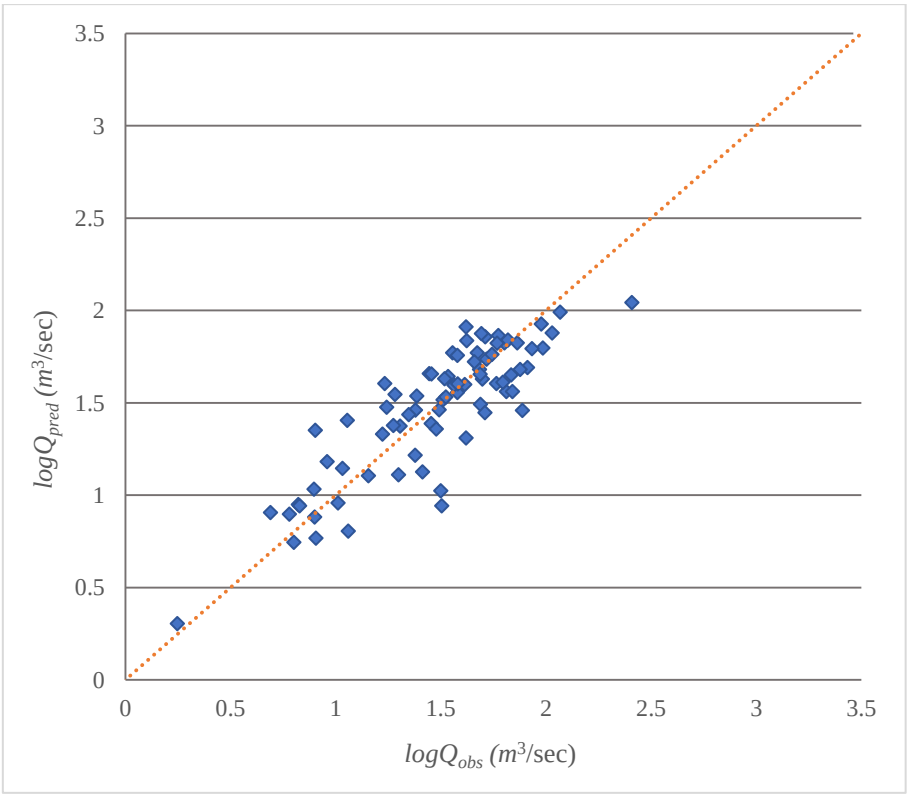

Figure C.6 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A1 for Q_2

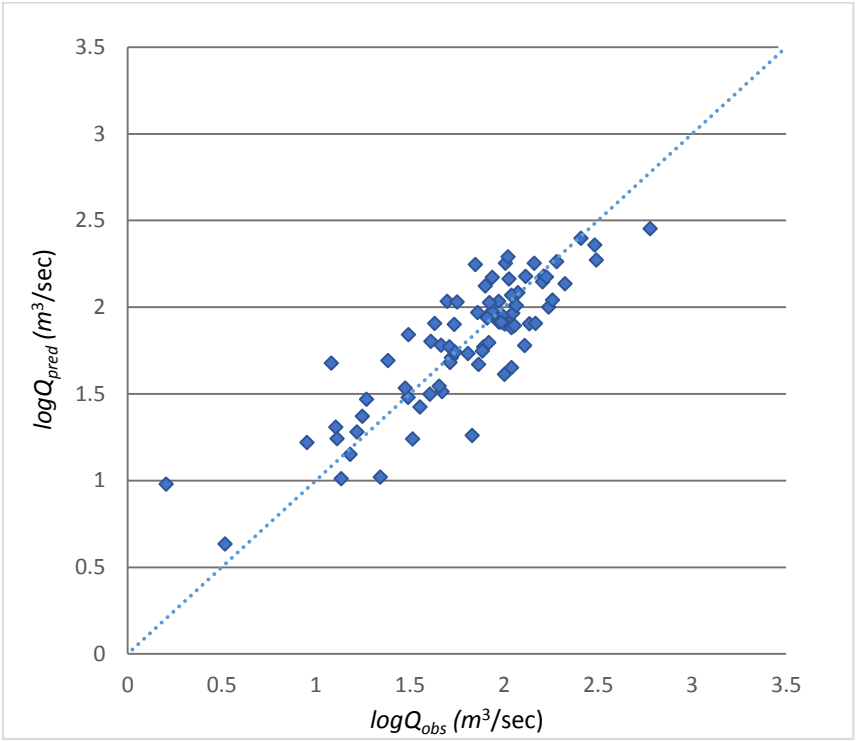

Figure C.7 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A1 for Q_5

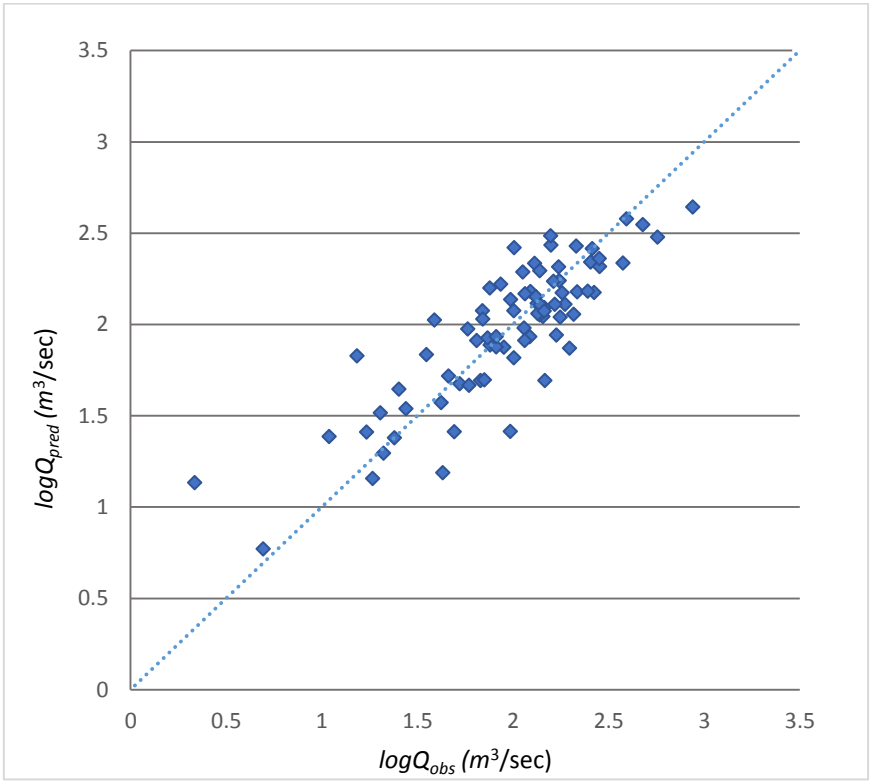

Figure C.8 Comparison of observed and predicted flood quantiles for for log-log linear model of clustering group A1 for Q_{10}

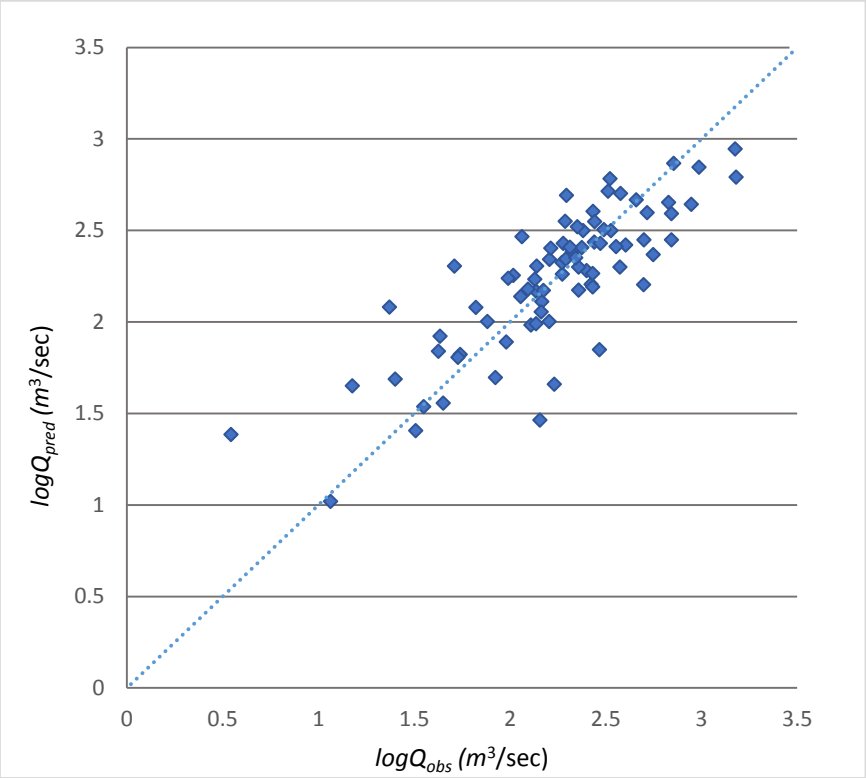

Figure C. 9 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A1 for Q_{50}

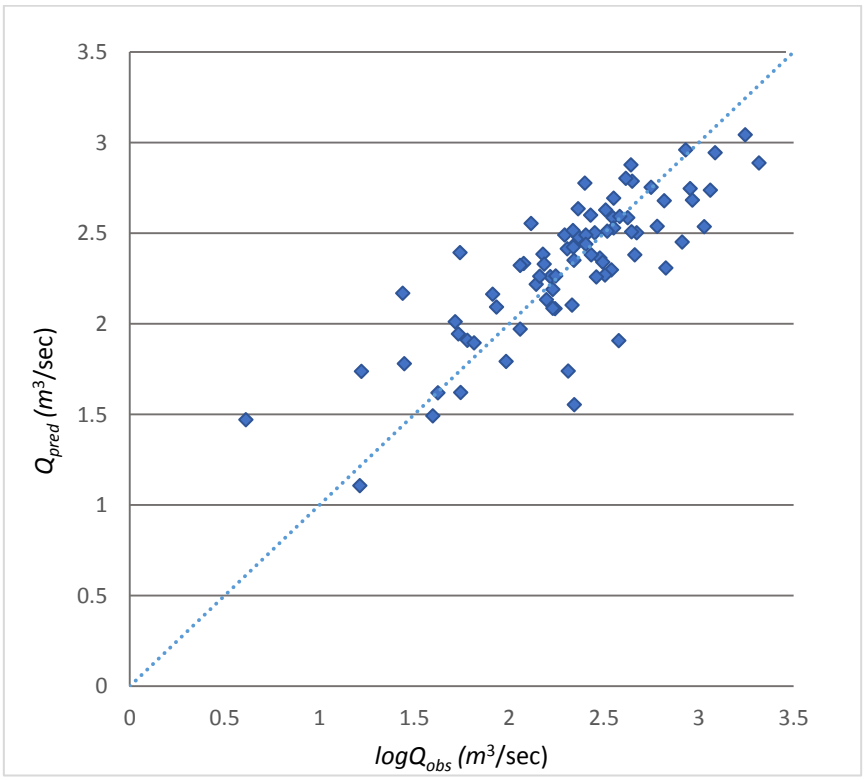

Figure C. 10 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A1 for Q_{100}

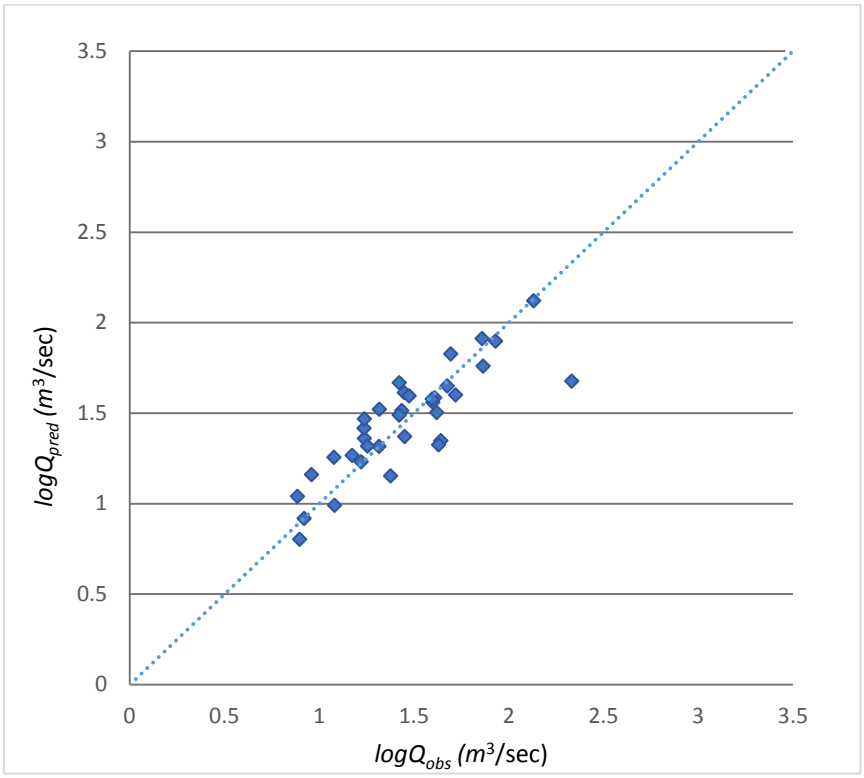

Figure C. 11 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A2 for Q_2

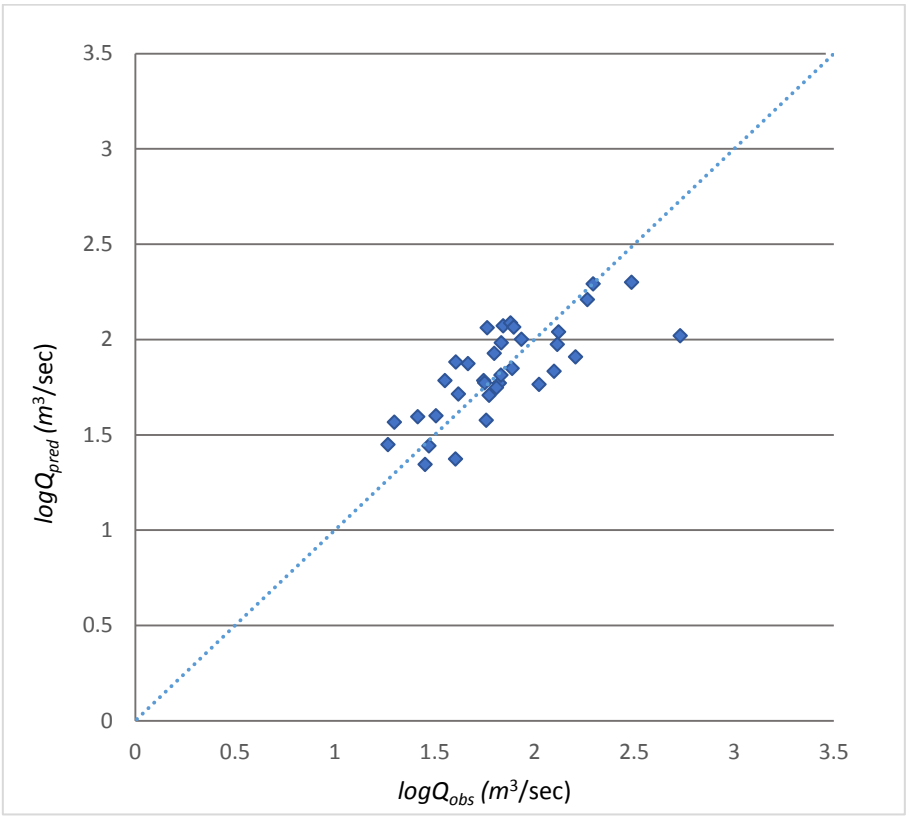

Figure C. 12 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A2 for Q_5

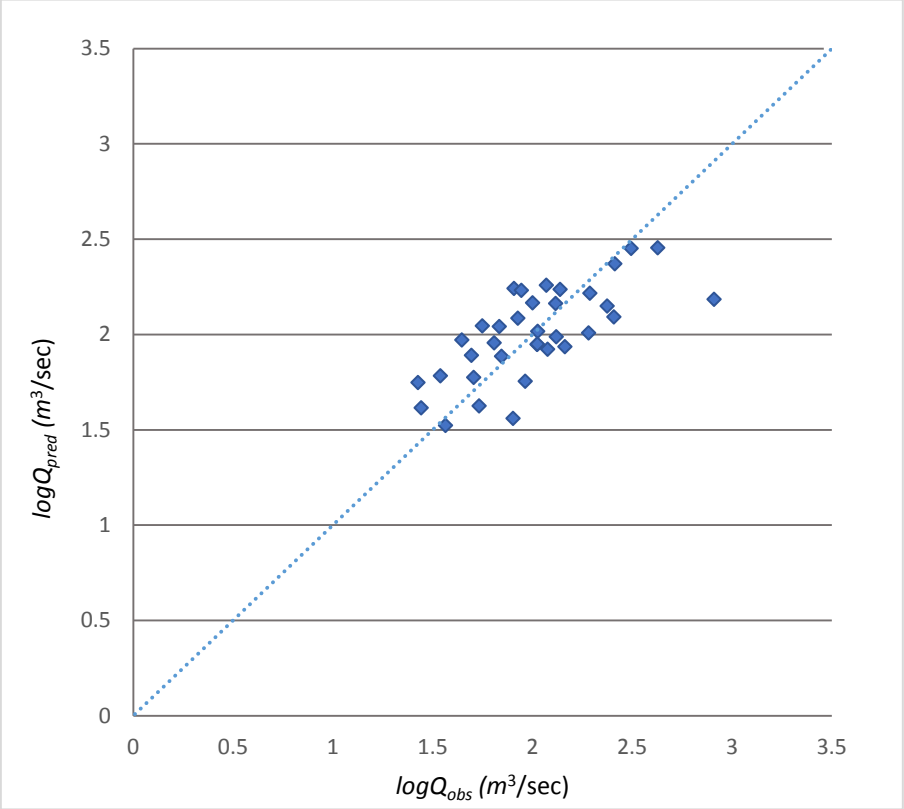

Figure C. 13 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A2 for Q_{10}

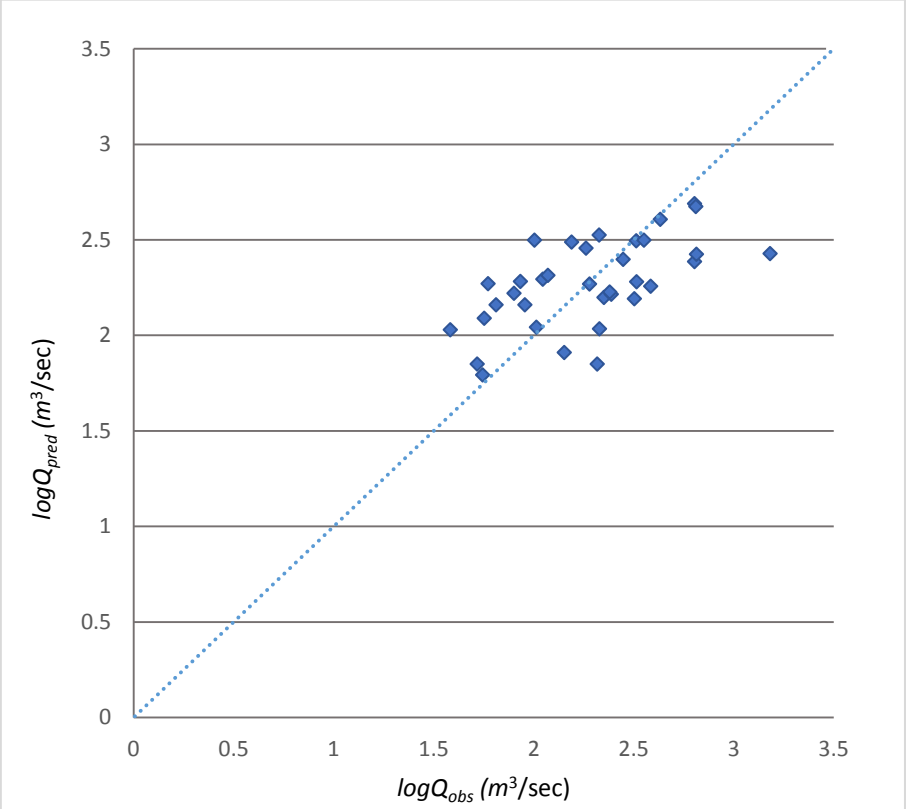

Figure C. 14 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A2 for Q_{50}

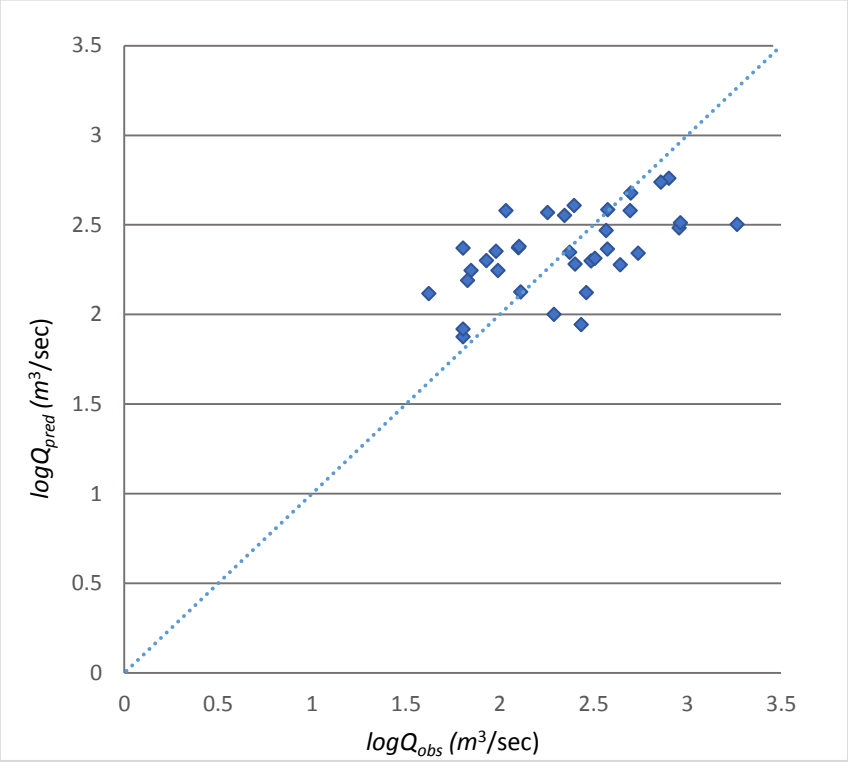

Figure C. 15 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group A2 for Q_{100}

