Stoch Environ Res Risk Assess (2018) 32:123–139
https://doi.org/10.1007/s00477-017-1384-1

ORIGINAL PAPER

Development of regional flood frequency analysis techniques using generalized additive models for Australia

A. Rahman[1] · C. Charron[2] · T. B. M. J. Ouarda[2,3] · F. Chebana[3]

Published online: 25 January 2017

Abstract Estimation of flood quantiles in ungauged catchments is a common problem in hydrology. For this, the log-linear regression model is widely adopted. However, in many cases, a simple log transformation may not be able to capture the complexity and nonlinearity in flood generation processes. This paper develops generalized additive model (GAM) to deal with nonlinearity between the dependent and predictor variables in regional flood frequency analysis (RFFA) problems. The data from 85 gauged catchments from New South Wales State in Australia is used to compare the performances of a number of alternative RFFA methods with respect to variable selection, variable transformation and delineation of regions. Four RFFA methods are compared in this study: GAM with fixed region, log-linear model, canonical correlation analysis (to form neighbourhood in the space catchment attributes) and region-of-influence approach. Based on the outcome from a leave-one-out validation approach, it has been found that the GAM method generally outperforms the other methods even without linking GAM with a neighbourhood/region-of-influence approach. The main strength of GAM is that it captures the non-linearity between the dependent and predictor variables without any restrictive assumption. The findings of this study will encourage other researchers worldwide to apply GAM in RFFA studies, allowing development of more flexible and realistic RFFA models and their wider adoption in practice.

Keywords GAM · Regression · Regional frequency analysis · Nonlinear models · Floods

1 Introduction

Flood is one of worst and costliest natural disasters. For example, in 2011 alone, the global flood damage was estimated to be worth $70 billion, plus over 6000 human deaths (Westra et al. 2014). Many studies have recently been conducted on various aspects of flooding to reduce flood damage (e.g. Motevalli and Vafakhah 2016; Kim et al. 2016; Kovalchuk et al. 2016). Flood estimate is needed in the design of hydraulic structures and many environmental and ecological studies, which aim to reduce the societal impacts of floods. At-site flood frequency analysis is the most direct method of design flood estimation, which however needs a long period of recorded streamflow data at the site of interest.

At many locations of interest, recorded streamflow data is unavailable or of short record length or of poor quality; and for these situations, regional flood frequency analysis (RFFA) is generally applied to estimate design floods (Ouarda et al. 2008a, b; Blöschl et al. 2013). RFFA essentially consists of two principal steps: identification of groups of hydrologically similar catchments and development of prediction equations (Ouarda 2013). In many previous RFFA studies, regions were formed based on geographic or administrative boundaries (IE Aust 1987), which often lacked in hydrological similarity (Bates et al.

✉ A. Rahman
a.rahman@westernsydney.edu.au

[1] School of Computing, Engineering and Mathematics, Western Sydney University, Building XB, Kingswood, Locked Bag 1797, Penrith, NSW 2751, Australia

[2] Institute Center for Water and Environment (iWATER), Masdar Institute of Science and Technology, P.O. Box 54224, Abu Dhabi, UAE

[3] INRS ETE, University of Quebec, 490 de la Couronne, Quebec, QC G1K 9A9, Canada