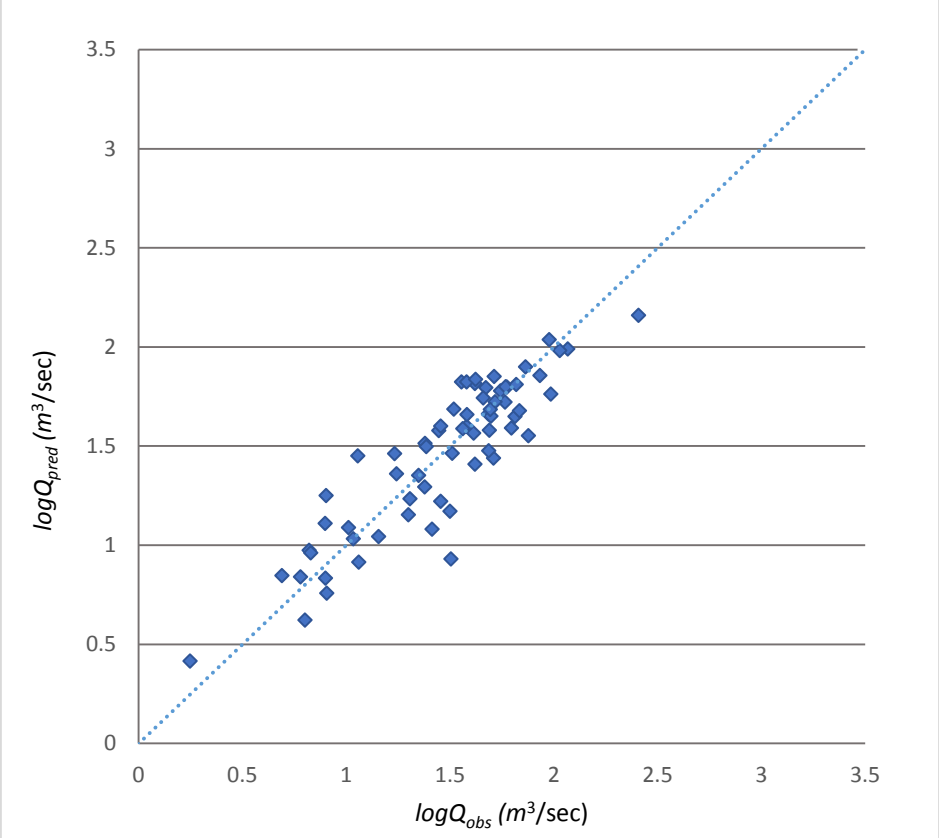

Figure C. 16 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B1 for Q_2

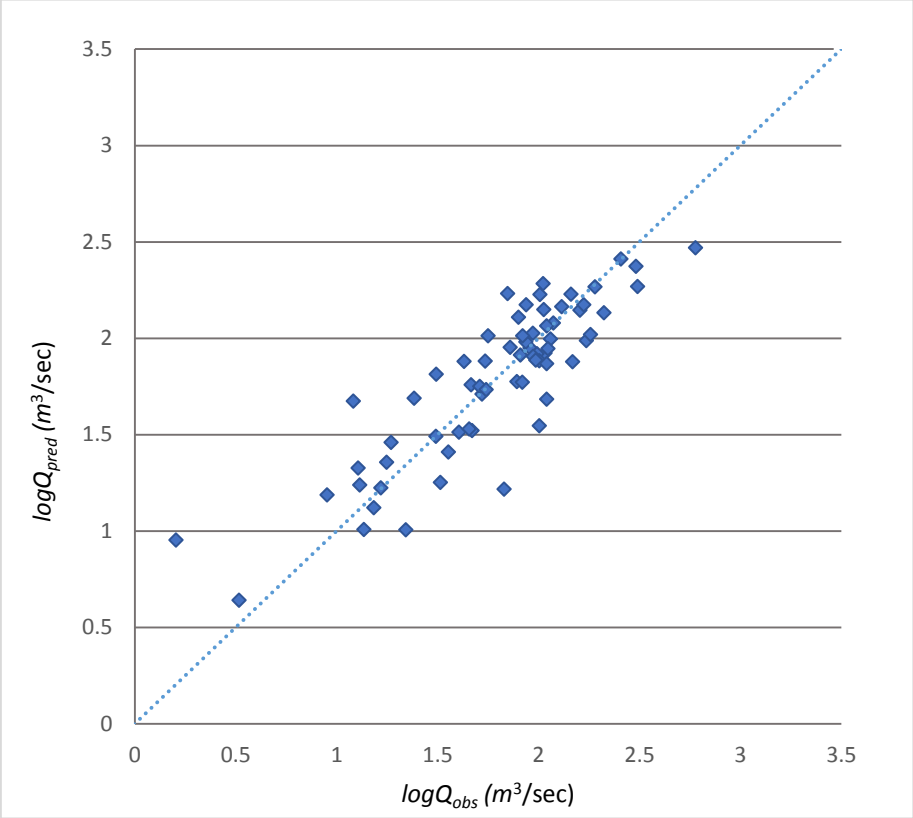

Figure C. 17 Comparison of observed and predicted flood quantiles for for log-log linear model of clustering group B1 for Q_5

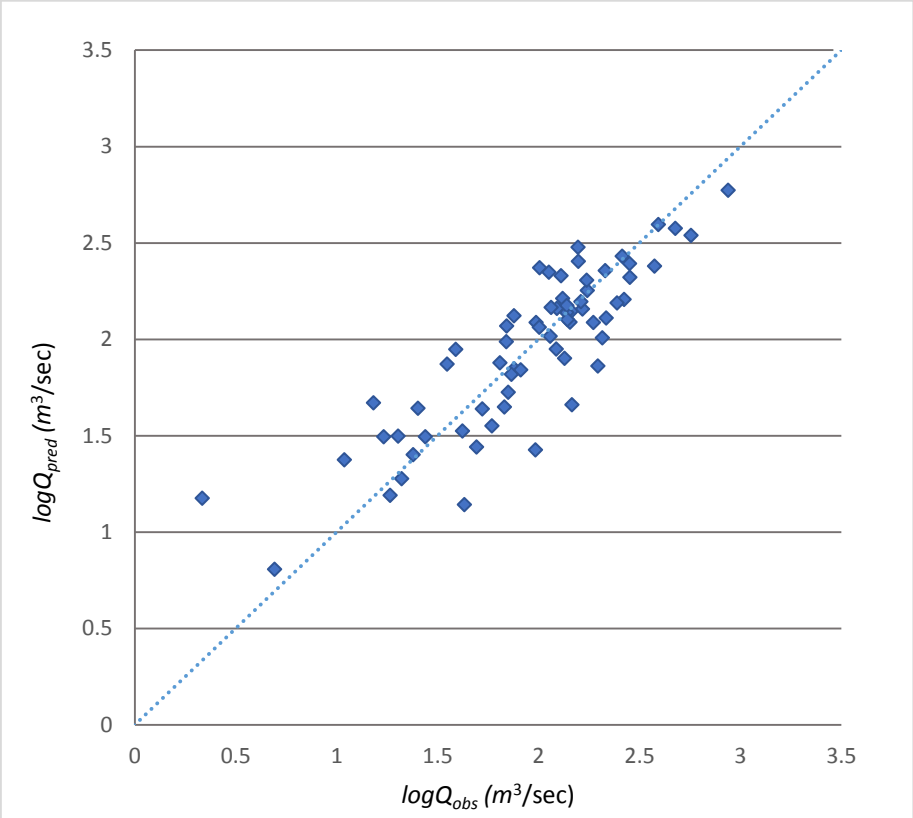

Figure C. 18 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B1 for Q_{10}

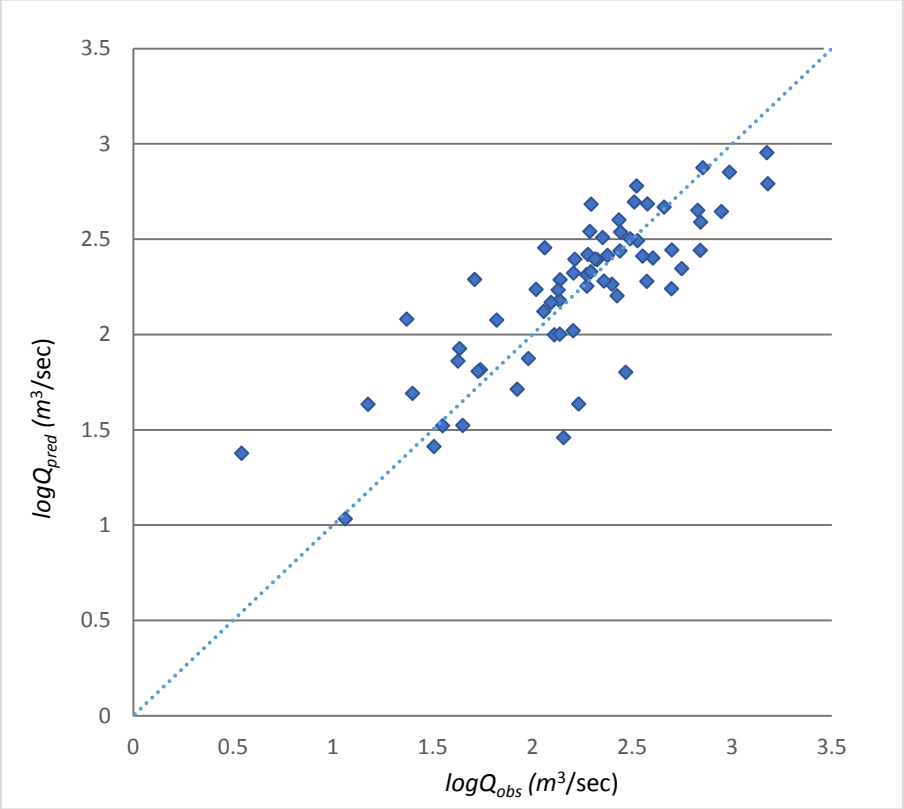

Figure C. 19 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B1 for Q_{50}

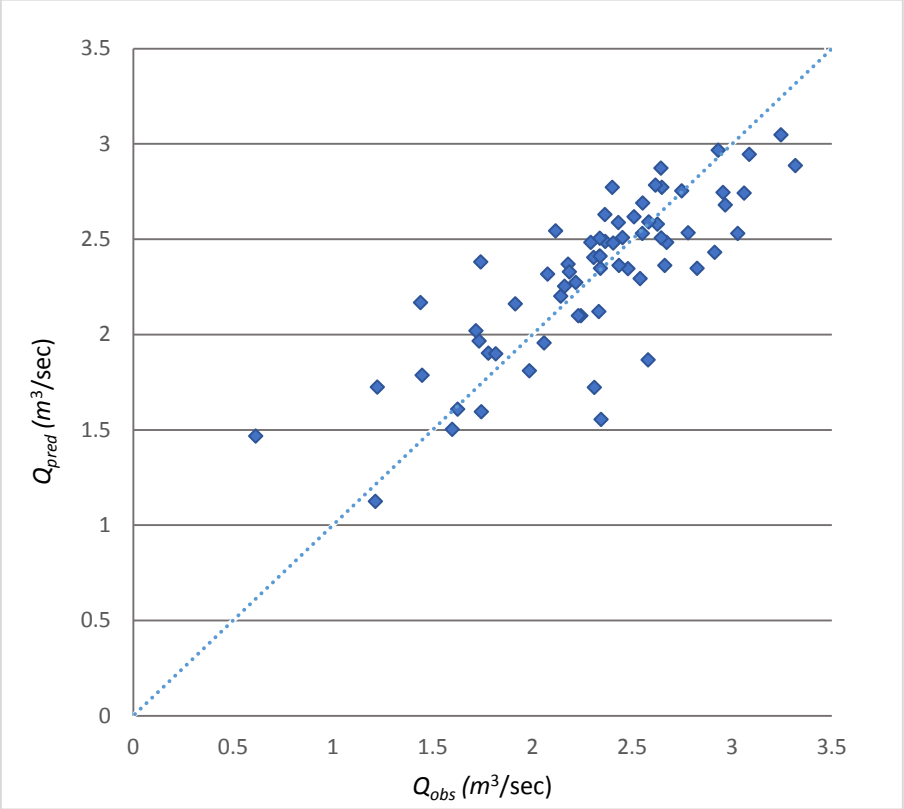

Figure C. 20 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B1 for Q_{100}

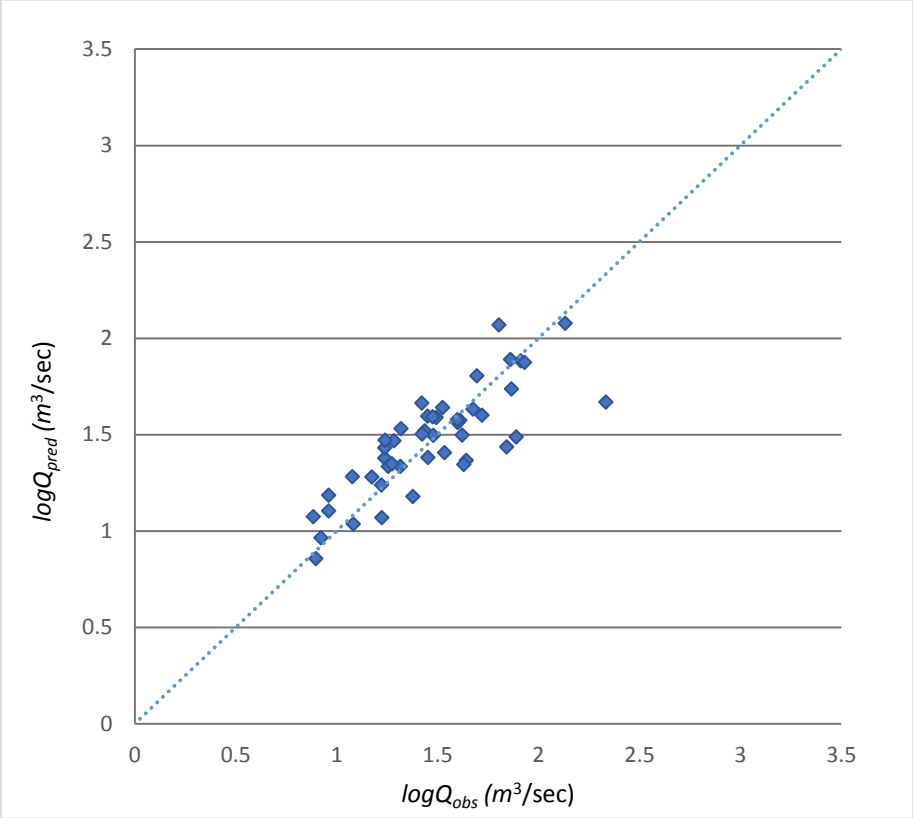

Figure C. 21 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B2 for Q_2

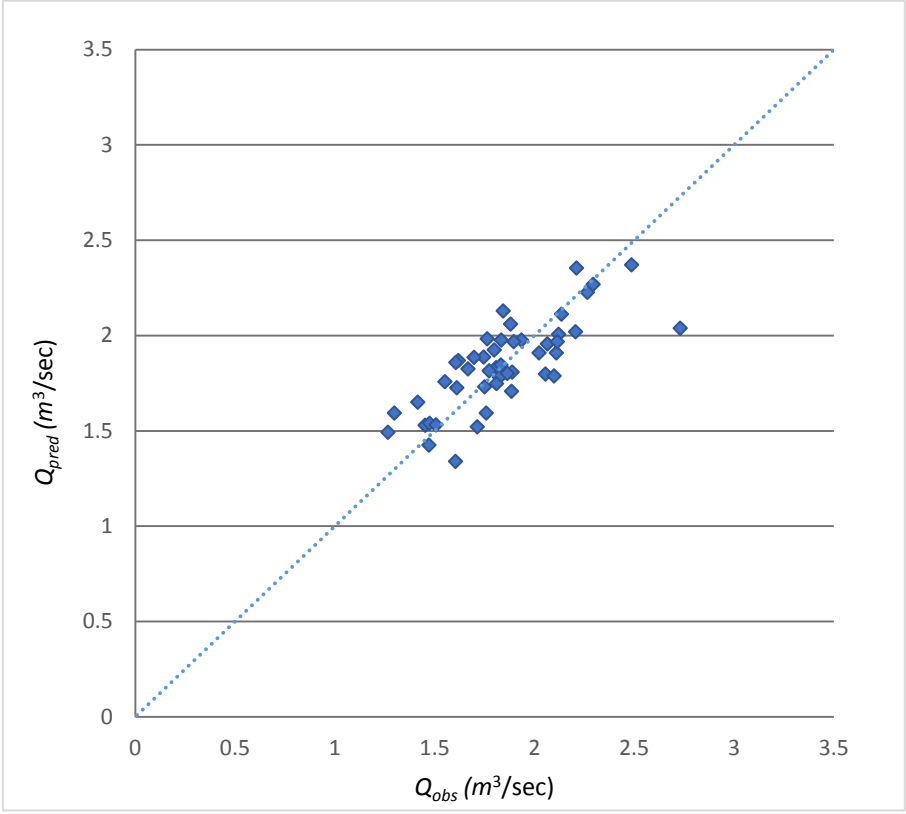

Figure C. 22 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B2 for Q_5

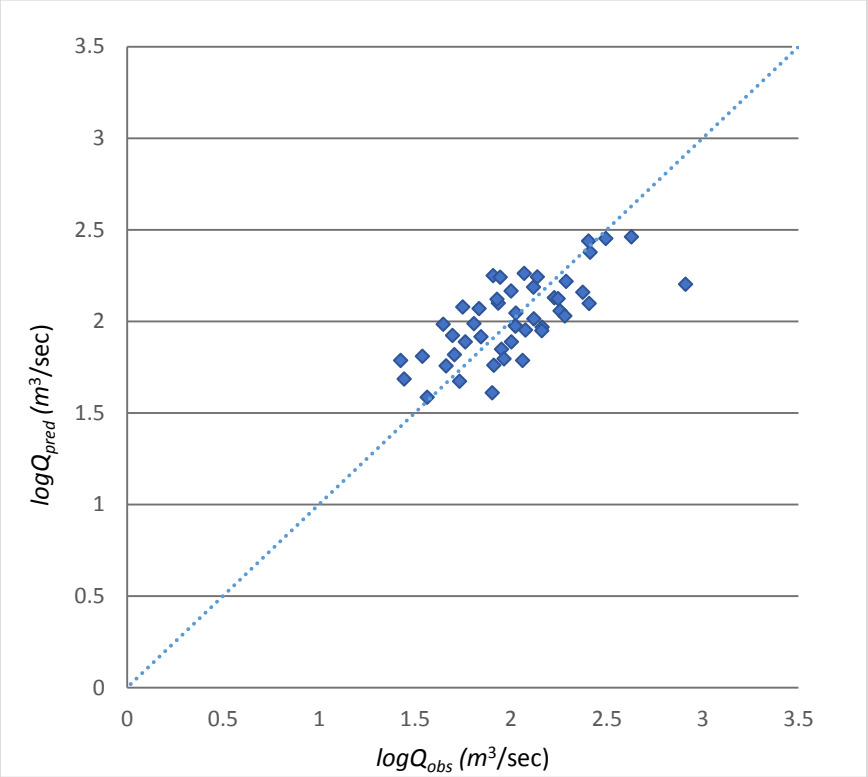

Figure C. 23 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B2 for Q_{10}

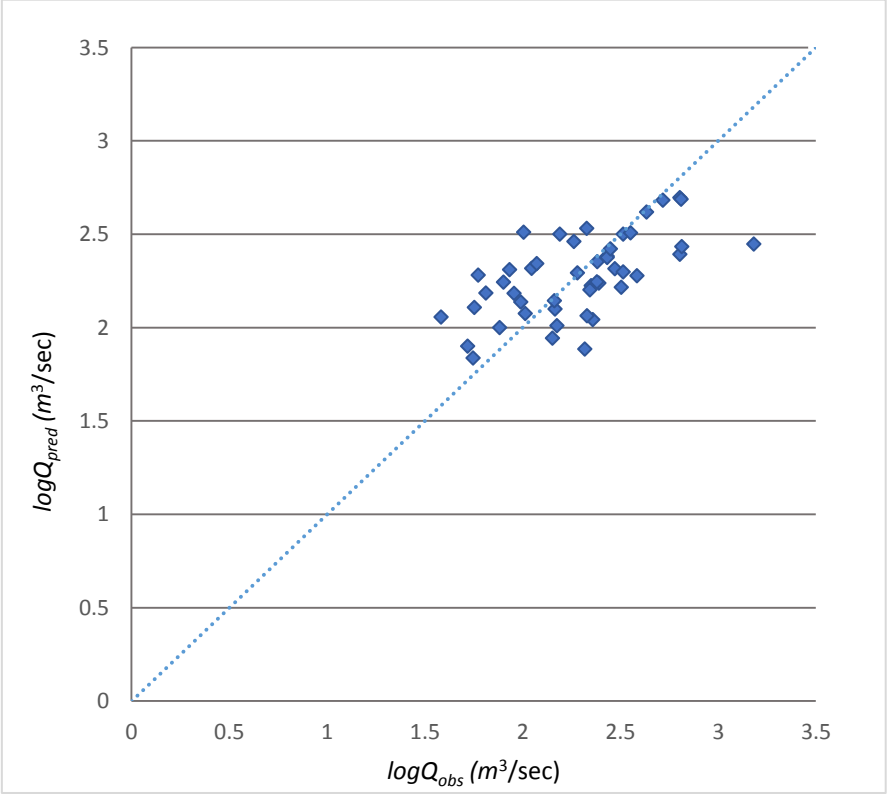

Figure C. 24 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B2 for Q_{50}

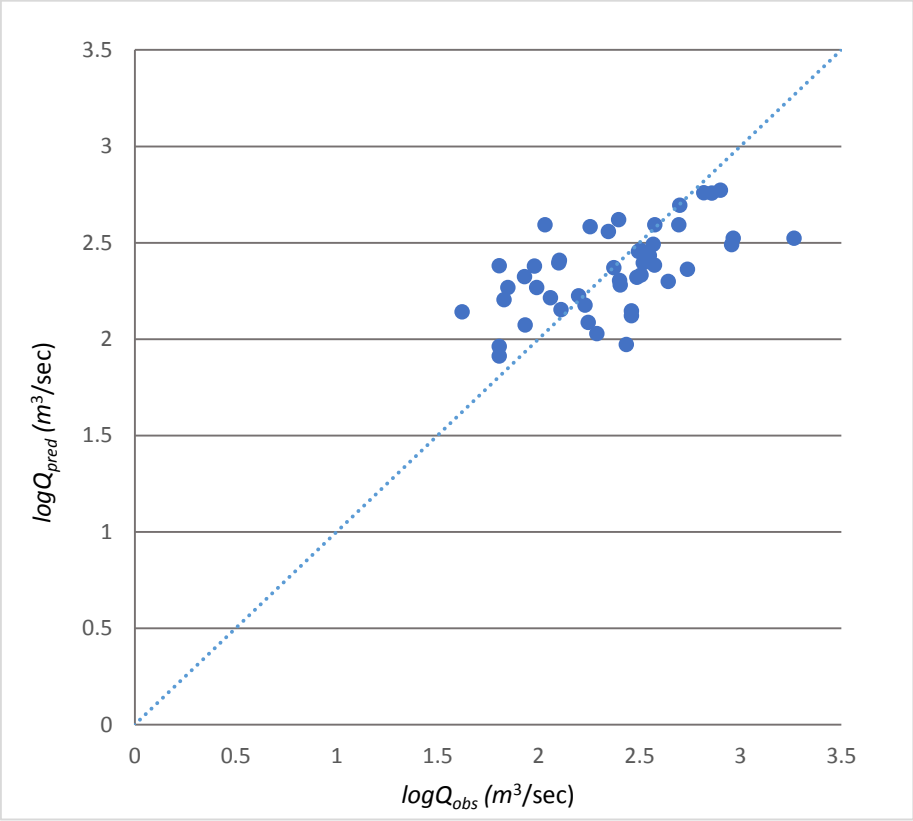

Figure C. 25 Comparison of observed and predicted flood quantiles for log-log linear model of clustering group B2 for Q_{100}

APPENDIX D

Additional results from GAM model

Q_2 model diagnostics

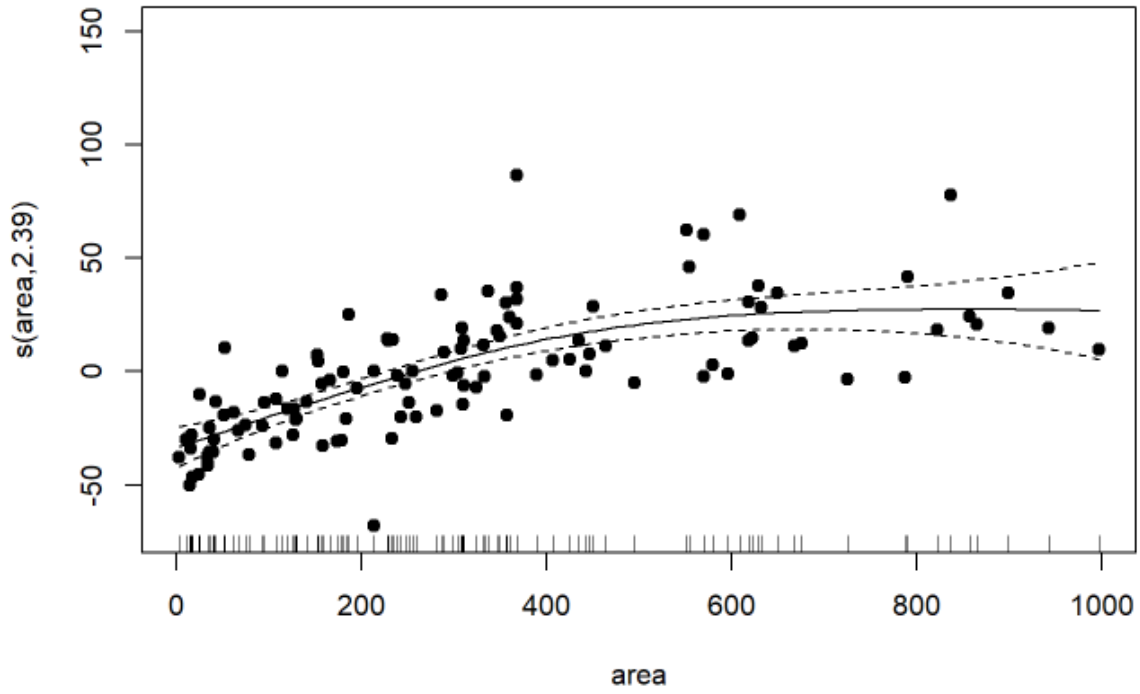

Figure D.1 Regression plot by smooth function for predictor variable *area* for Q_2 GAM model

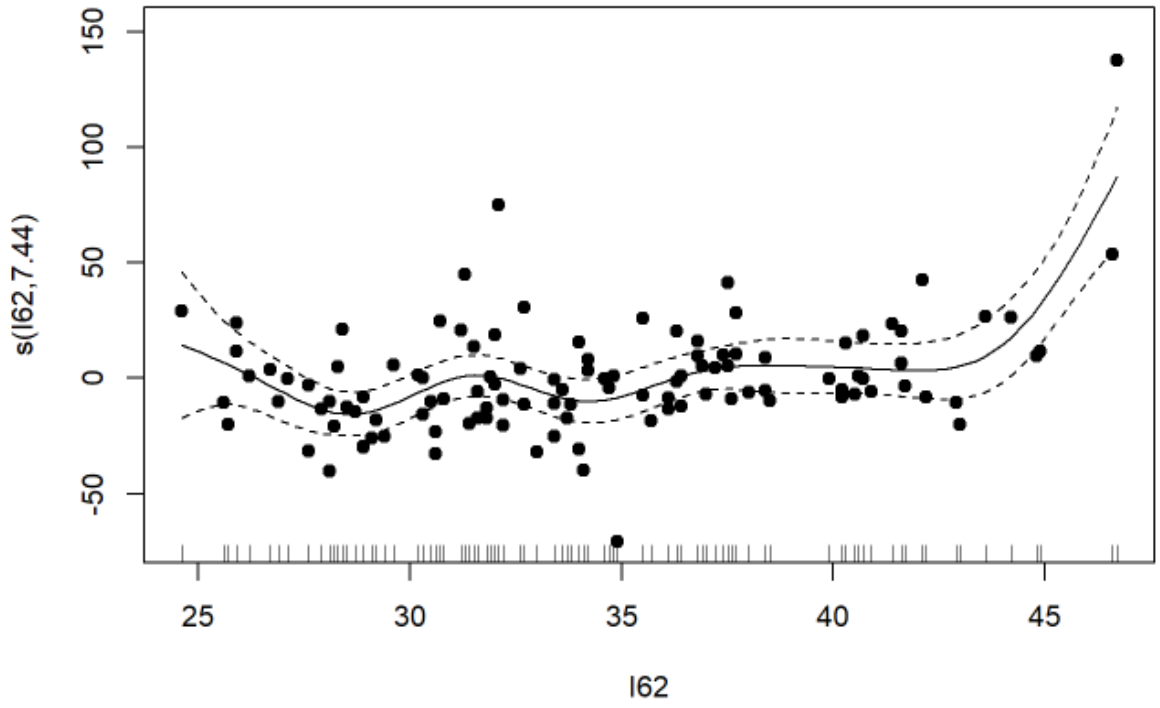

Figure D.2 Regression plot by smooth function for predictor variable $I_{6,2}$ for Q_2 GAM model

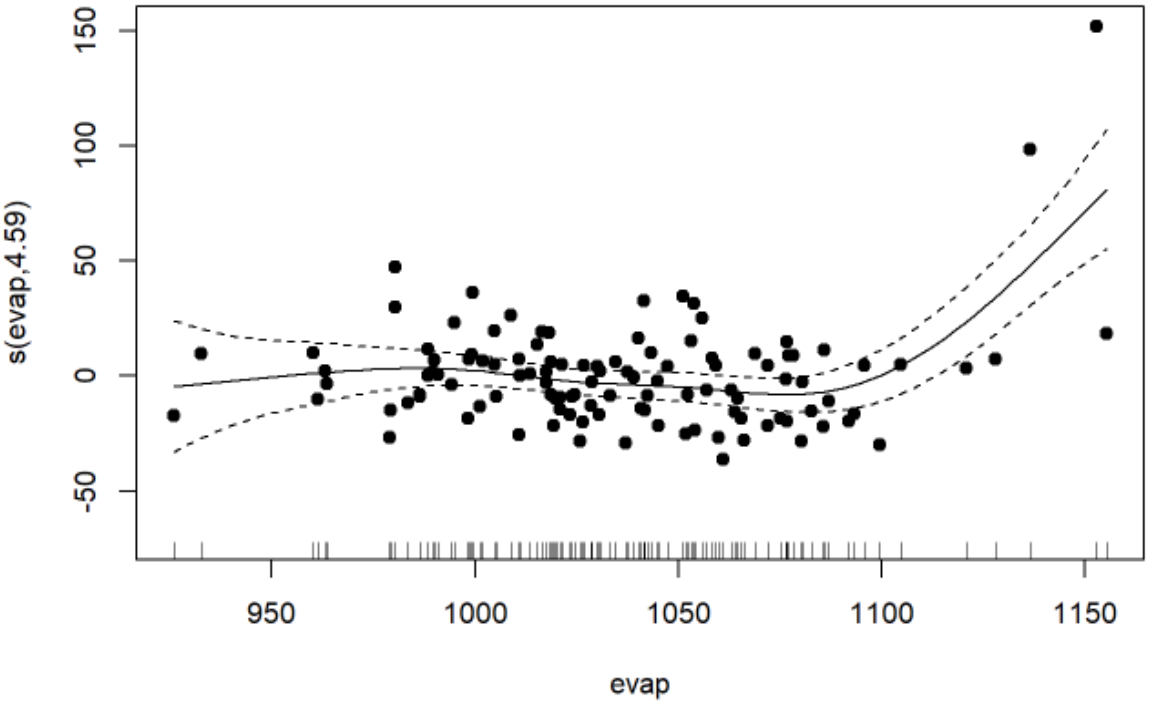

Figure D. 3 Regression plot by smooth function for predictor variable *evap* for Q_2 GAM model

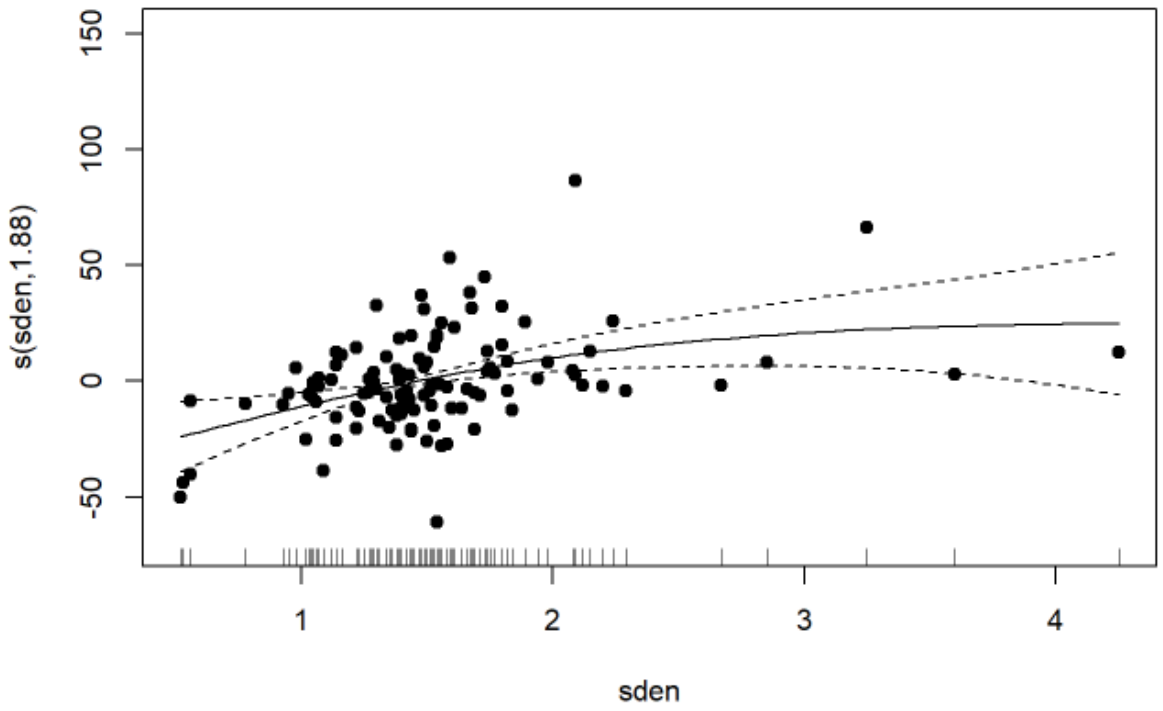

Figure D.4 Regression plot by smooth function for predictor variable *sden* for Q_2 GAM model

Q_5 model diagnostics

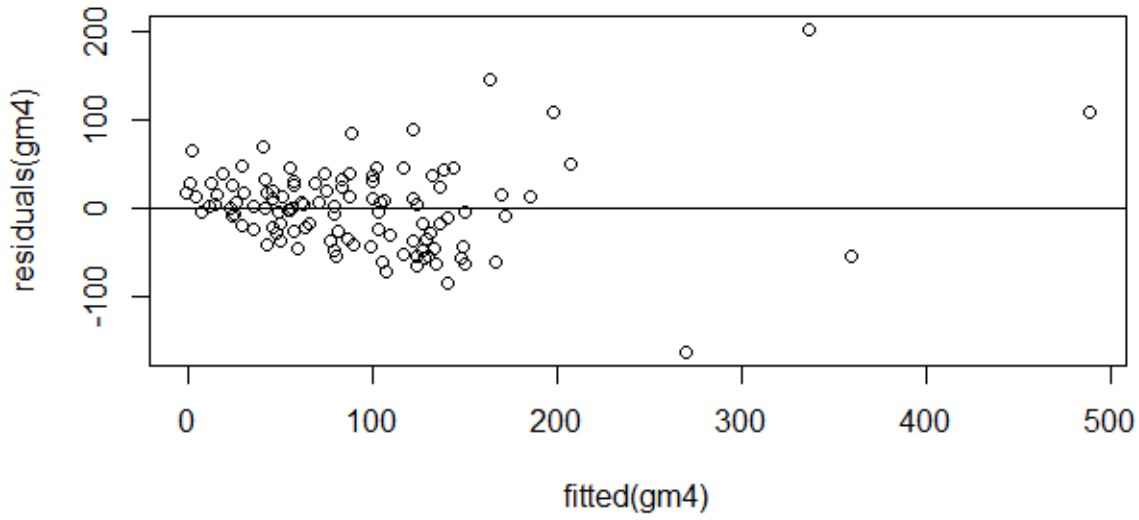

Figure D.5 Standardised residual vs fitted predicted values for the Q_5 GAM model

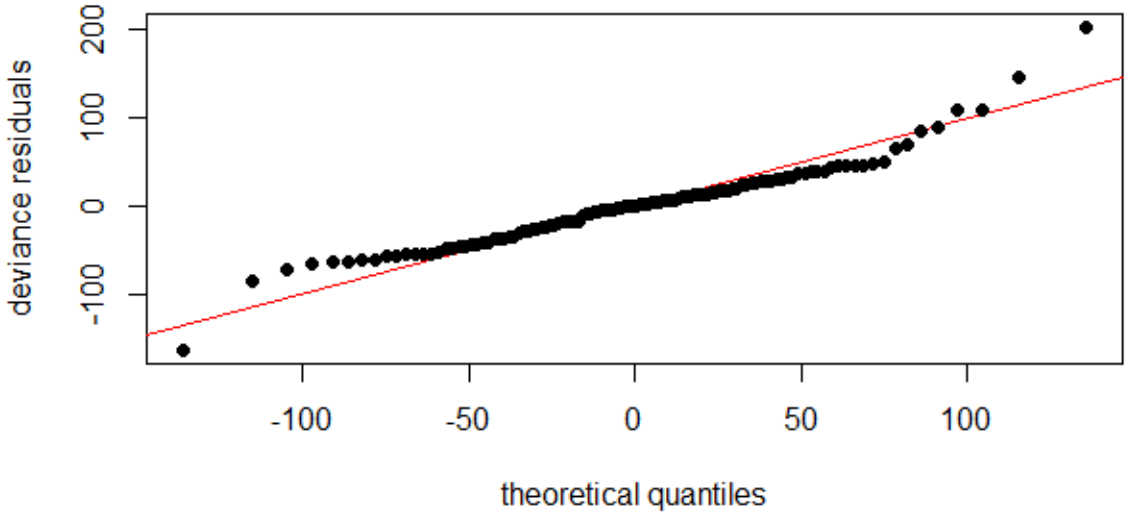

Figure D.6 Normal Q-Q plot of the standardised residuals for the Q_5 GAM model

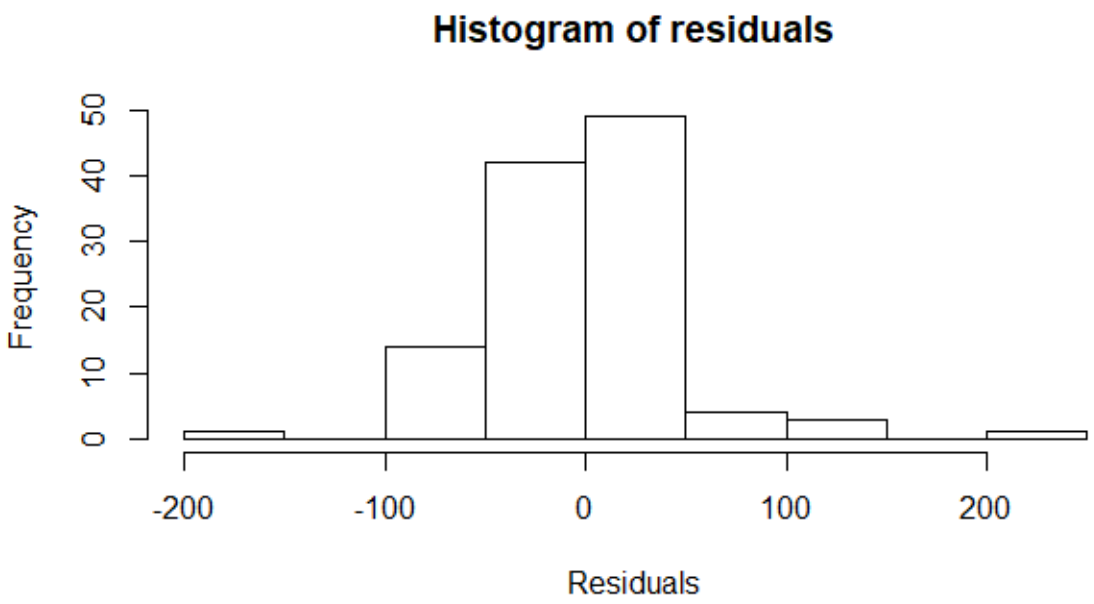

Figure D.7 Histogram of the standardised residuals for Q_5 GAM model

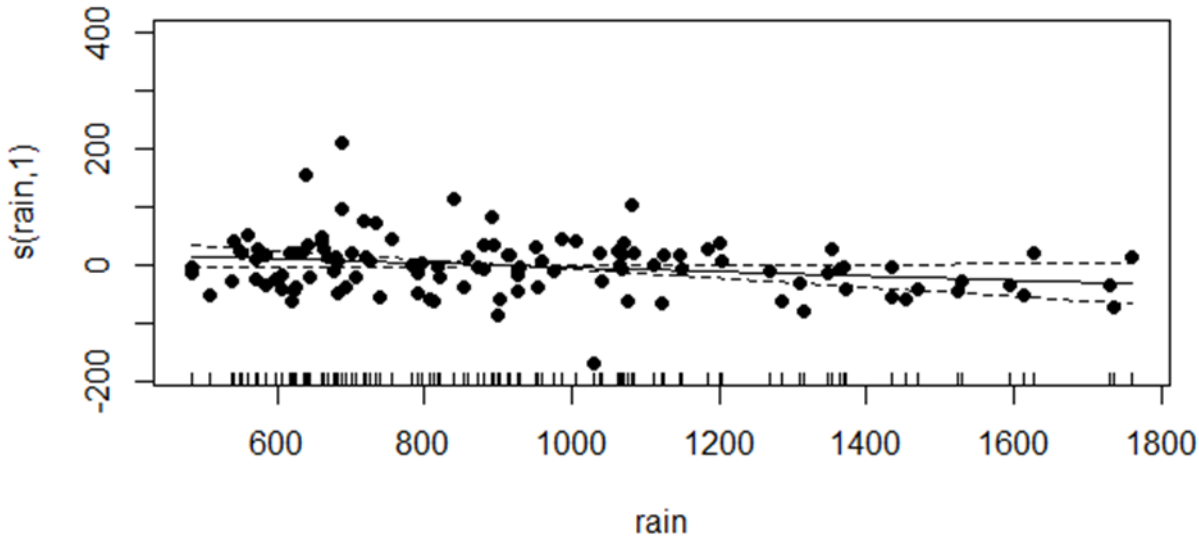

Figure D.8 Regression plot by smooth function for predictor variable *rain* for Q_5 GAM model

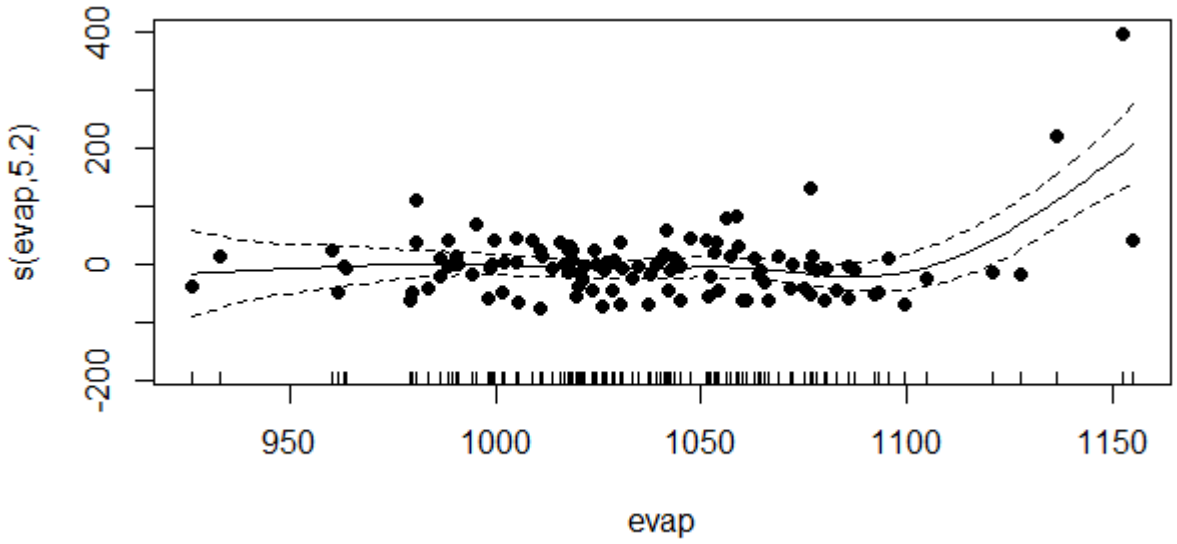

Figure D.9 Regression plot by smooth function for predictor variable *evap* for Q_5 GAM model

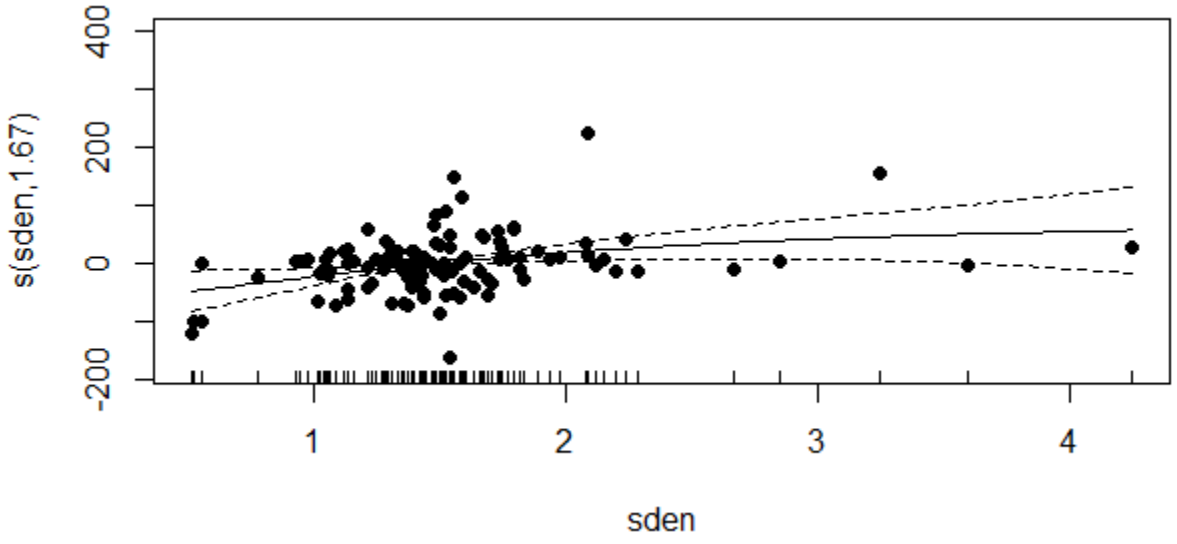

Figure D.10 Regression plot by smooth function for predictor variable *sden* for Q_5 GAM model

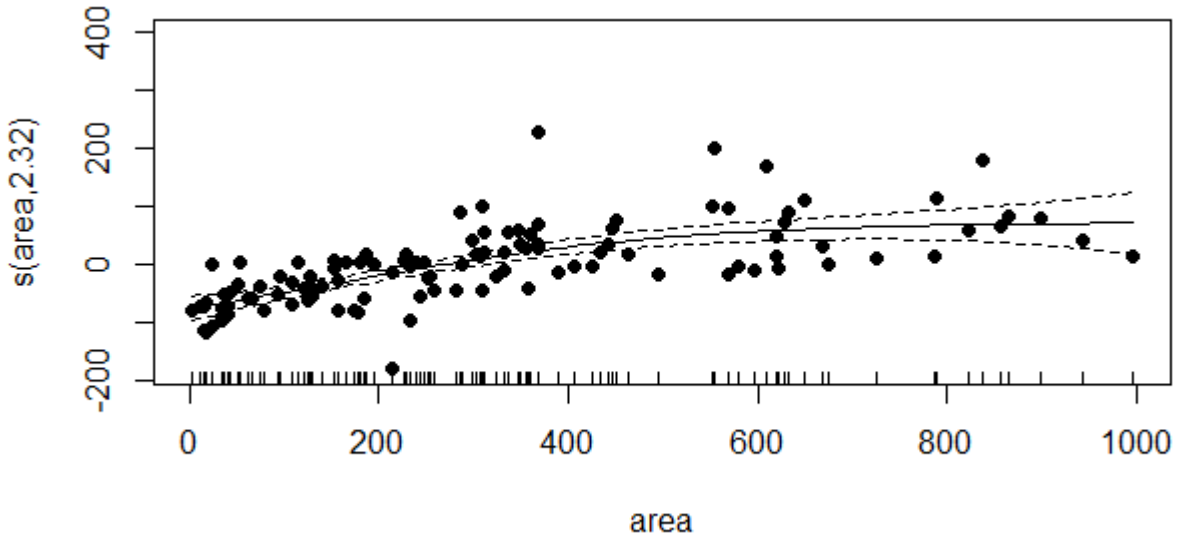

Figure D.11 Regression plot by smooth function for predictor variable *area* for Q_5 GAM model