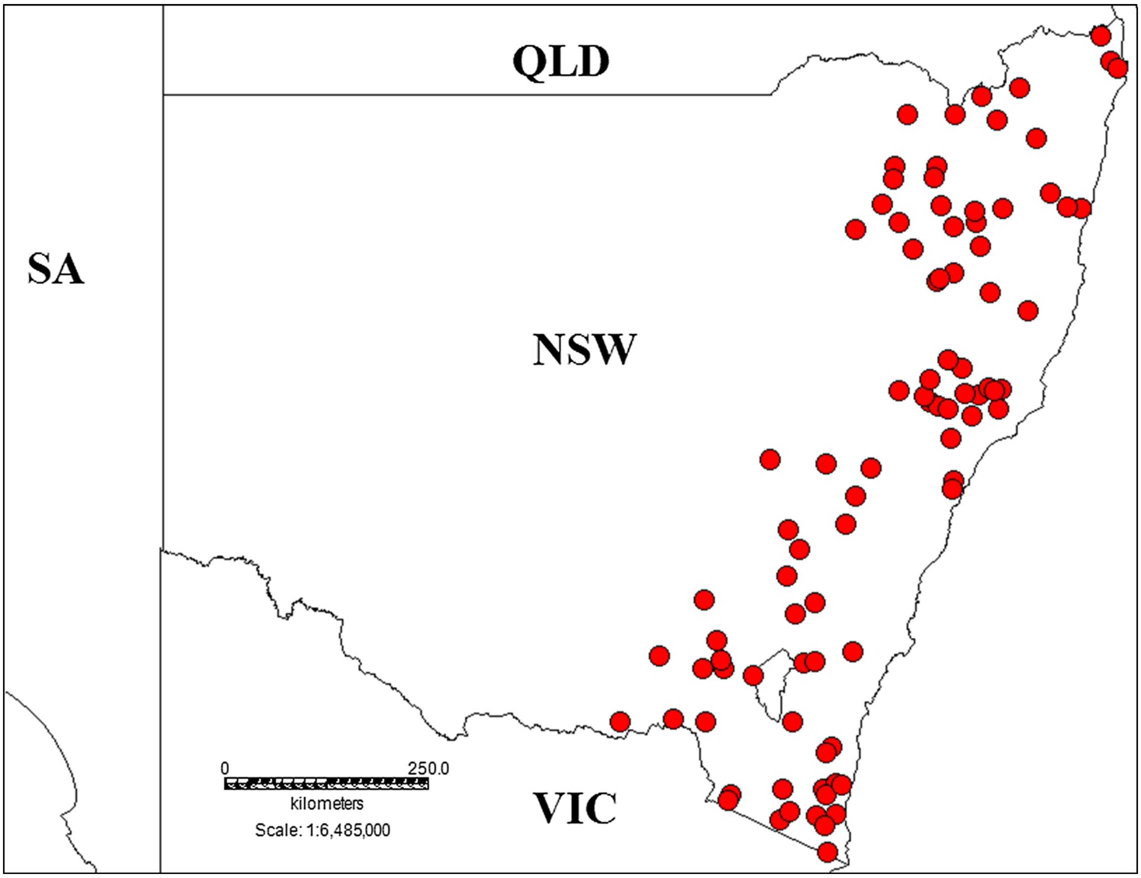

Fig. 1 Location of study catchments in NSW, Australia

1998; Chebana and Ouarda 2007). To overcome the limitations of fixed regions, a region-of-influence (ROI) approach was applied successfully in many studies (e.g. Burn 1990; Eng et al. 2007; Merz and Blöschl 2005; Haddad and Rahman 2012; Micevski et al. 2015; Rahman et al. 2015a, b). In ROI approach, allocation of an ungauged catchment to pre-defined homogeneous region(s) is problematic and moreover, model parameters vary from station to station, and hence practical application of ROI-based RFFA methods is not straight forward.

There have been number of RFFA studies which compared alternative RFFA methods. For example, Ouarda et al. (2008a, b) compared four different approaches to form homogeneous regions using data of 29 stream gauging stations from several Mexican River Basins. These include (1) hierarchical cluster analysis that delivered fixed regions; (2) canonical correlation analysis (CCA) that allowed formation of regions, or neighbourhoods which are specific to the site of interest; (3) a modified form of CCA, which did not require parameter optimization; and (4) canonical kriging that allowed interpolation of hydrological variables in the canonical physiographical space. The study revealed the advantages of the neighbourhood type of approach and the superiority of the CCA method. In another study, Ouarda et al. (2006) compared three flood seasonality regionalization methods to a flood dataset from Québec (Canada). Using a leave-one-out (LOO) validation (Haddad et al. 2013), the seasonality method was compared with a traditional regionalization approach based on similarities in catchment physiographic data space. It was found that the seasonality method based on the peaks over-threshold approach outperformed the traditional regionalization approach.

In relation to the development of regional prediction equations, the index flood method has been applied widely, which is dependent on the criteria of statistical homogeneity (Hosking and Wallis 1993; Fill and Stedinger 1995; Rahman et al. 1999; Cunderlik and Burn 2006; Castellarin et al. 2008, Chebana and Ouarda 2009; Wazneh et al. 2013). In contrast, the quantile regression technique (QRT) relaxes the criteria of statistical homogeneity. With QRT both ordinary least squares and generalized least squares regression methods have been adopted to estimate regression parameters/coefficients (e.g. Pandey and Nguyen 1999; Stedinger and Tasker 1985; Rahman 2005; Griffis and Stedinger 2007; Micevski and Kuczera 2009; Haddad et al. 2012, 2015; Ouali et al. 2016). Durocher et al. (2015) proposed a RFFA approach based on

Table 1 Descriptive statistics of hydrological and physio-meteorological variables of selected 85 catchments in New South Wales, Australia