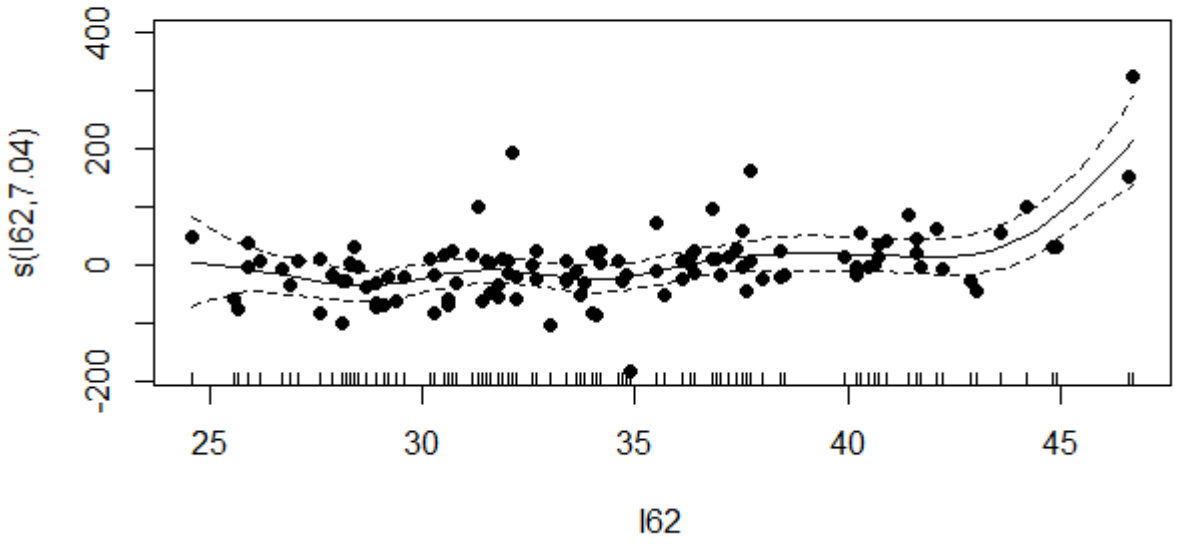

Figure D.12 Regression plot by smooth function for predictor variable $I_{6,2}$ for Q_5 GAM model

Q_{10} model diagnostics

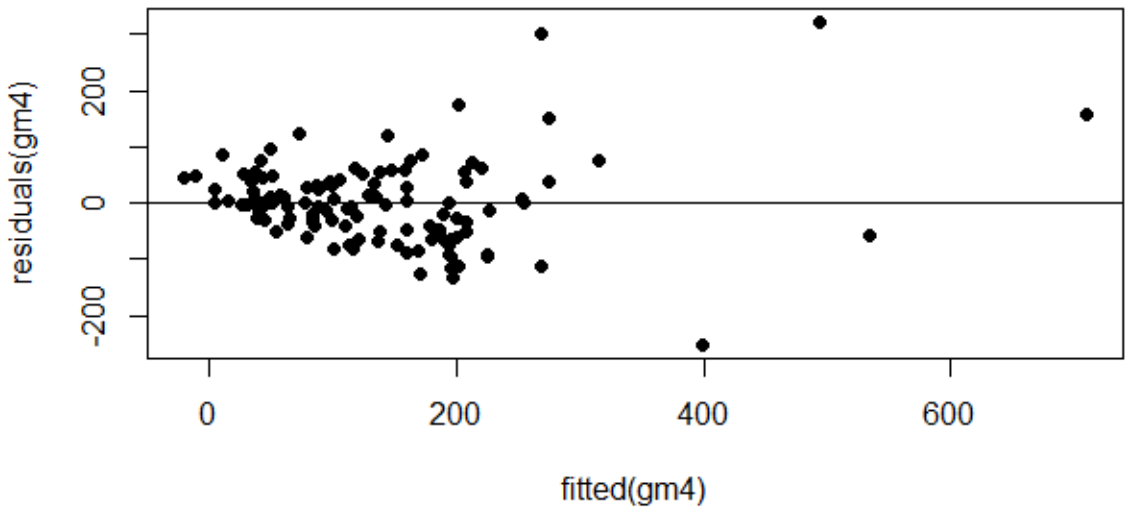

Figure D.13 Standardised residual vs fitted predicted values for the Q_{10} GAM model

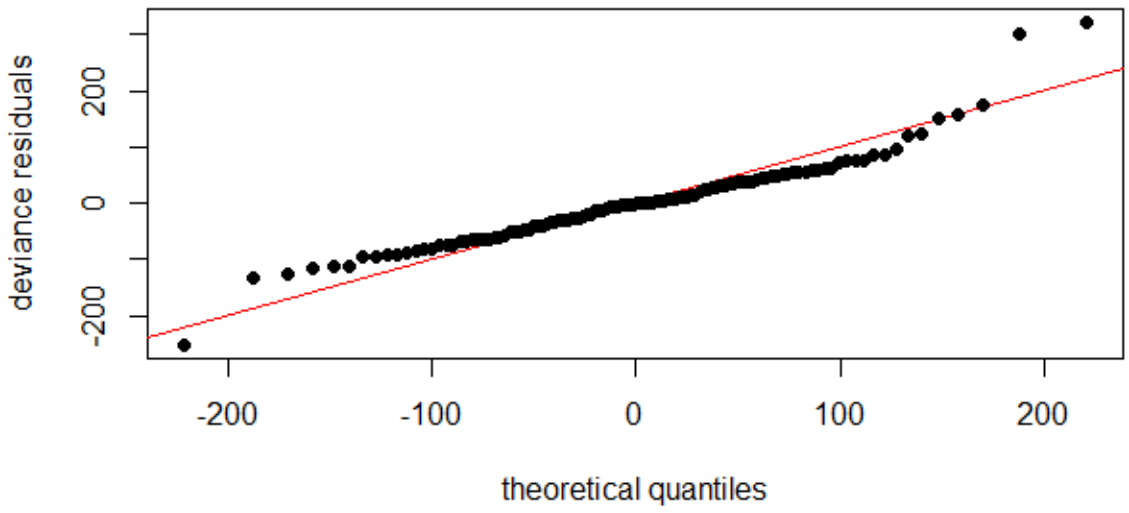

Figure D.14 Normal Q-Q plot of the standardised residuals for the Q_{10} GAM model

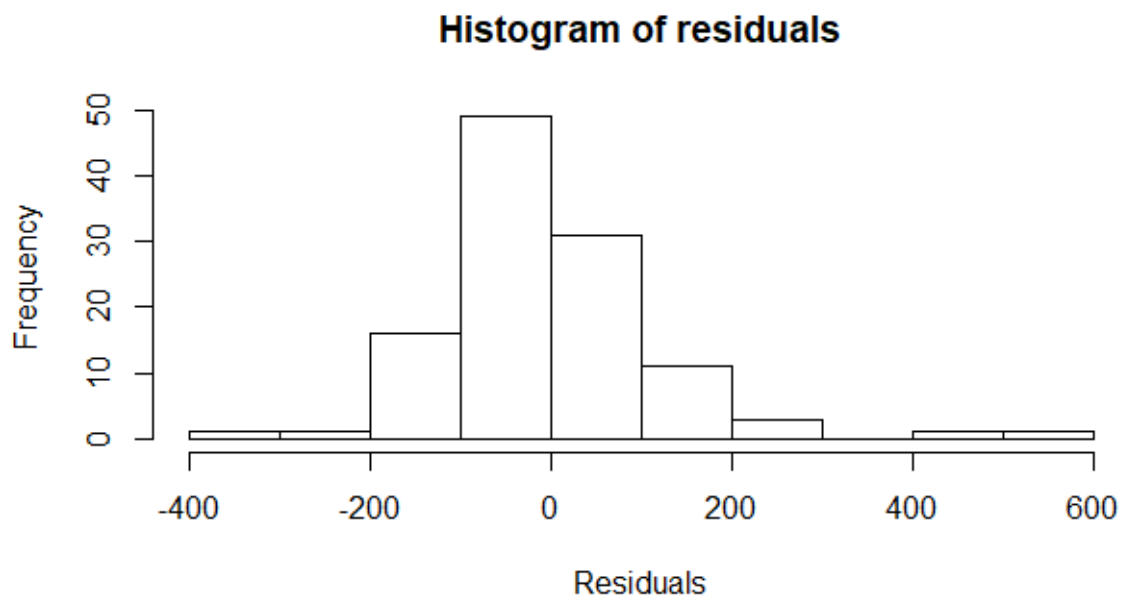

Figure D.15 Histogram of the standardised residuals for Q_{10} GAM model

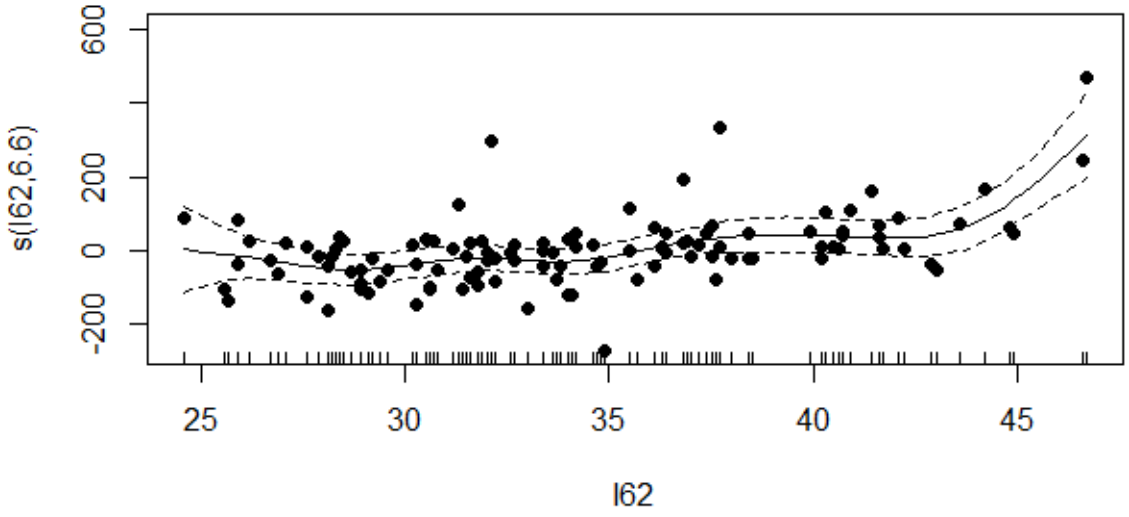

Figure D.16 Regression plot by smooth function for predictor variable $I_{6,2}$ for Q_{10} GAM model

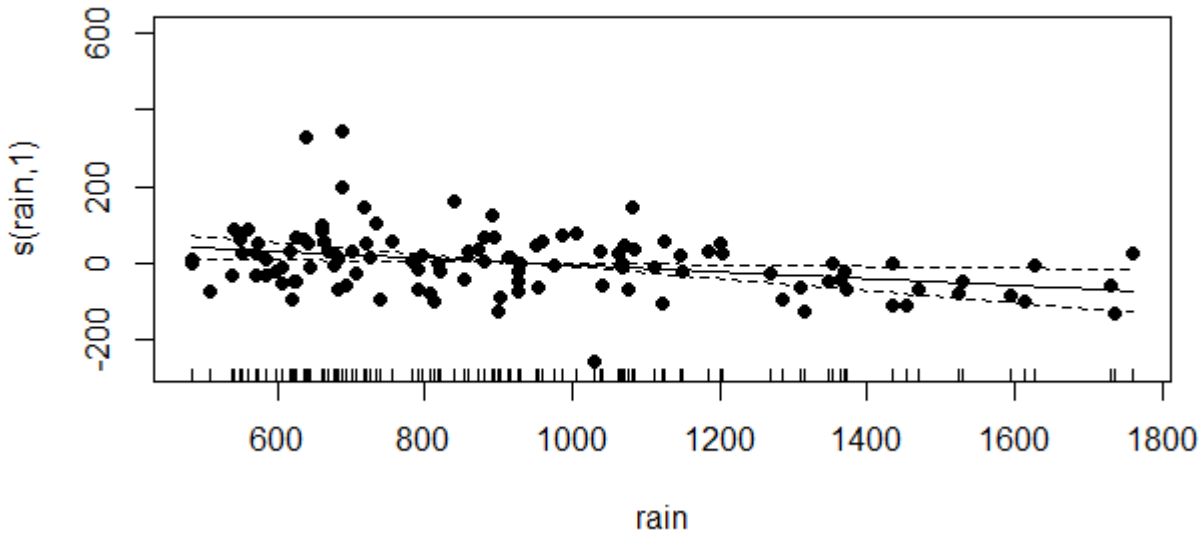

Figure D.17 Regression plot by smooth function for predictor variable *rain* for Q_{10} GAM model

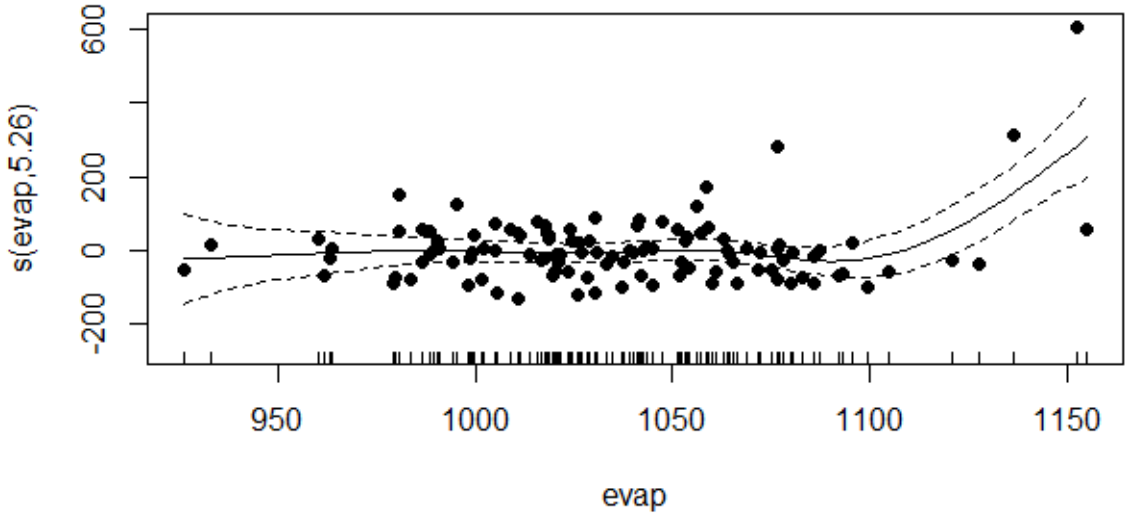

Figure D.18 Regression plot by smooth function for predictor variable *evap* for Q_{10} GAM model

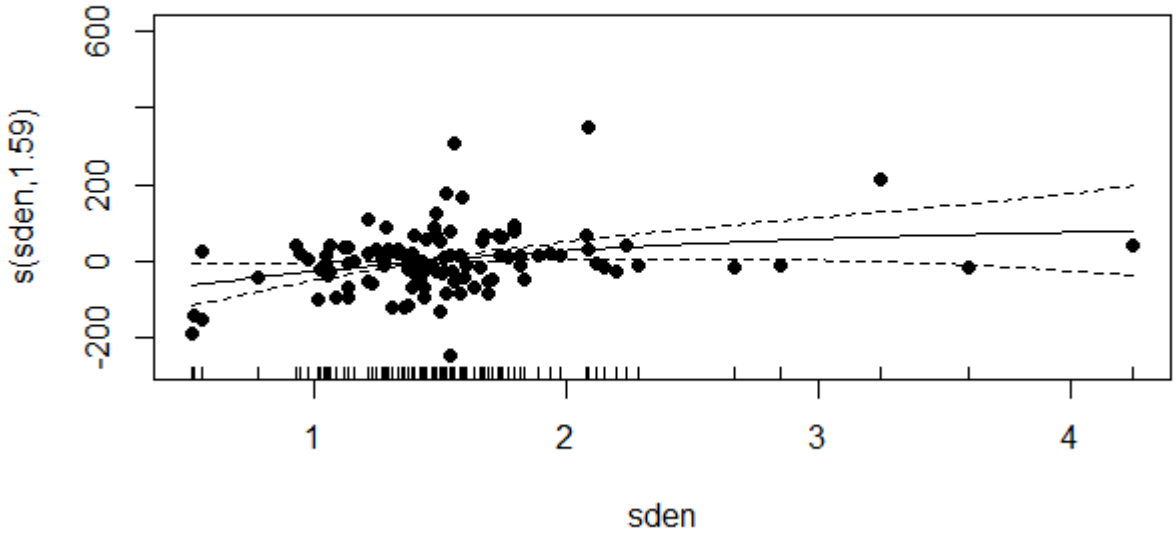

Figure D.19 Regression plot by smooth function for predictor variable *sden* for Q_{10} GAM model

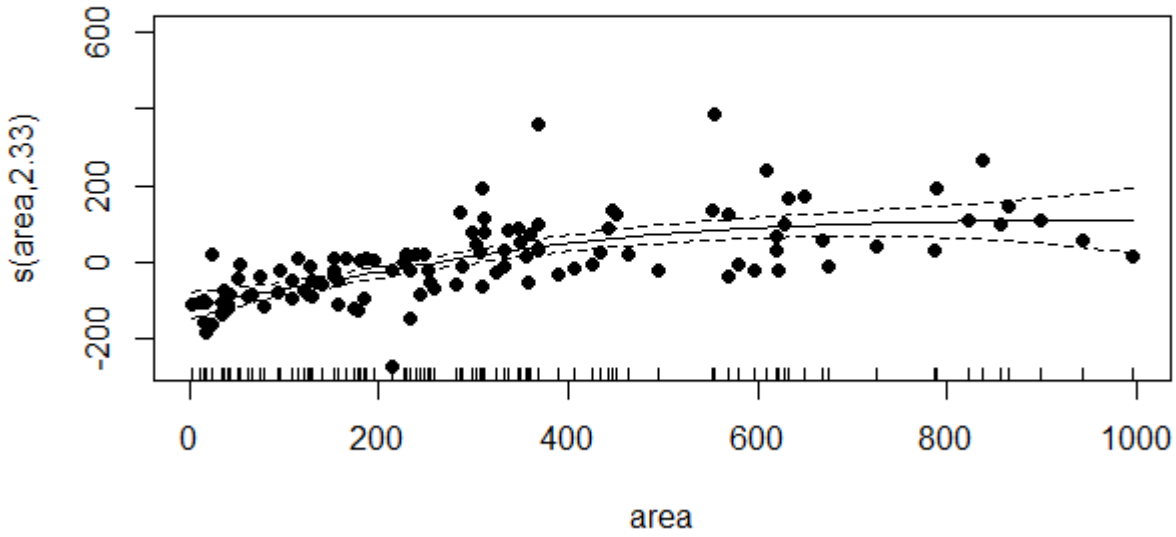

Figure D.20 Regression plot by smooth function for predictor variable *area* for Q_{10} GAM model

Q_{20} model diagnostics

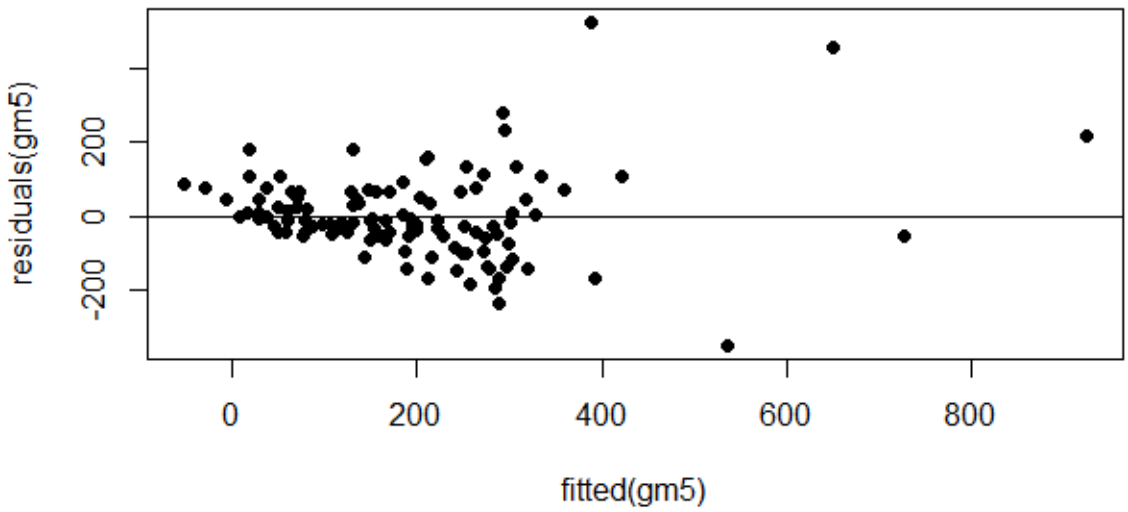

Figure D.21 Standardised residual vs fitted predicted values for the Q_{20} GAM model

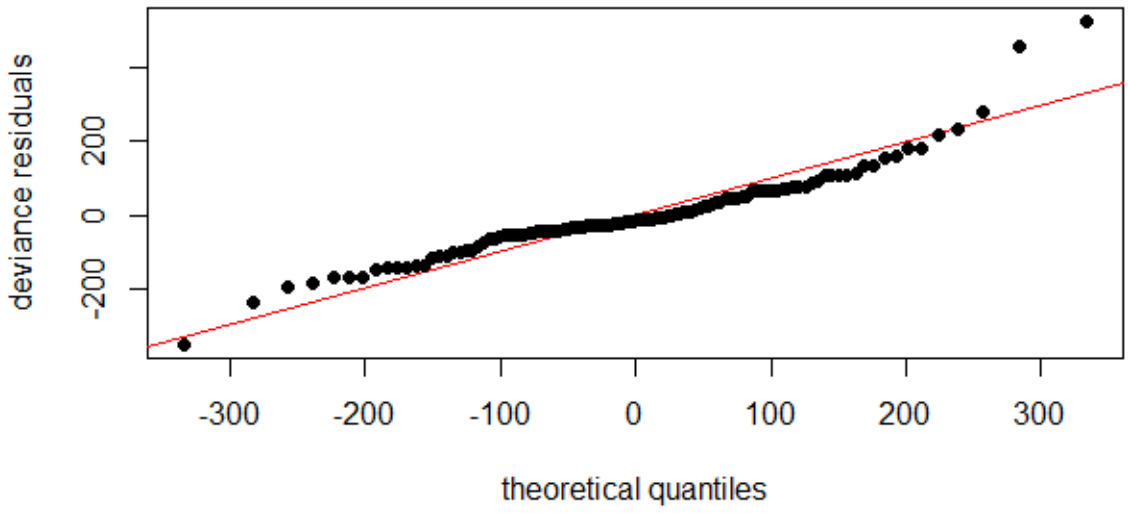

Figure D.22 Normal Q-Q plot of the standardised residuals for the Q_{20} GAM model

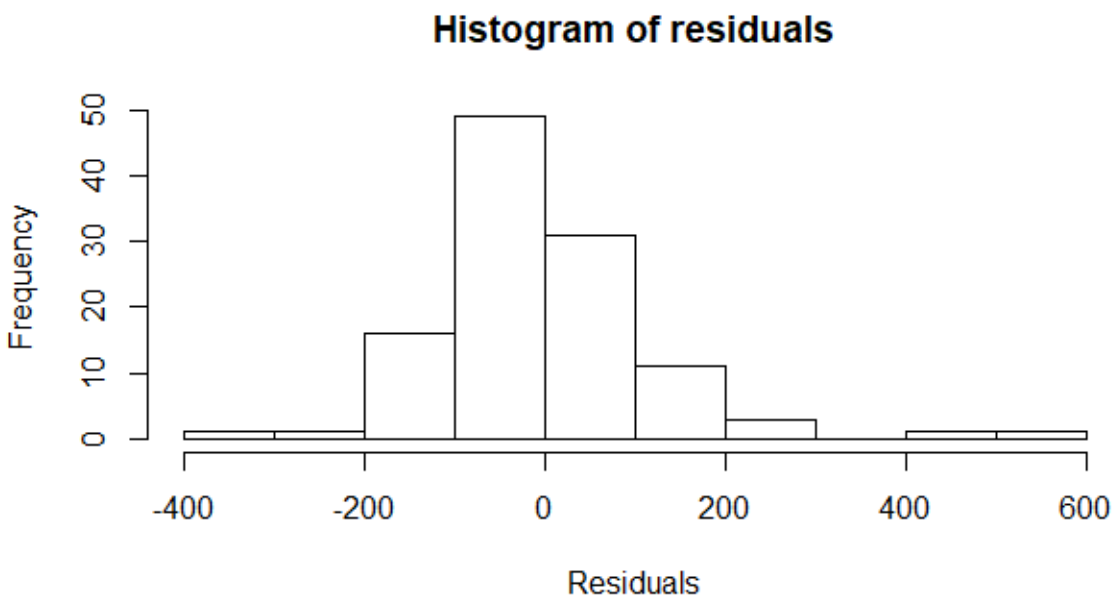

Figure D.23 Histogram of the standardised residuals for Q_{20} GAM model

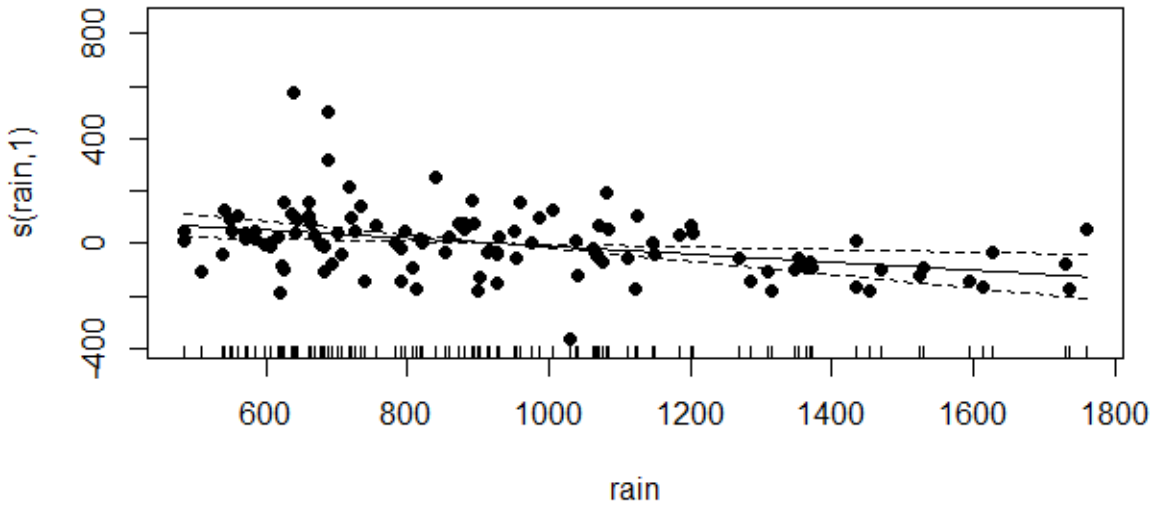

Figure D.24 Regression plot by smooth function for predictor variable *rain* for Q_{20} GAM model

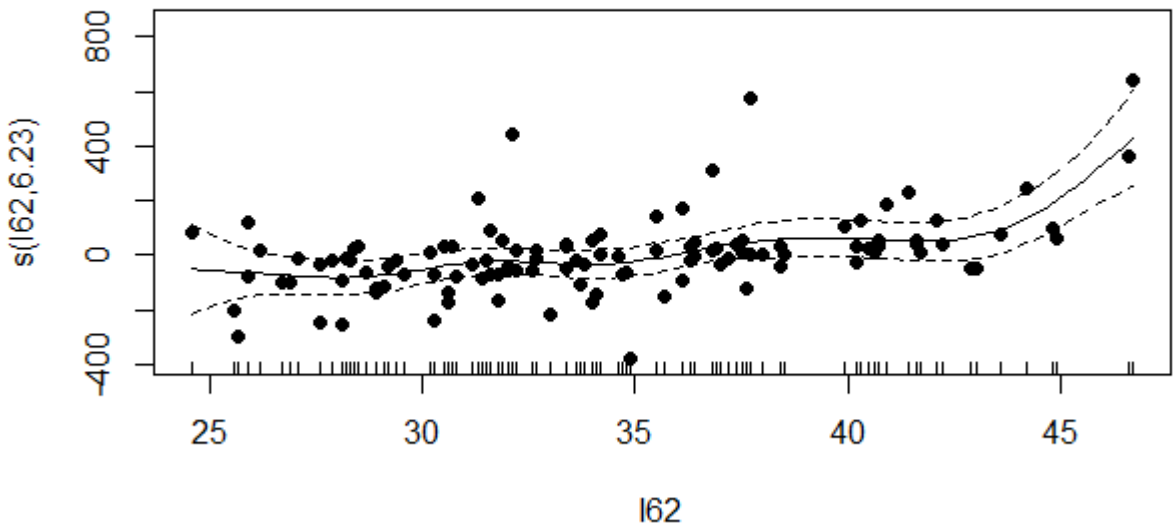

Figure D. 25 Regression plot by smooth function for predictor variable $I_{6,2}$ for Q_{20} GAM model

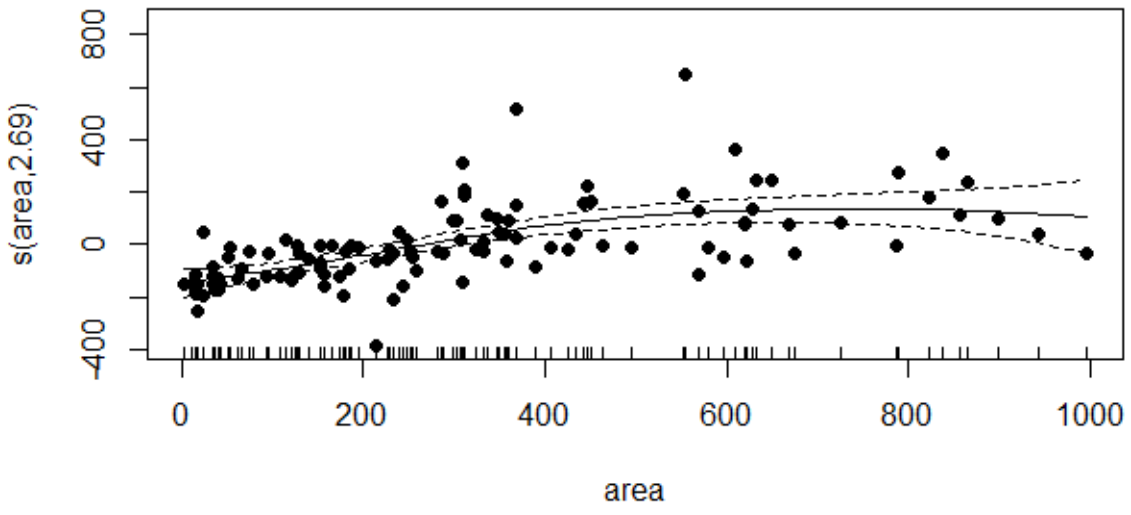

Figure D.26 Regression plot by smooth function for predictor variable *area* for Q_{20} GAM model

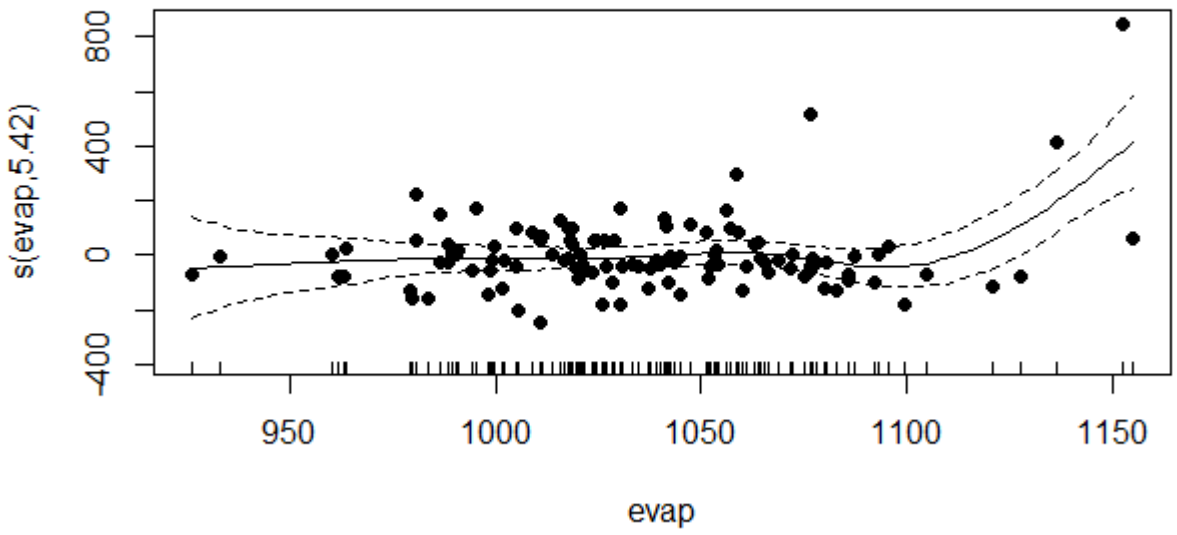

Figure D.27 Regression plot by smooth function for predictor variable *evap* for Q_{20} GAM model

Q_{50} model diagnostics

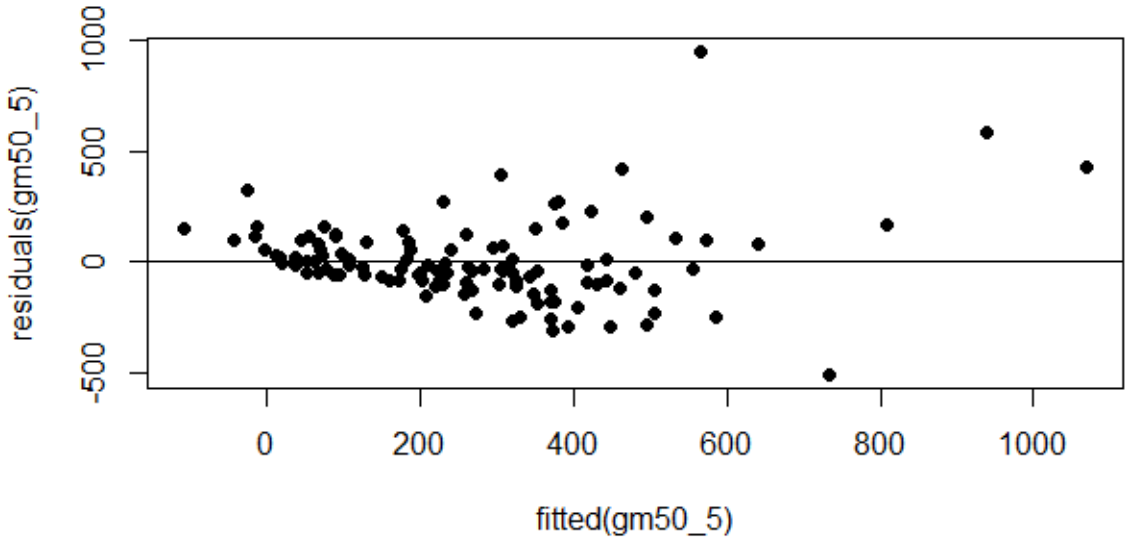

Figure D.28 Standardised residual vs fitted predicted values for the Q_{50} GAM model

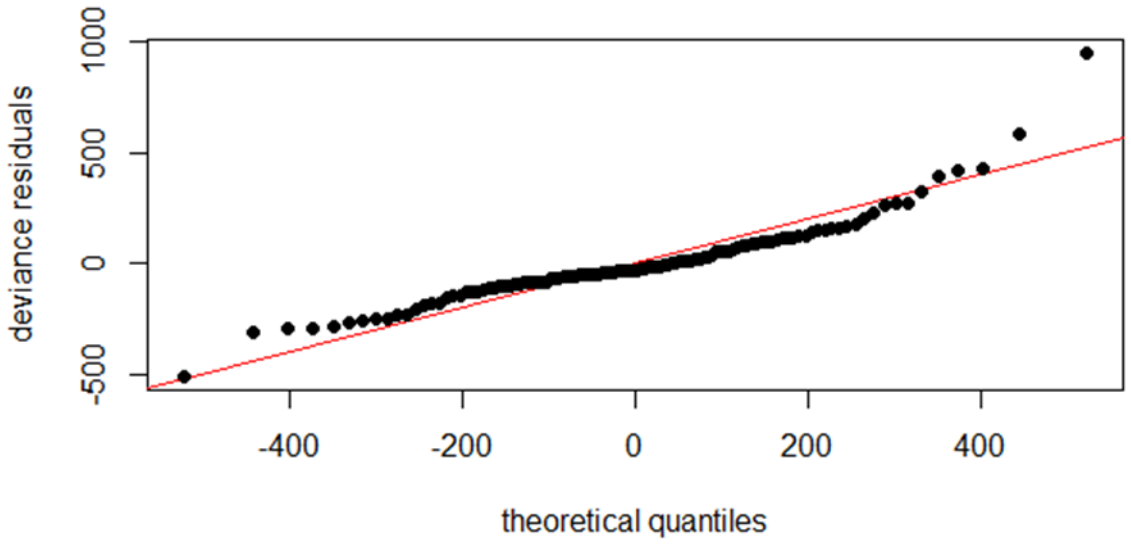

Figure D.29 Normal Q-Q plot of the standardised residuals for the Q_{50} GAM model

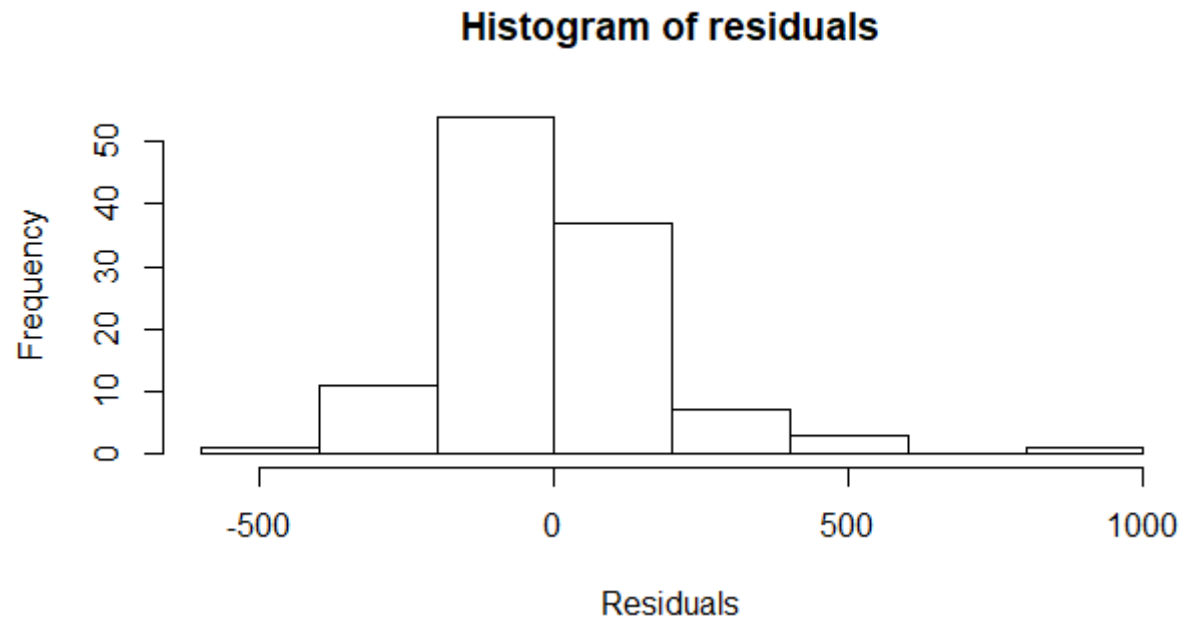

Figure D.30 Histogram of the standardised residuals for Q_{50} GAM model

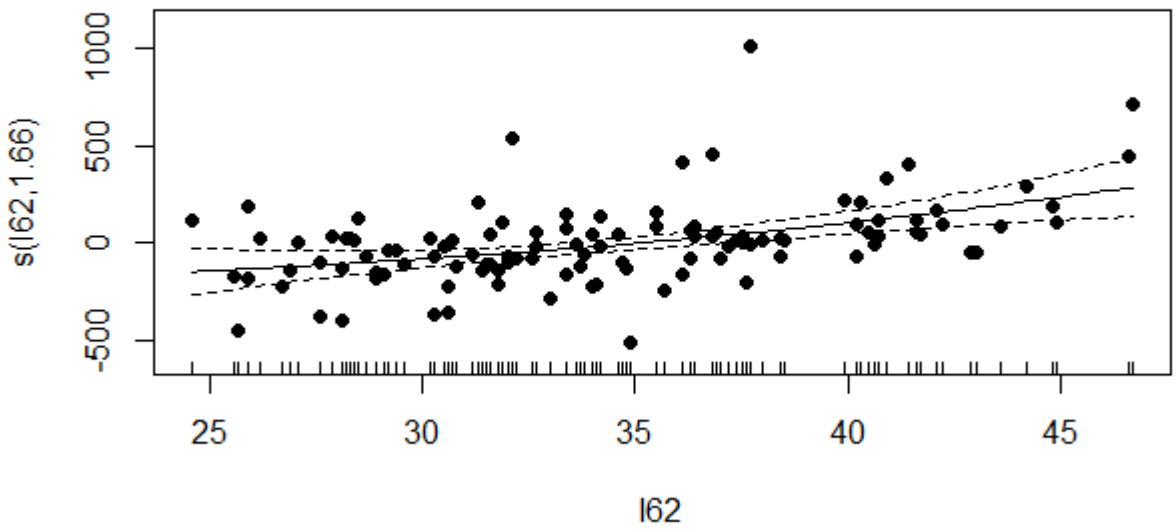

Figure D.31 Regression plot by smooth function for predictor variable $I_{6,2}$ for Q_{50} GAM model

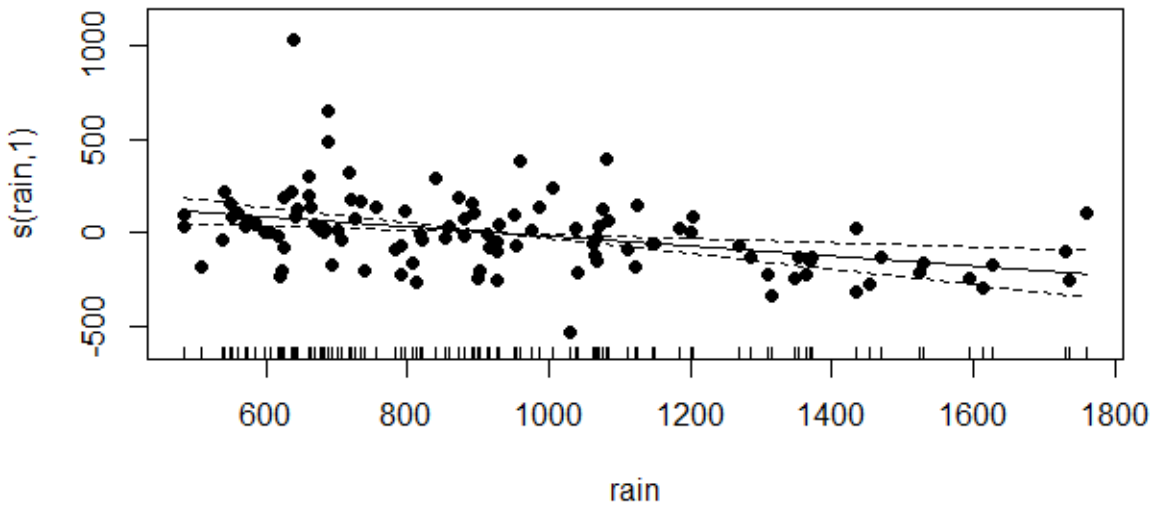

Figure D.32 Regression plot by smooth function for predictor variable *rain* for Q_{50} GAM model

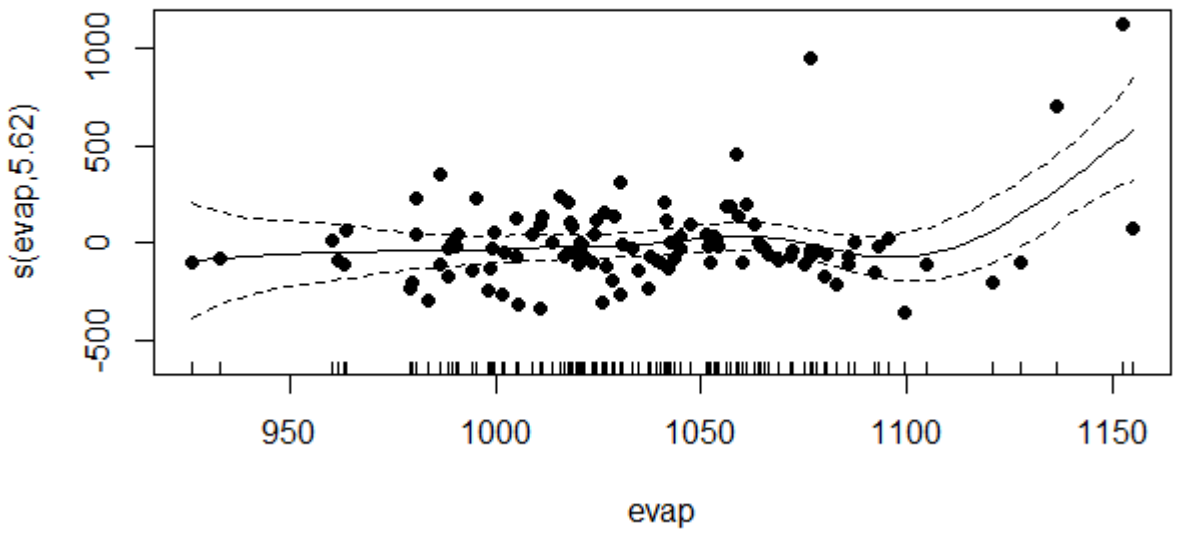

Figure D. 33 Regression plot by smooth function for predictor variable *evap* for Q_{50} GAM model

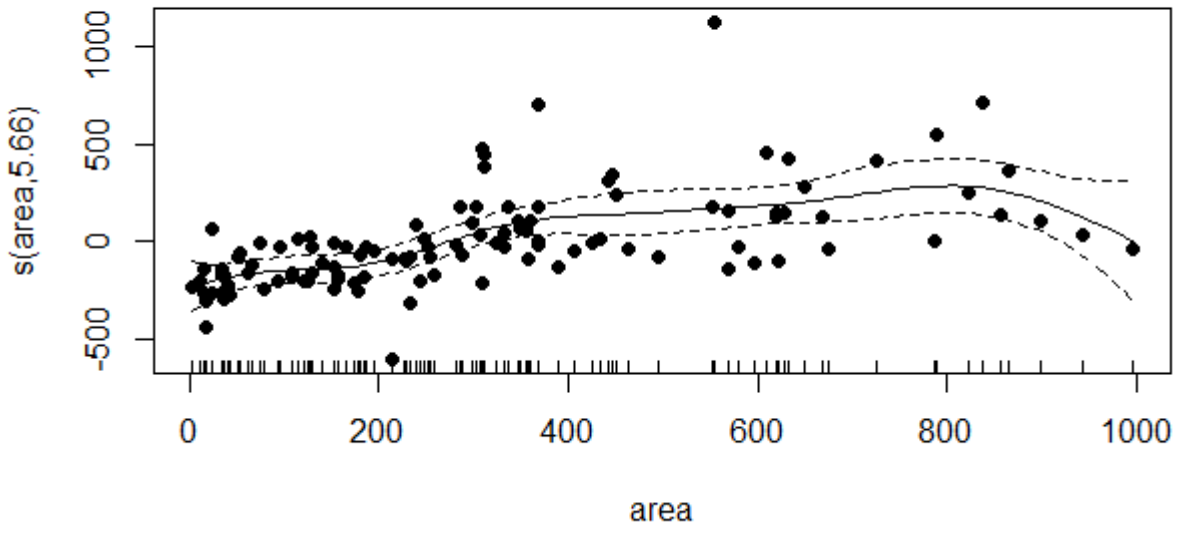

Figure D. 34 Regression plot by smooth function for predictor variable *area* for Q_{50} GAM model