Variable	Unit	Notation	Min	Mean	Max	SD
Flood quantile of 10 year return period	m^3/s	Q_{10}	8.48	415.52	2028.36	377.52
Flood quantile of 50 year return period	m^3/s	Q_{50}	15.79	833.35	4306.82	721.81
Catchment area	km^2	AREA	8.00	351.98	1010.00	281.43
Catchment shape factor	–	SF	0.26	0.76	1.63	0.21
Main stream slope	m/km	S10,85	1.54	12.92	49.86	10.80
Stream density	km/km^2	SDEN	0.52	2.85	5.47	1.10
Percentage of catchment covered by forest	%	FOREST	0.00	0.51	0.99	0.32
Rainfall intensity (6 h duration and 2 year return period)	mm/h	I6,2	31.30	45.40	87.30	11.27
Mean annual rainfall	mm	MAR	626.17	1000.28	1953.23	304.48
Mean annual potential evapotranspiration	mm	MAE	980.40	1223.69	1543.30	126.30

projection pursuit regression (PPR). PPR is a family of regression models that applies smooth functions on intermediate predictors to fit complex patterns. Results indicated that the procedure is efficient in modelling nonlinearities and handles efficiently problematic stations. Durocher et al. (2016) noted that the regression of flood quantiles in RFFA is often carried out at the logarithmic scale, which introduces a bias and leads to suboptimal estimates. They examined the use of spatial copulas to offer proper corrections in this framework. Spatial copulas are the formulation of traditional geostatistics by copulas. Their results showed that spatial copulas can deal with the problem of bias, are simple to apply and are robust to the presence of problematic stations. Ouali et al. (2015) presented a RFFA procedure based on non-linear canonical correlation analysis (NL-CCA). Their results indicated that NL-CCA is more robust that most commonly used linear RFFA procedures and can reproduce efficiently the non-linear relationshop structures between physiographical and hydrological variables.

Most of the regression-based RFFA models are based on linearity assumption. The linear models assume that the relationship between the dependent variable (e.g. flood quantile) and the predictor variables (physio-meteorological) are linear. Hydrological processes are naturally complex in several aspects including nonlinearity (Chebana et al. 2014). The linearity assumption in hydrology may not be satisfied in many cases (for example, larger catchments behave differently than smaller ones and drier antecedent catchment state produces relatively smaller runoff than wetter one). The application of non-linear methods in RFFA problems is rather limited. A number of studies applied artificial intelligence based methods to RFFA problems (e.g. Dawson et al. 2006; Shu and Ouarda 2007, 2008; Ouarda and Shu 2009; Aziz et al. 2014, 2015, 2016, Alobaidi et al. 2015) and these studies have found that non-linear methods generally outperform the linear ones.

The application of more general non-linear methods such as the generalized additive model (GAM) (Hastie and Tibshirani 1986; Wood 2006) has increased in recent years due to the development of new statistical tools and computer programs (e.g. Wood 2003; Kauermann and Opsomer 2003; Morlini 2006; Schindeler et al. 2009). GAMs have been applied successfully in environmental studies (e.g. Wood and Augustin 2002; Wen et al. 2011), in renewable energy assessment (e.g. Ouarda et al. 2016) and also in public health and epidemiological research (Leitte et al. 2009; Vieira et al. 2009; Bayentin et al. 2010; Clifford et al. 2011). There have been numbers 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 been only limited of applications of GAM. Tisseuil et al. (2010) applied generalized linear model (GLM), GAM, aggregated boosted trees 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. 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 Québec, 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.

GAM allows for the inclusion and presentation of non-linear 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 as compared with a wetter catchment. Hence, 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 relationship 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. This study focuses on the development and testing of GAM in RFFA and comparison with other more established RFFA methods. This study uses data from New South Wales (NSW) State in Australia. Australian hydrology is known to have a higher degree of variability and non-linearity, and hence testing the applicability of GAM is worthwhile for Australian conditions, as done in this study.

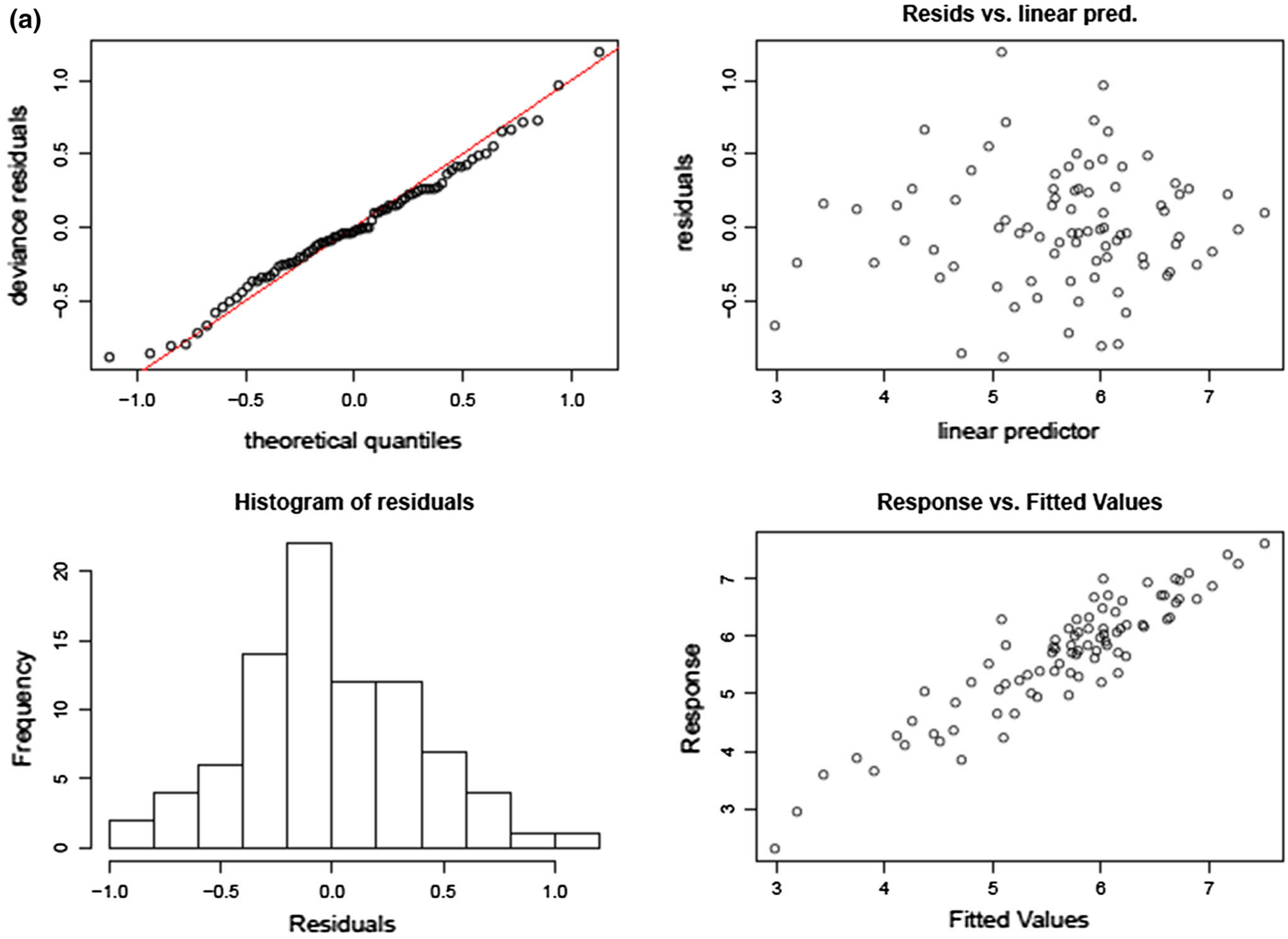

Fig. 2 Graphs of validation of the fit of **a** Q10 for the model GAM (Figure generated with the package in R), **b** Q50 for the model GAM (Figure generated with the package in R)