Q_{100} model diagnostics

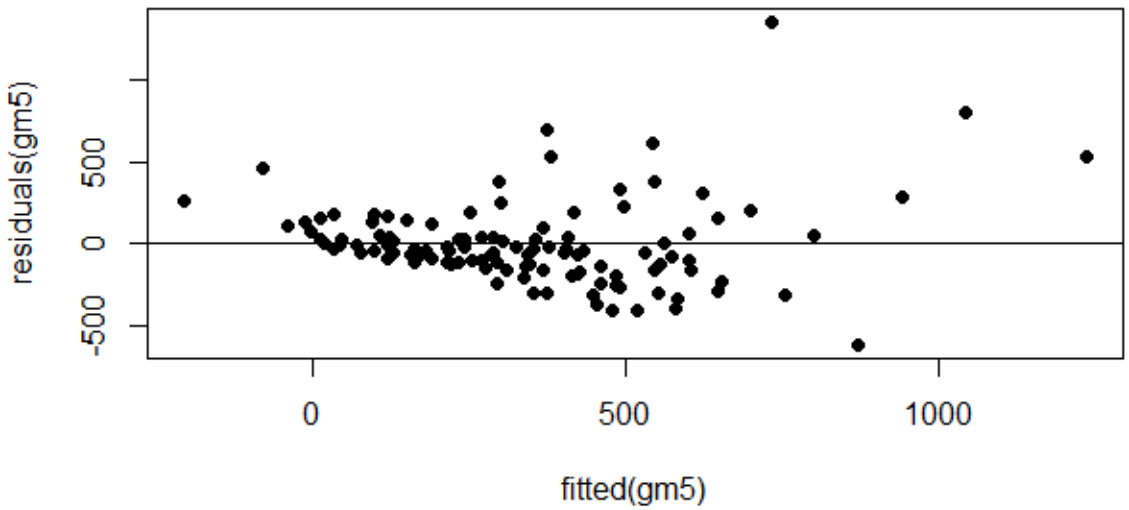

Figure D.35 Standardised residual vs fitted predicted values for the Q_{100} GAM model

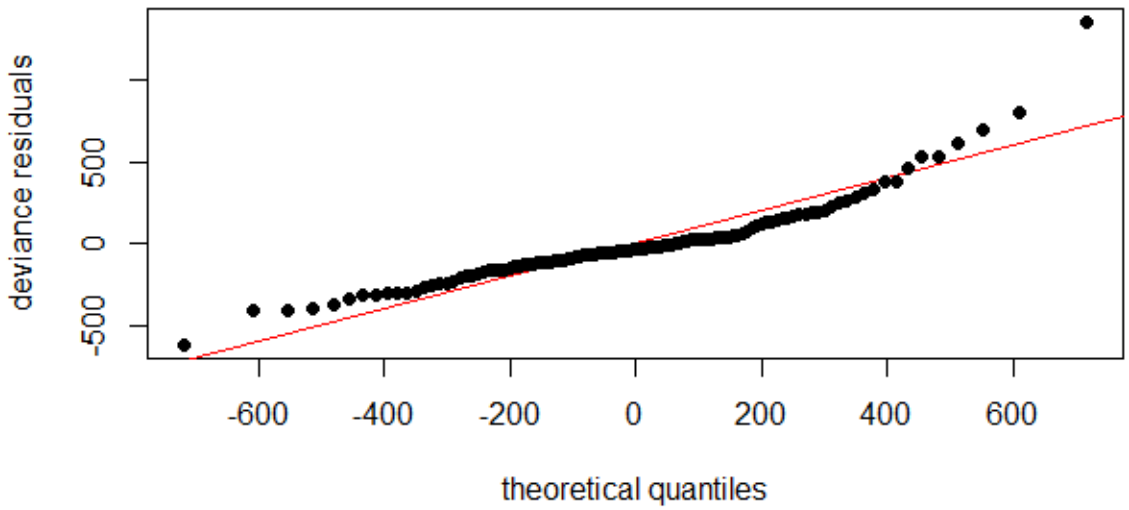

Figure D.36 Normal Q-Q plot of the standardised residuals for the Q_{100} GAM model

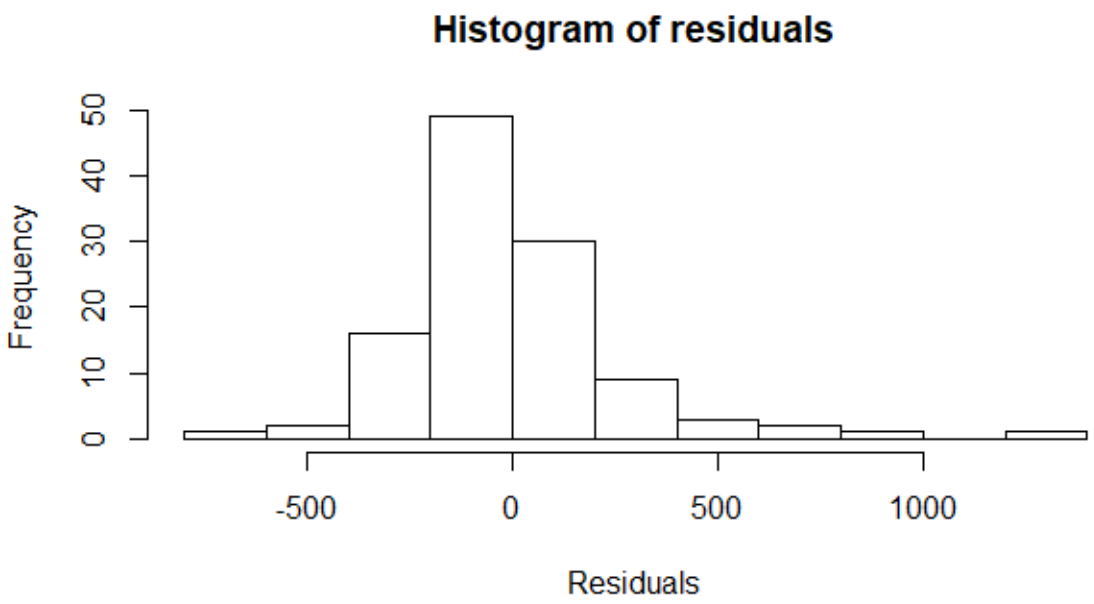

Figure D.37 Histogram of the standardised residuals for Q_{50} GAM model

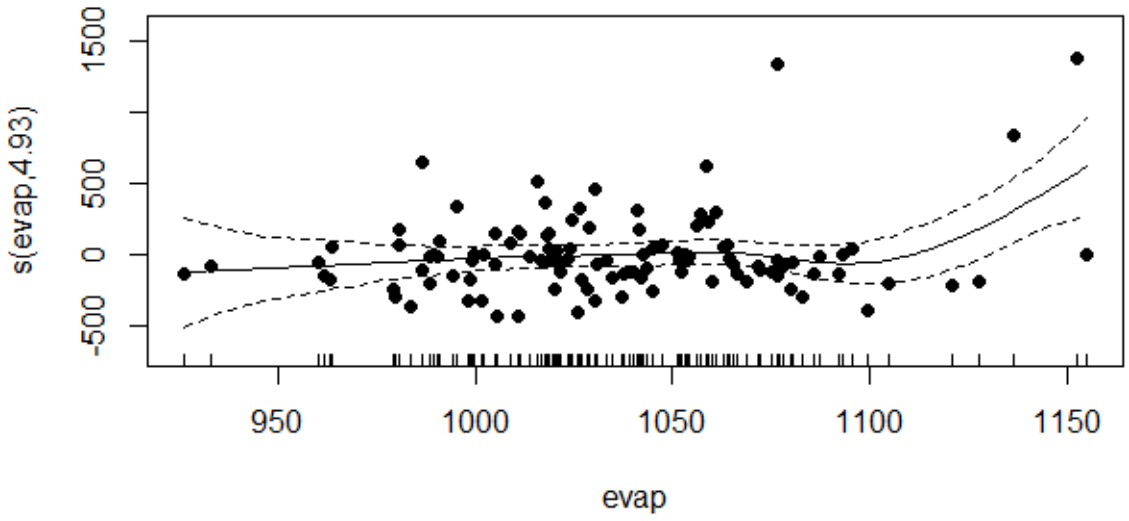

Figure D.38 Regression plot by smooth function for predictor variable *evap* for Q_{100} GAM model

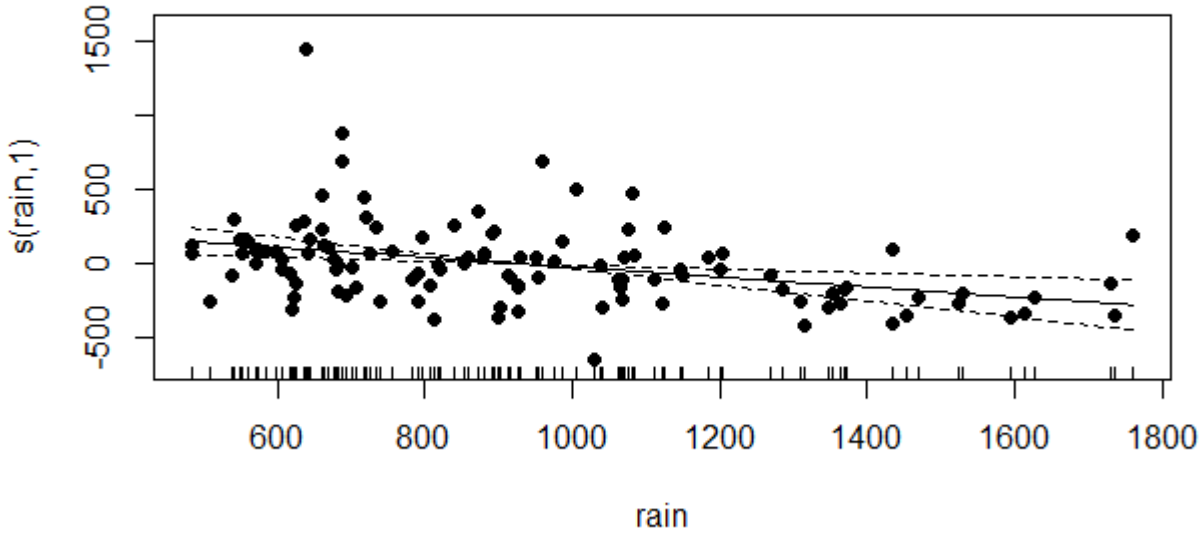

Figure D. 39 Regression plot by smooth function for predictor variable *rain* for Q_{100} GAM model

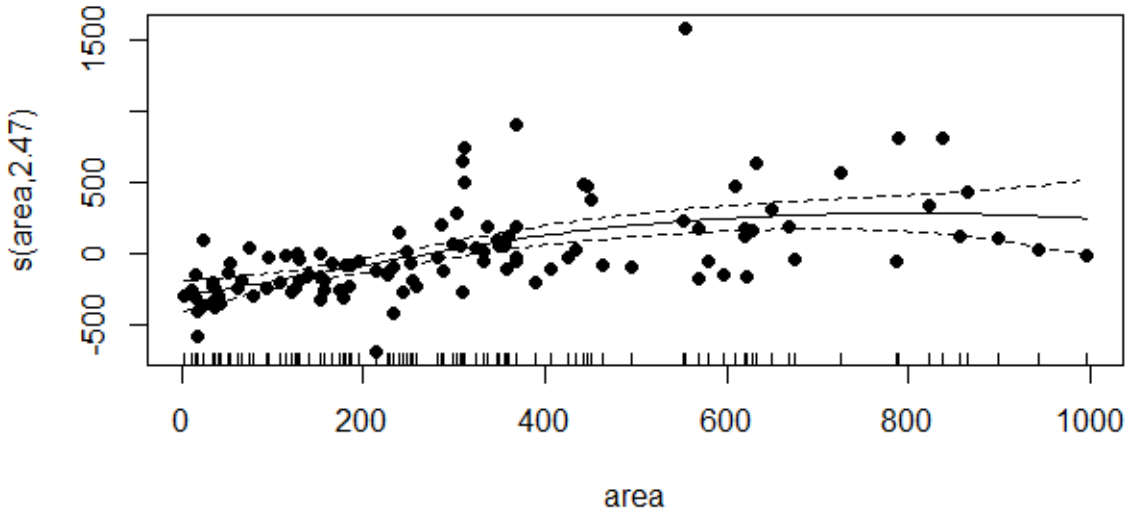

Figure D. 40 Regression plot by smooth function for predictor variable *area* for Q_{100} GAM model

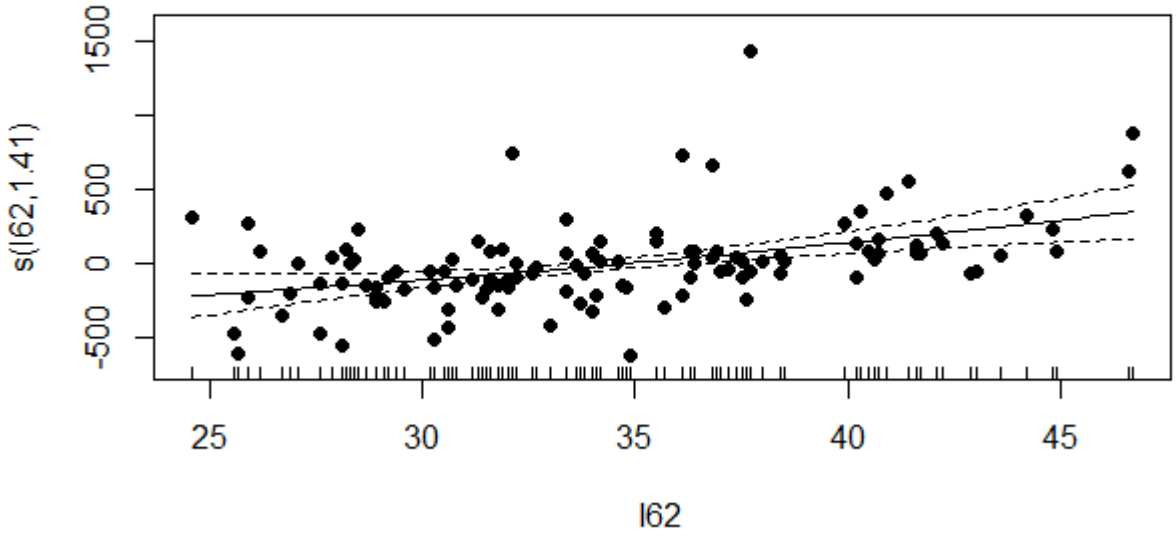

Figure D.41 Regression plot by smooth function for predictor variable $I_{6,2}$ for Q_{100} GAM model

APPENDIX E

Additional results from GAM models (scatter plot of Q_{obs} vs Q_{pred})

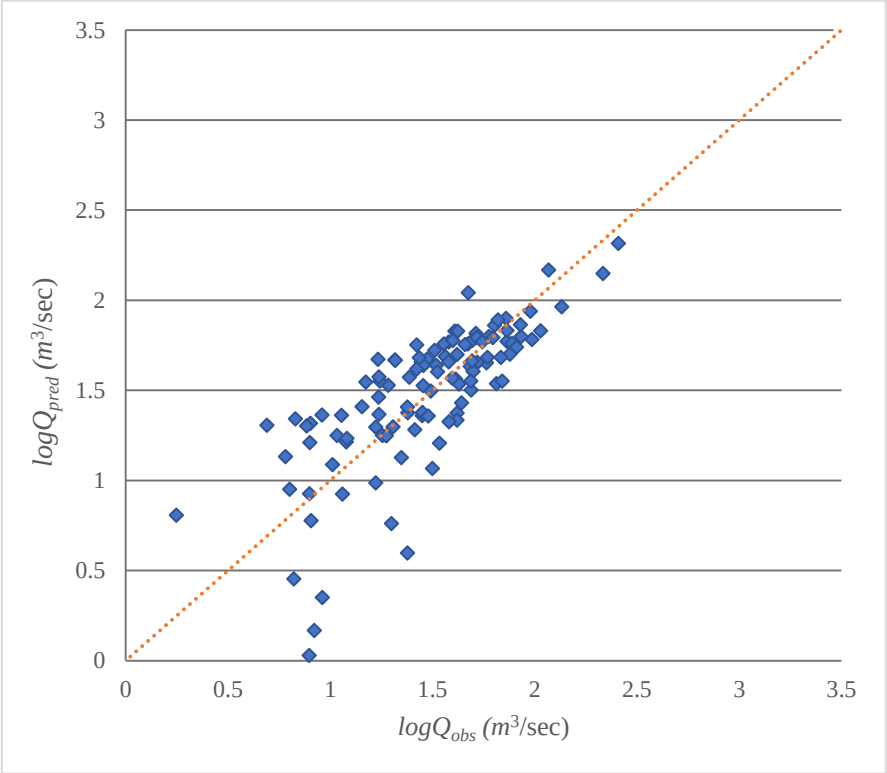

Figure E.1 Comparison of observed and predicted flood quantiles for GAM model for Q_2

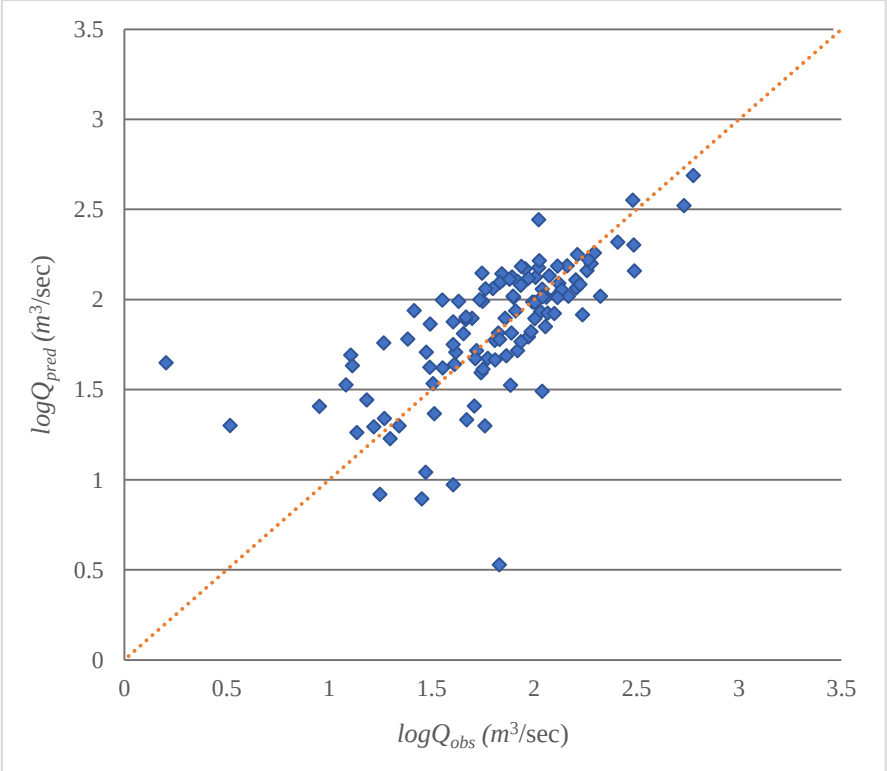

Figure E.2 Comparison of observed and predicted flood quantiles for GAM model for Q_5

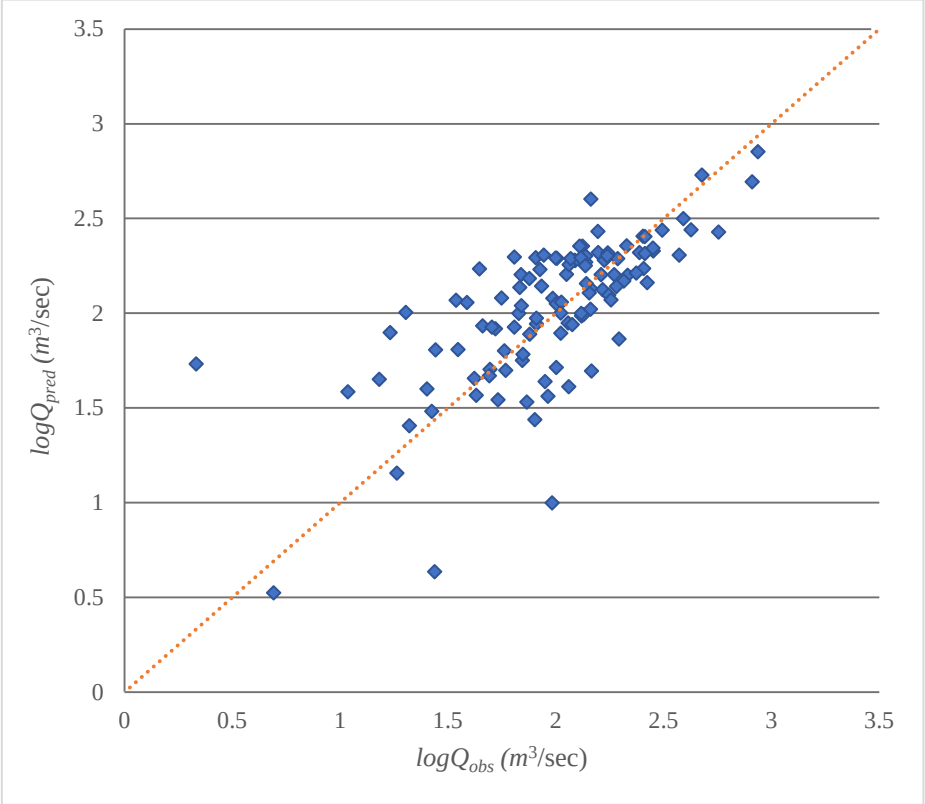

Figure E.3 Comparison of observed and predicted flood quantiles for GAM model for Q_{10}

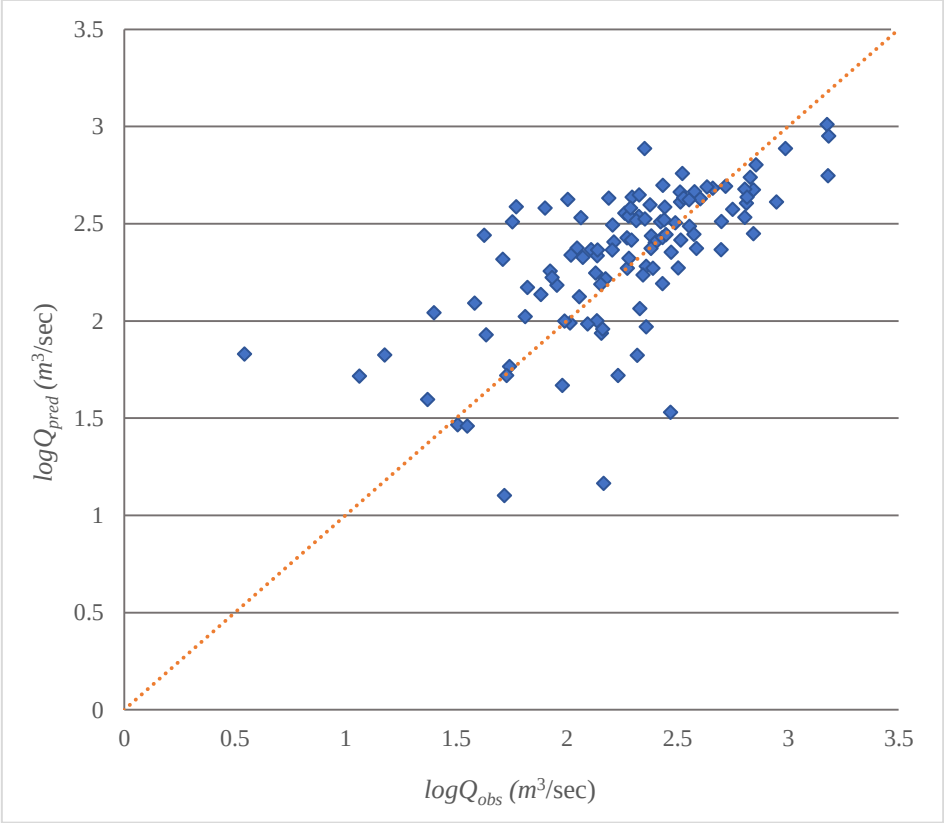

Figure E.4 Comparison of observed and predicted flood quantiles for GAM model for Q_{50}

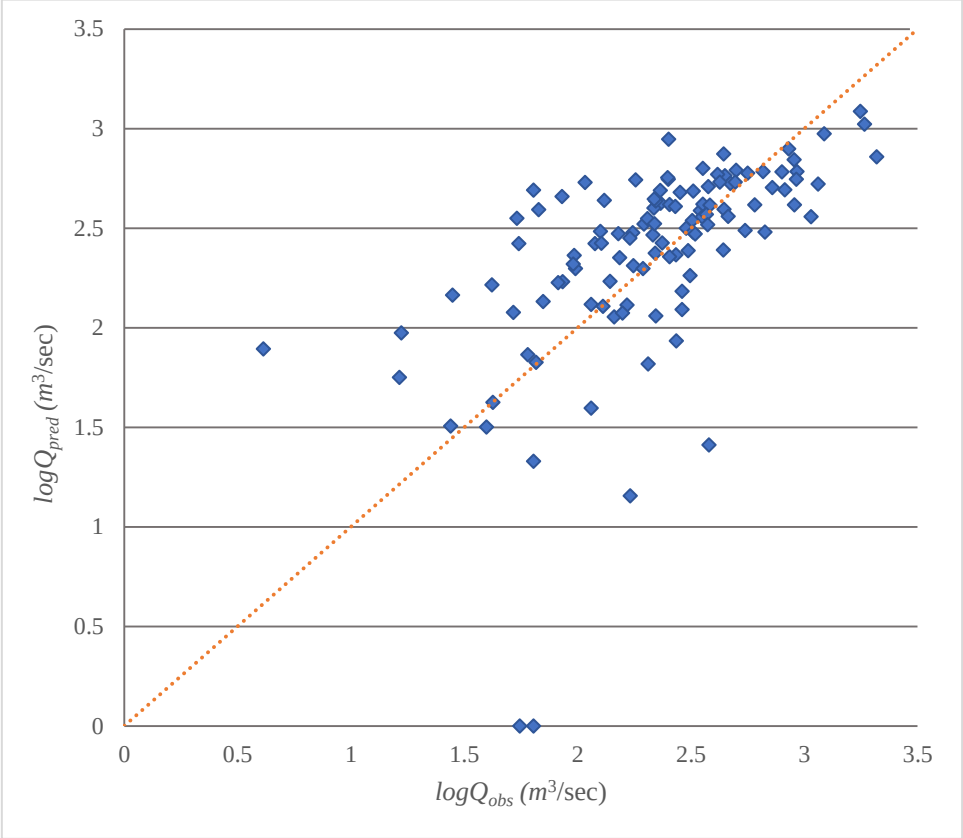

Figure E.5 Comparison of observed and predicted flood quantiles for GAM model for Q_{100}

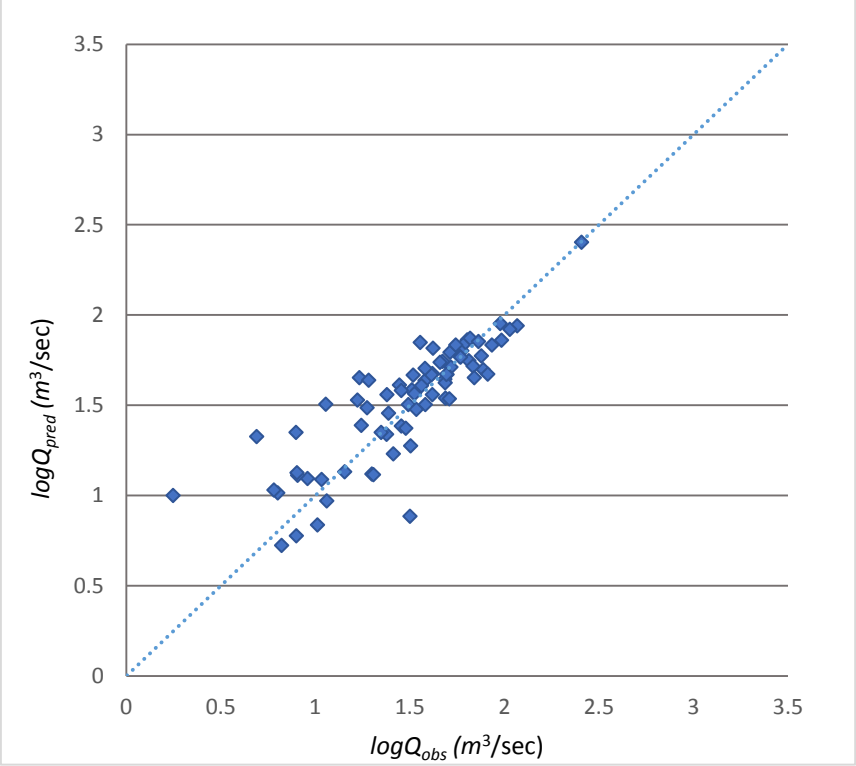

Figure E.6 Comparison of observed and predicted flood quantiles for GAM model for Q_2(A1 group)

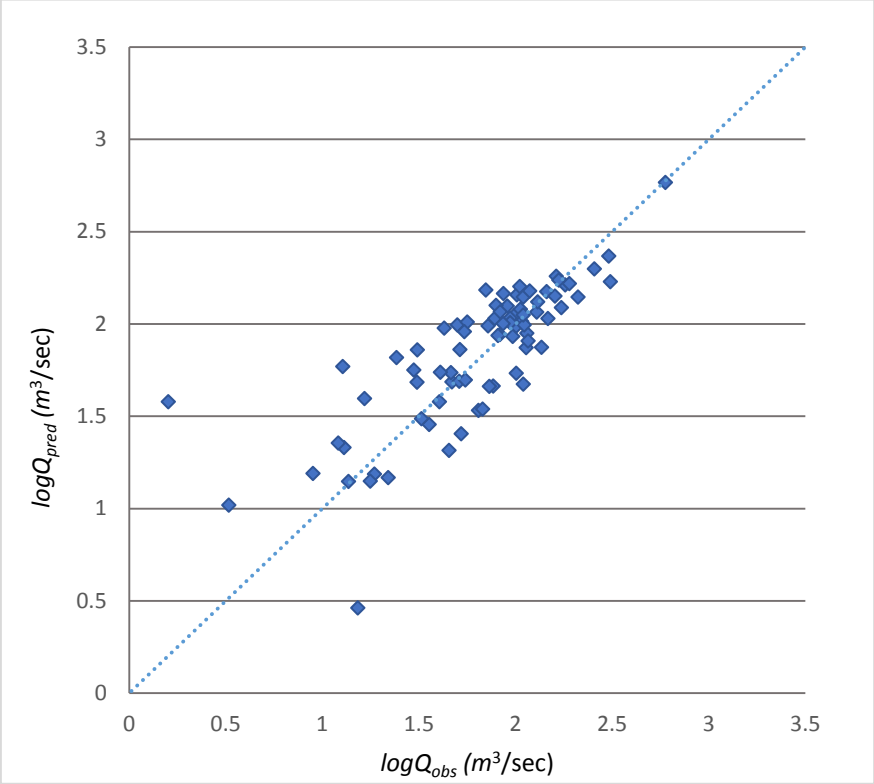

Figure E.7 Comparison of observed and predicted flood quantiles for GAM based RFFA model for Q_5(A1 group)

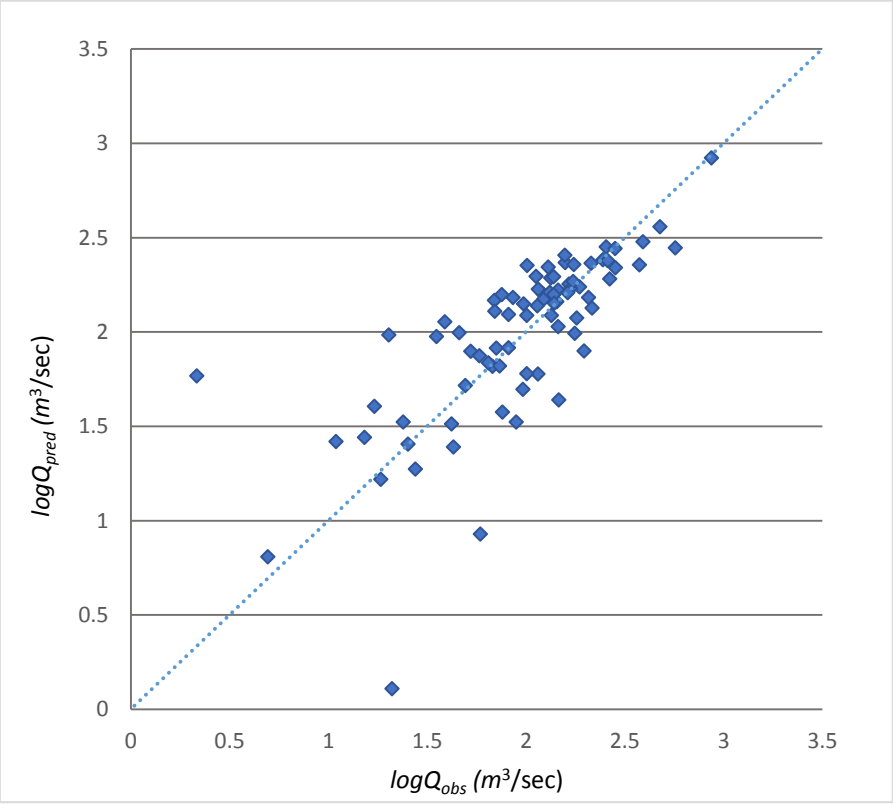

Figure E.8 Comparison of observed and predicted flood quantiles for GAM model for Q_{10}(A1 group)

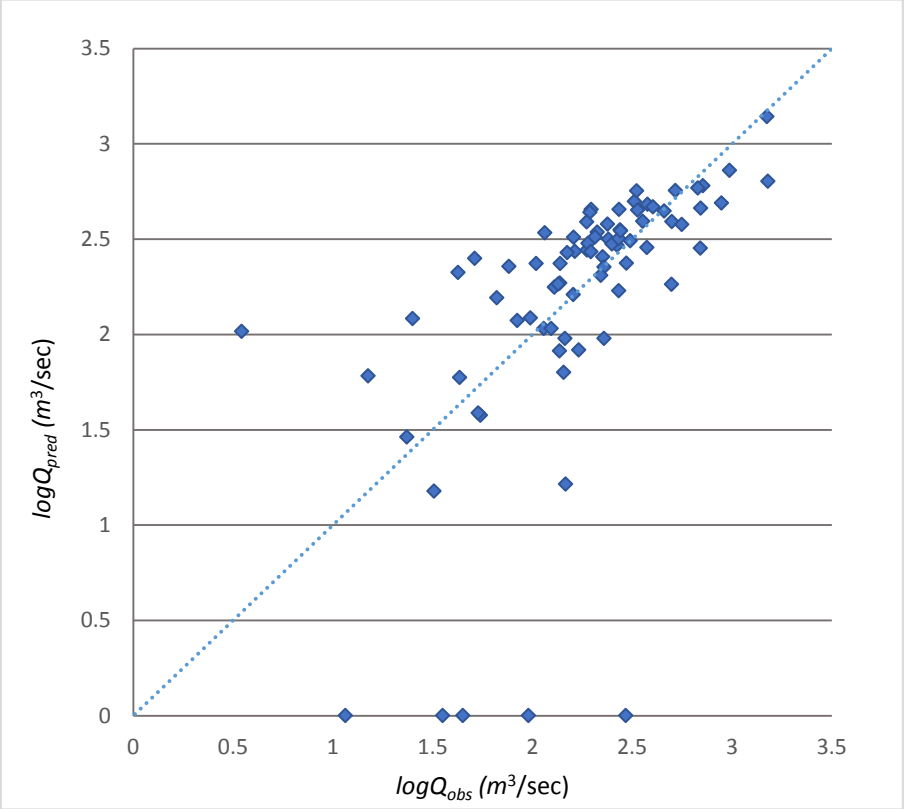

Figure E.9 Comparison of observed and predicted flood quantiles for GAM model for Q_{50}(A1 group)

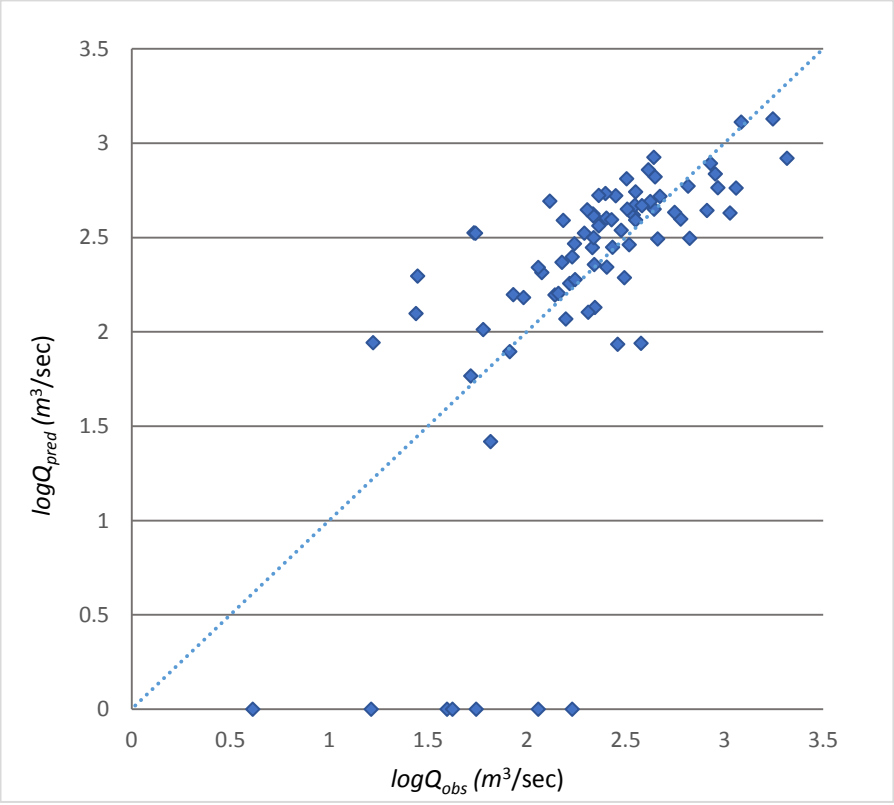

Figure E.10 Comparison of observed and predicted flood quantiles for GAM model for Q_{100}(A1 group)

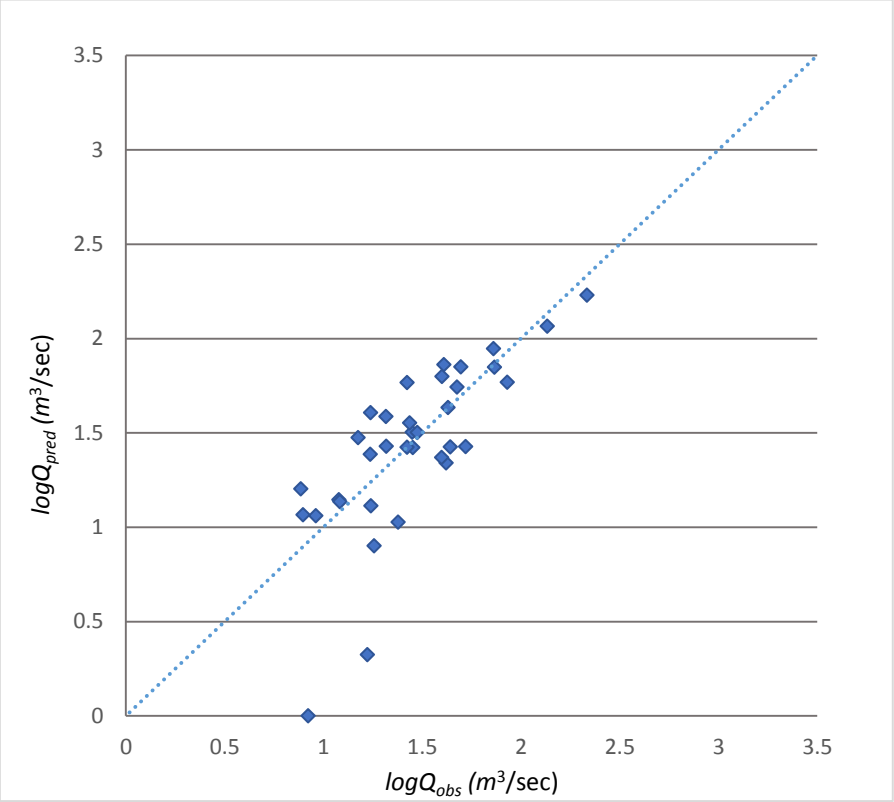

Figure E.11 Comparison of observed and predicted flood quantiles for GAM based RFFA model for Q_2(A2 group)

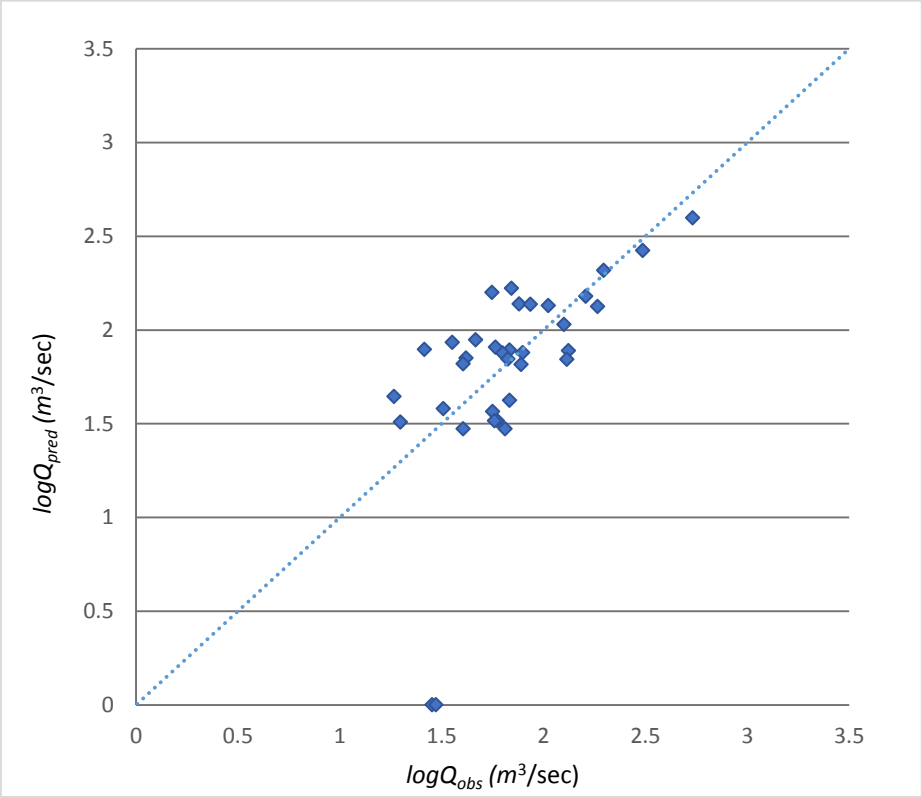

Figure E.12 Comparison of observed and predicted flood quantiles for GAM model for Q_5(A2 group)

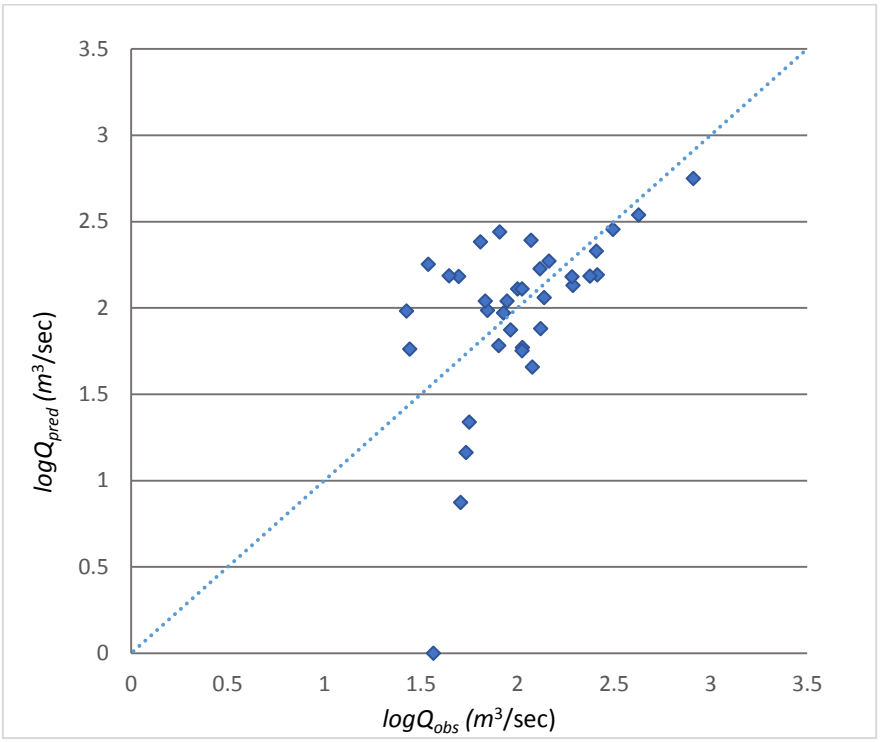

Figure E.13 Comparison of observed and predicted flood quantiles for GAM model for Q_{10}(A2 group)

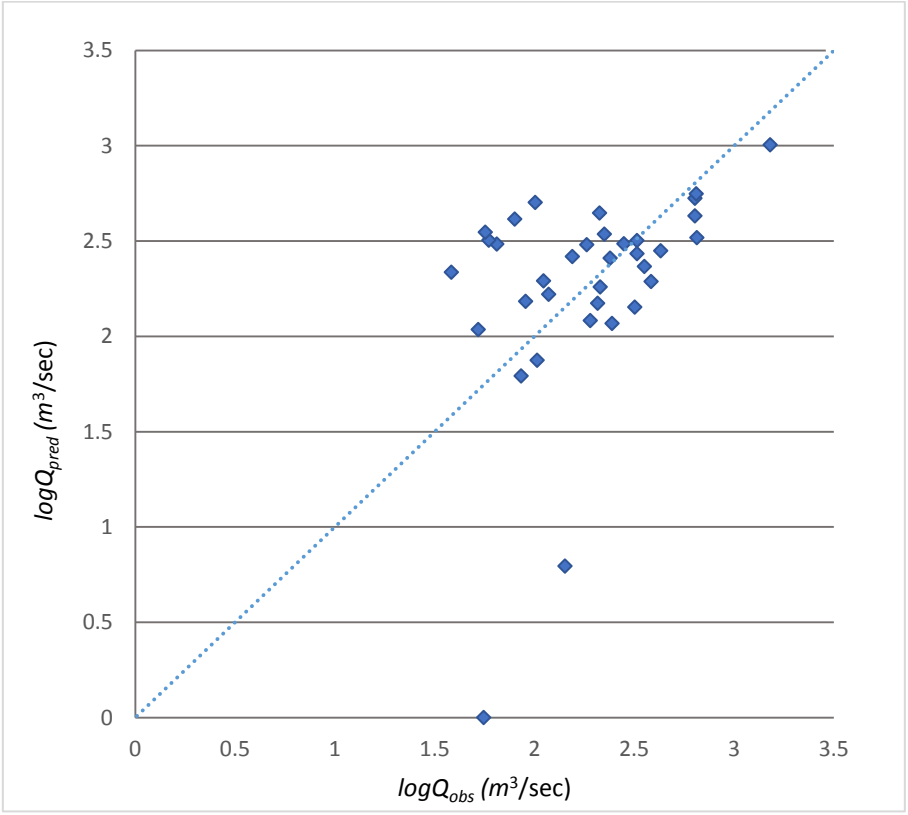

Figure E.14 Comparison of observed and predicted flood quantiles for GAM model for Q_{50}(A2 group)

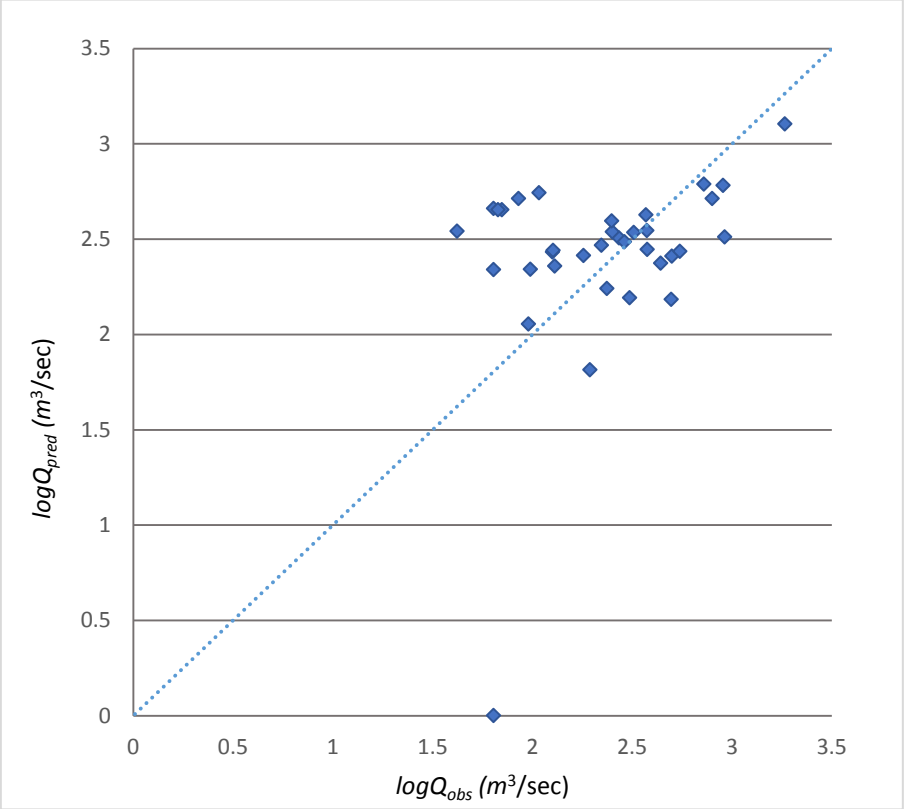

Figure E.15 Comparison of observed and predicted flood quantiles for GAM model for Q_{100}(A2 group)

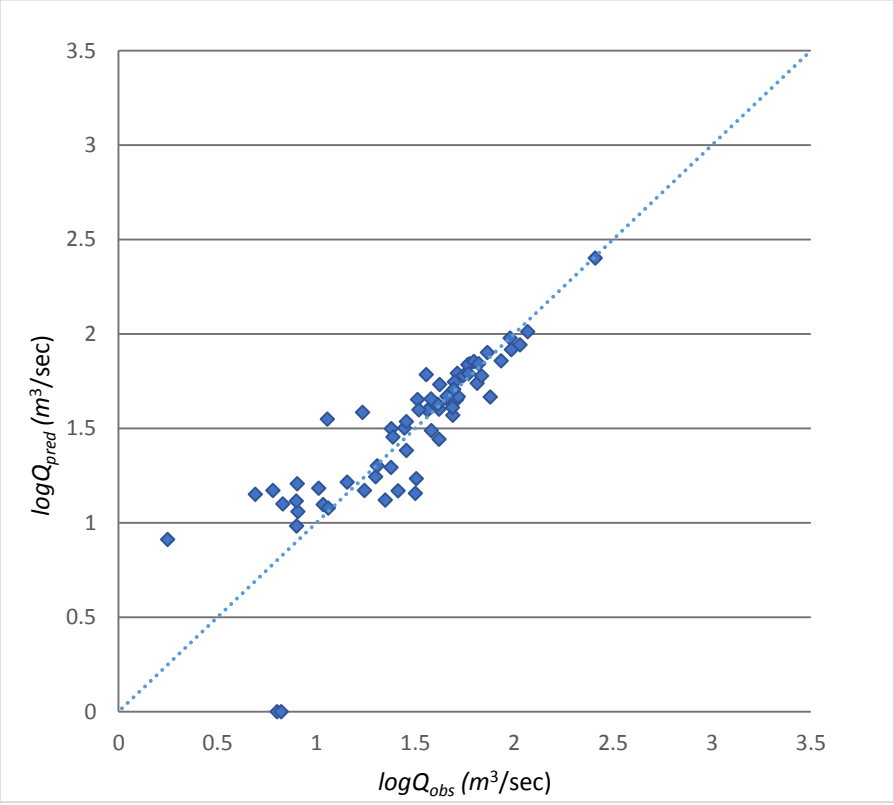

Figure E.16 Comparison of observed and predicted flood quantiles for GAM based RFFA model for Q_2(B1 group)

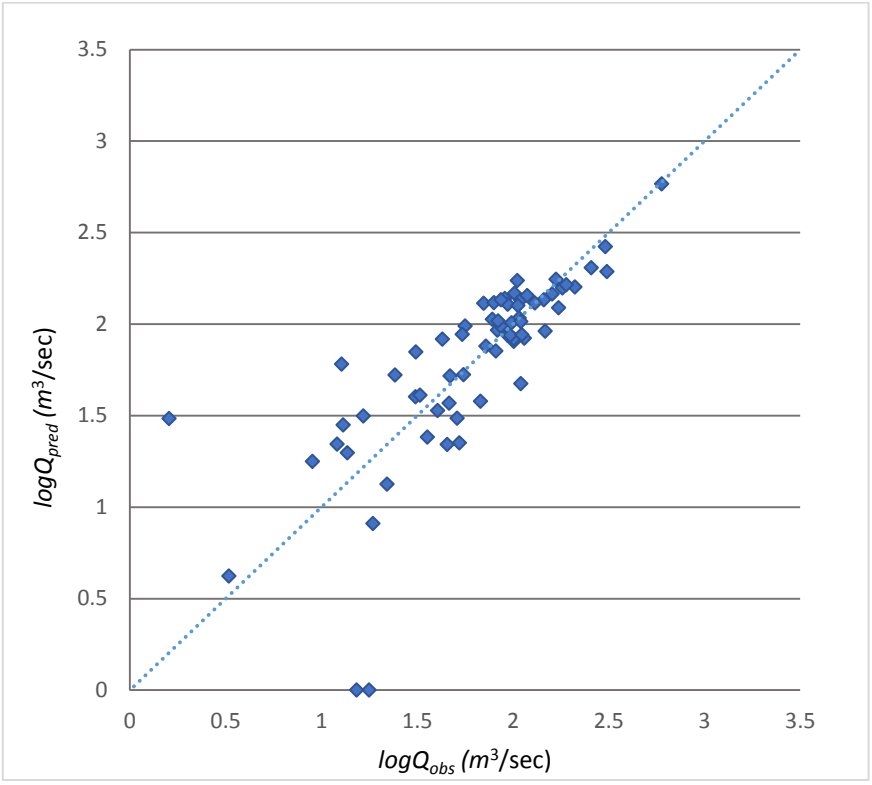

Figure E.17 Comparison of observed and predicted flood quantiles for GAM model for Q_5(B1 group)

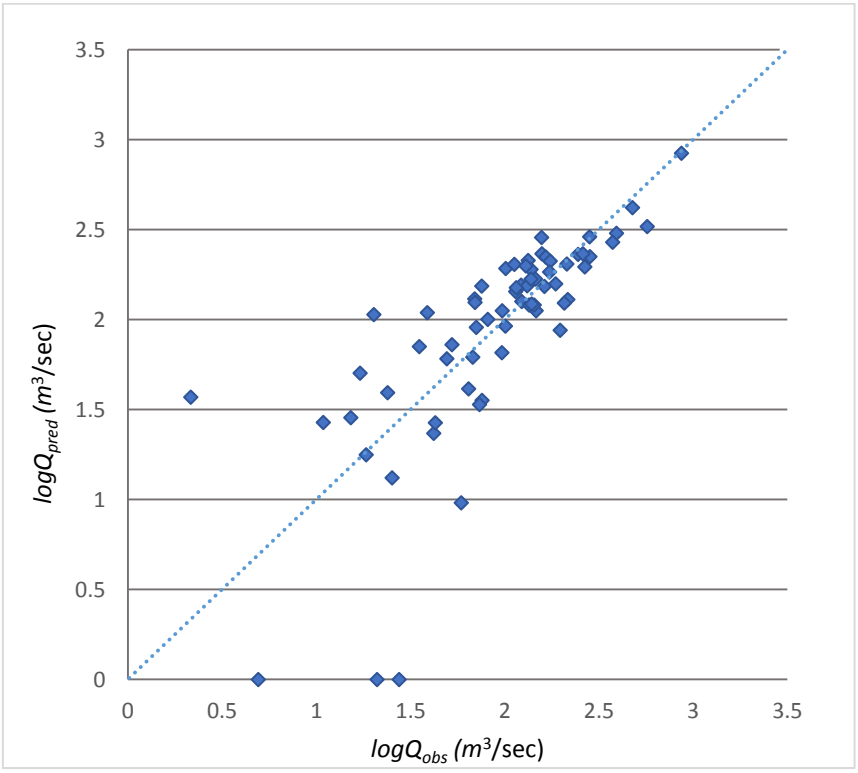

Figure E.18 Comparison of observed and predicted flood quantiles for GAM model for Q_{10}(B1 group)

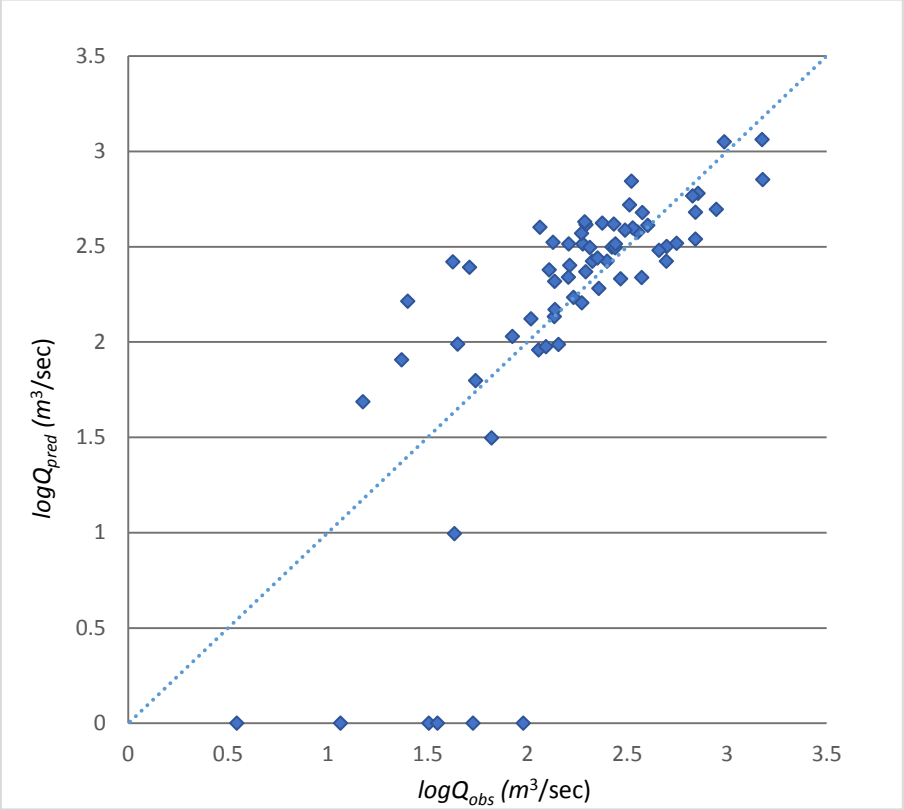

Figure E.19 Comparison of observed and predicted flood quantiles for GAM model for Q_{50}(B1 group)

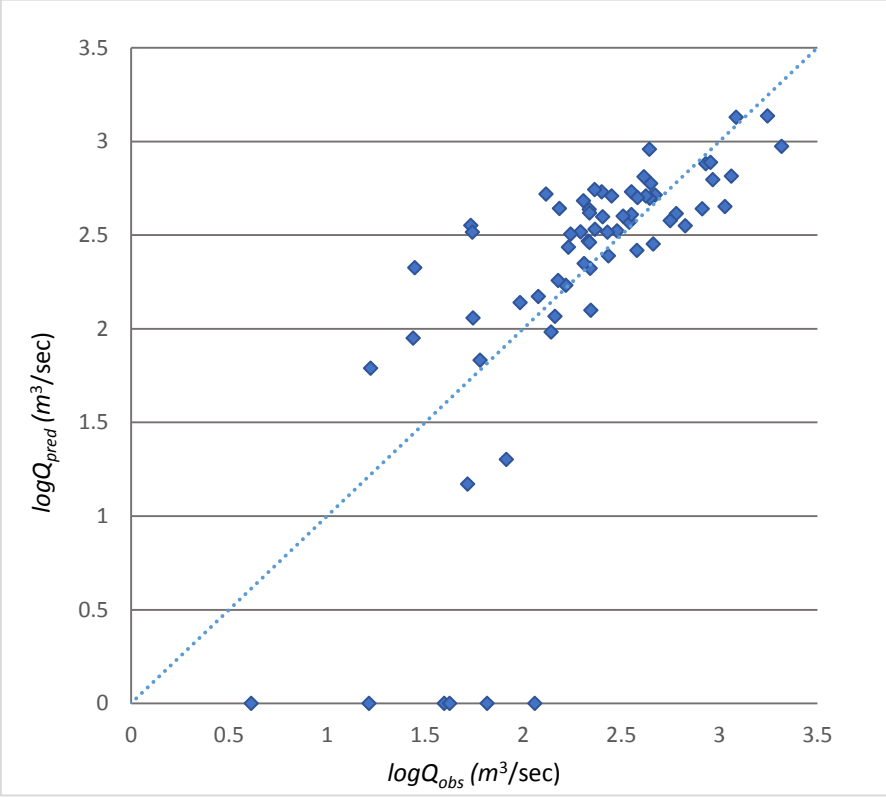

Figure E.20 Comparison of observed and predicted flood quantiles for GAM based RFFA model for Q_{100}(B1 group)

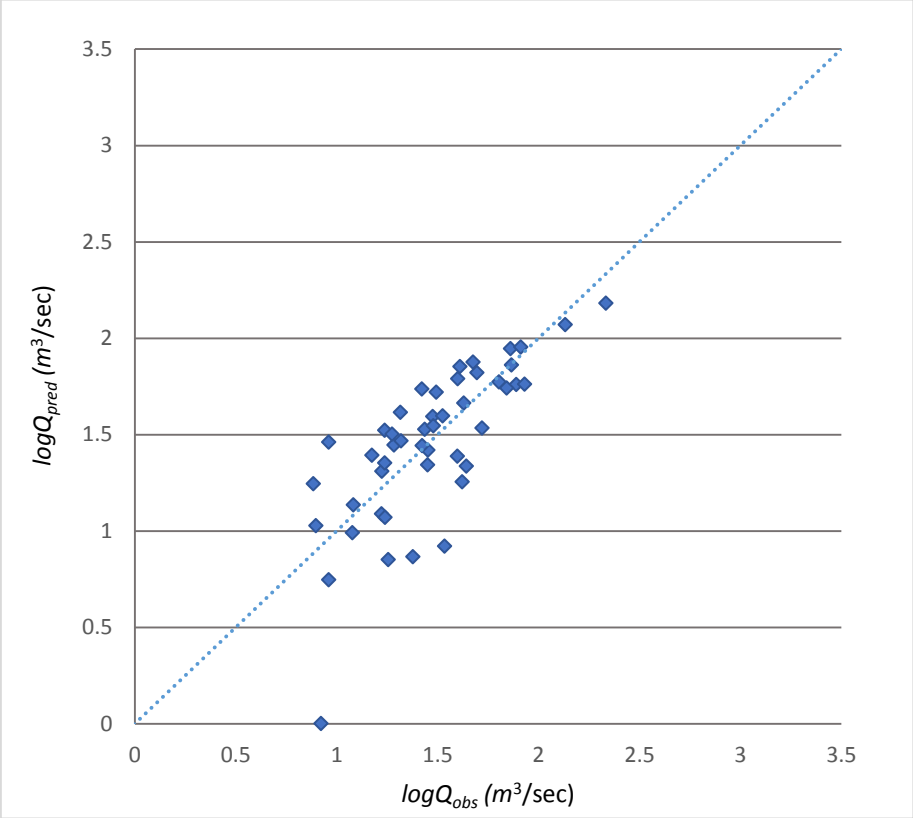

Figure E.21 Comparison of observed and predicted flood quantiles for GAM model for Q_2(B2 group)

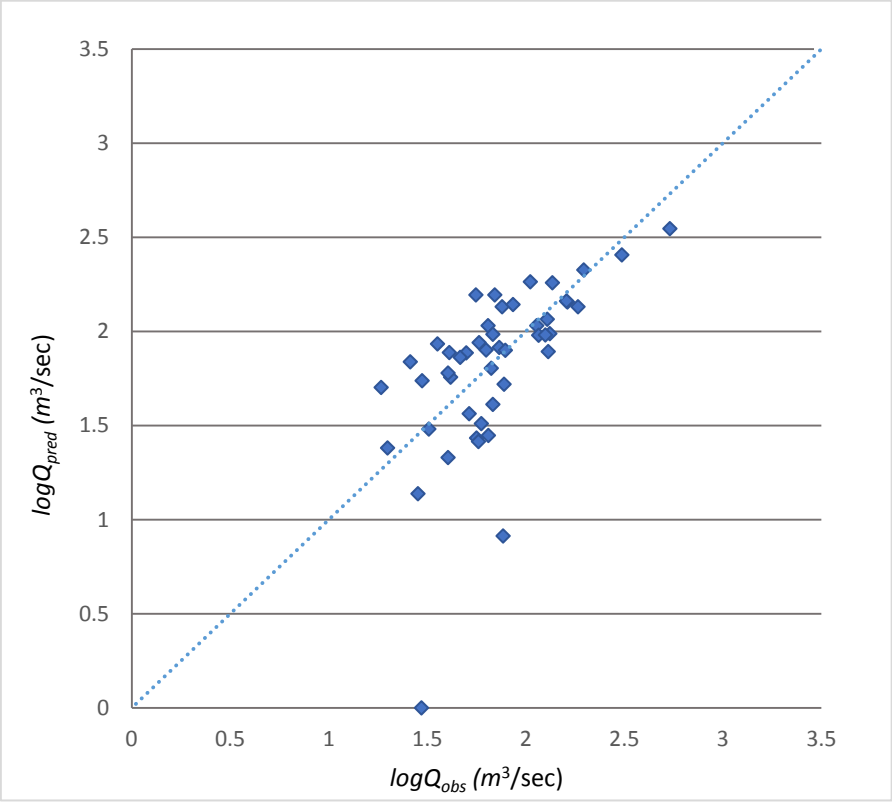

Figure E.22 Comparison of observed and predicted flood quantiles for GAM model for Q_5(B2 group)

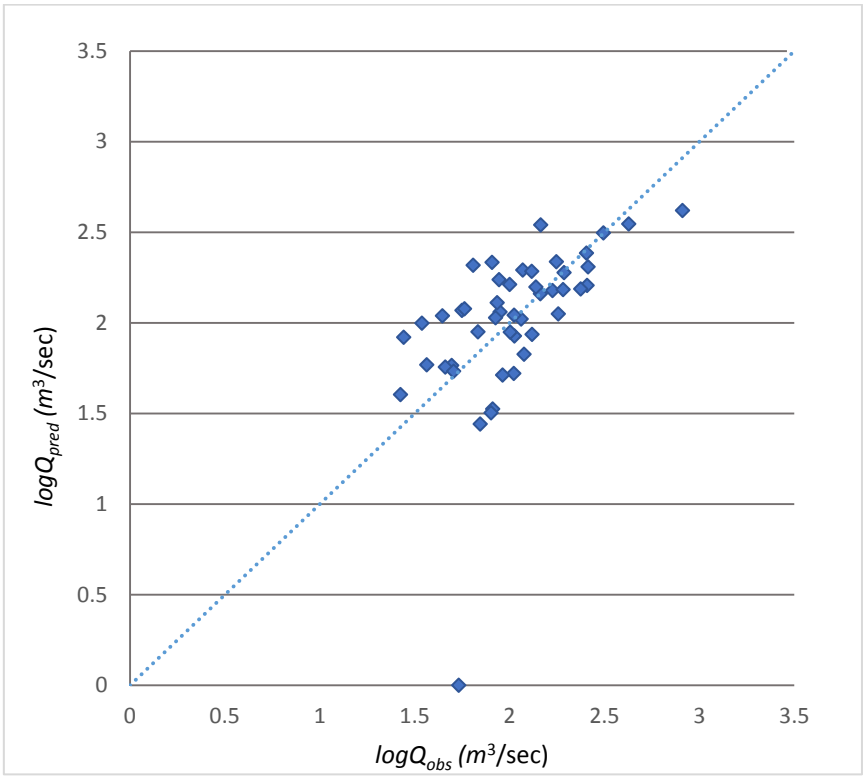

Figure E. 23 Comparison of observed and predicted flood quantiles for GAM model for Q_{10}(B2 group)

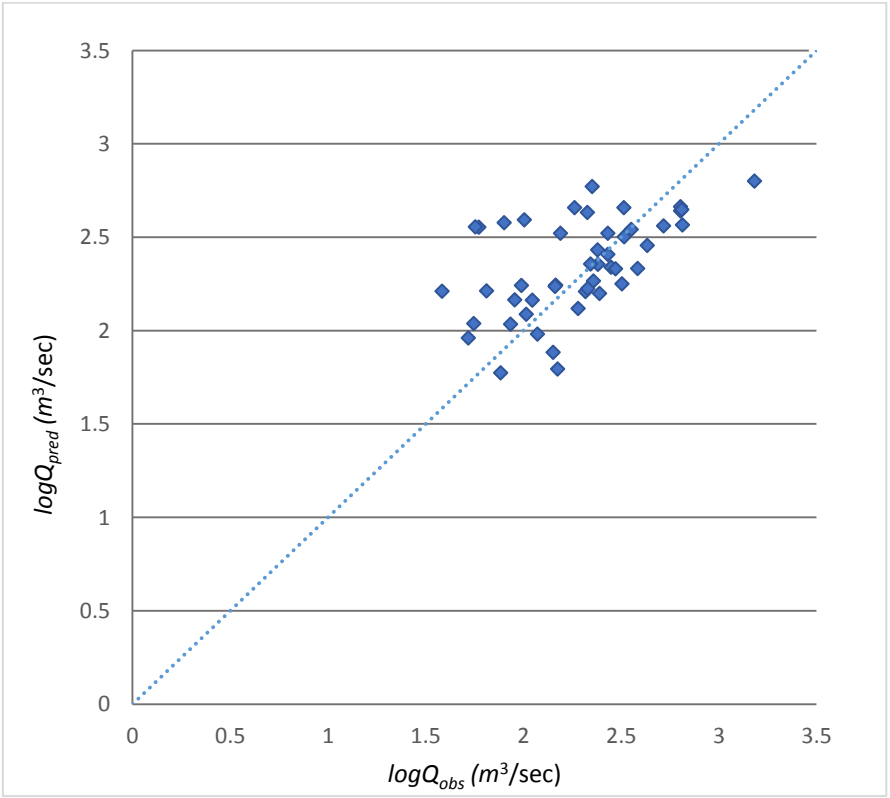

Figure E. 24 Comparison of observed and predicted flood quantiles for GAM model for Q_{50}(B2 group)

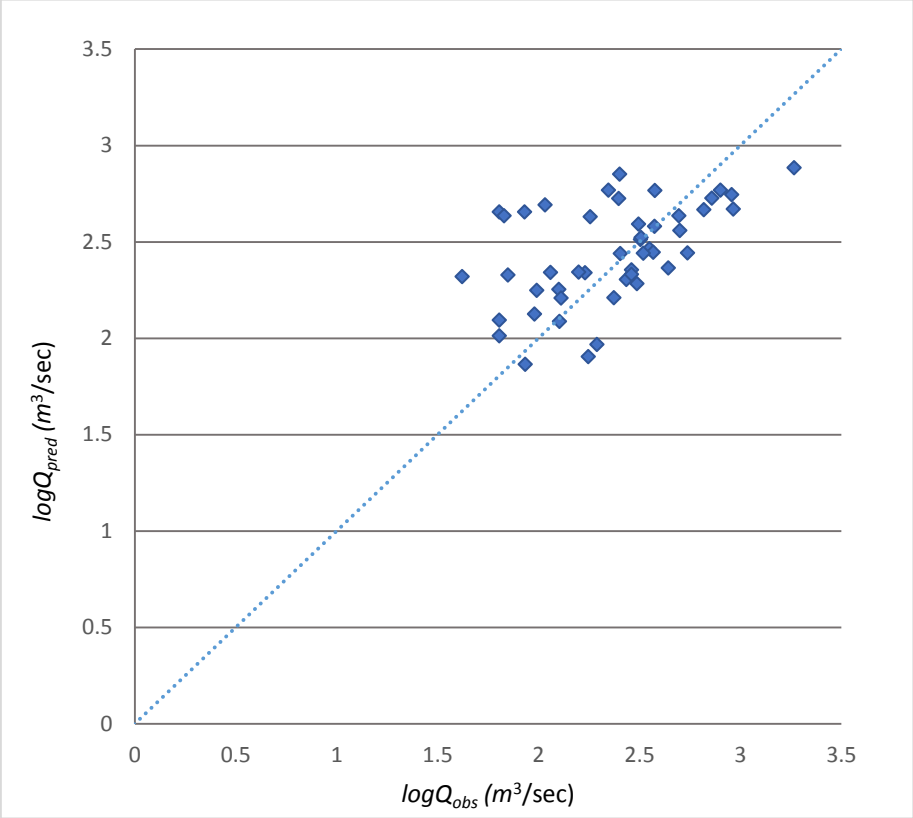

Figure E. 25 Comparison of observed and predicted flood quantiles for GAM model for Q_{100}(B2 group)