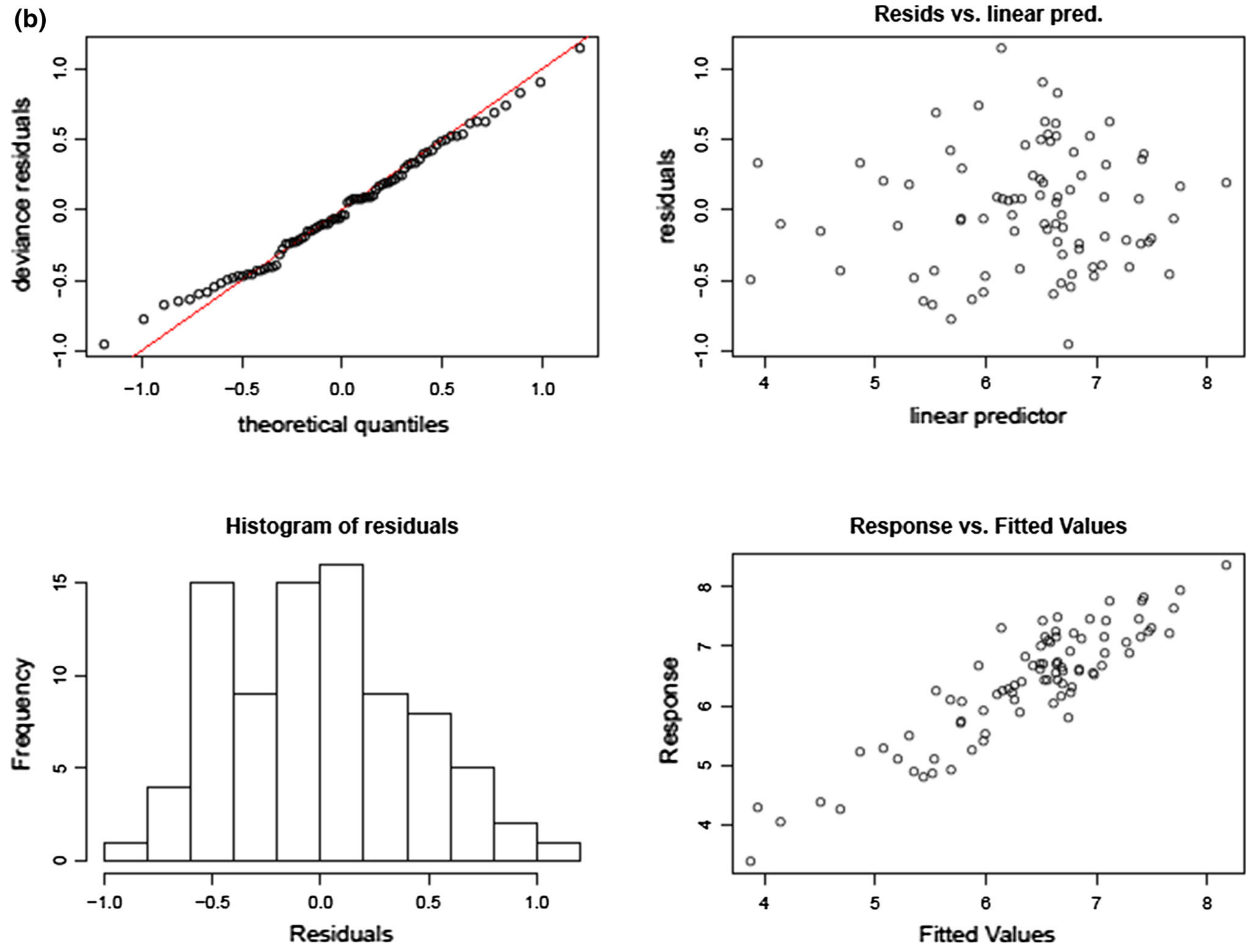

Fig. 2 continued

2 Methods

2.1 Multiple linear regression

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

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad (1)$$

where β_0 and β_j ($j = 1, 2, \ldots, p$) are unknown parameters and ε_i is the error term associated with i-th observation ($i = 1, 2, \ldots, n$), where $n =$ number of observations. The error term in Eq. 1 is assumed to be normally distributed $N(0, \sigma^2)$ and the model parameters are generally estimated by the method of least squares. In RFFA, a log-linear model is widely adopted where both the dependent and predictor variables are log-transformed in building the regression model under the assumption that it will achieve normality of the predictors and linearity between the dependent variable and predictor variables. Girard et al. (2004) presented a procedure for the correction of the bias that results from the use of the log-linear multiple regression model in RFFA.

2.2 Generalized additive model

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

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \varepsilon_i \qquad (2) \end{aligned}$$

GAM allows fitting a non-linear function f_j to each X_j i.e. one does not need to manually try out numerous transformations on each of the predictor variables. Since GAM is an additive model, one can examine the impact of each X_j on Y individually. In this model, the smoothness of function f_j for the variable X_j is summarized via degrees of freedom. In GAM, the linear predictor predicts a known smooth monotonic function of the expected value of the response, and the response may follow any distribution from exponential family or may have a known mean variance relationship, allowing a quasi-likelihood approach (Wood 2006).

In GAM, to estimate the smooth function f_j a spline is adopted. A number of spline types are available (e.g. P-splines, cubic splines and B-splines). In this study, thin plate regression splines are adopted as they provide fast computation, do not require selection of knot locations and have optimality in approximating smoothness (Wood 2003, 2006). Further information on GAM can be found in Wood (2008) and Chebana et al. (2014).

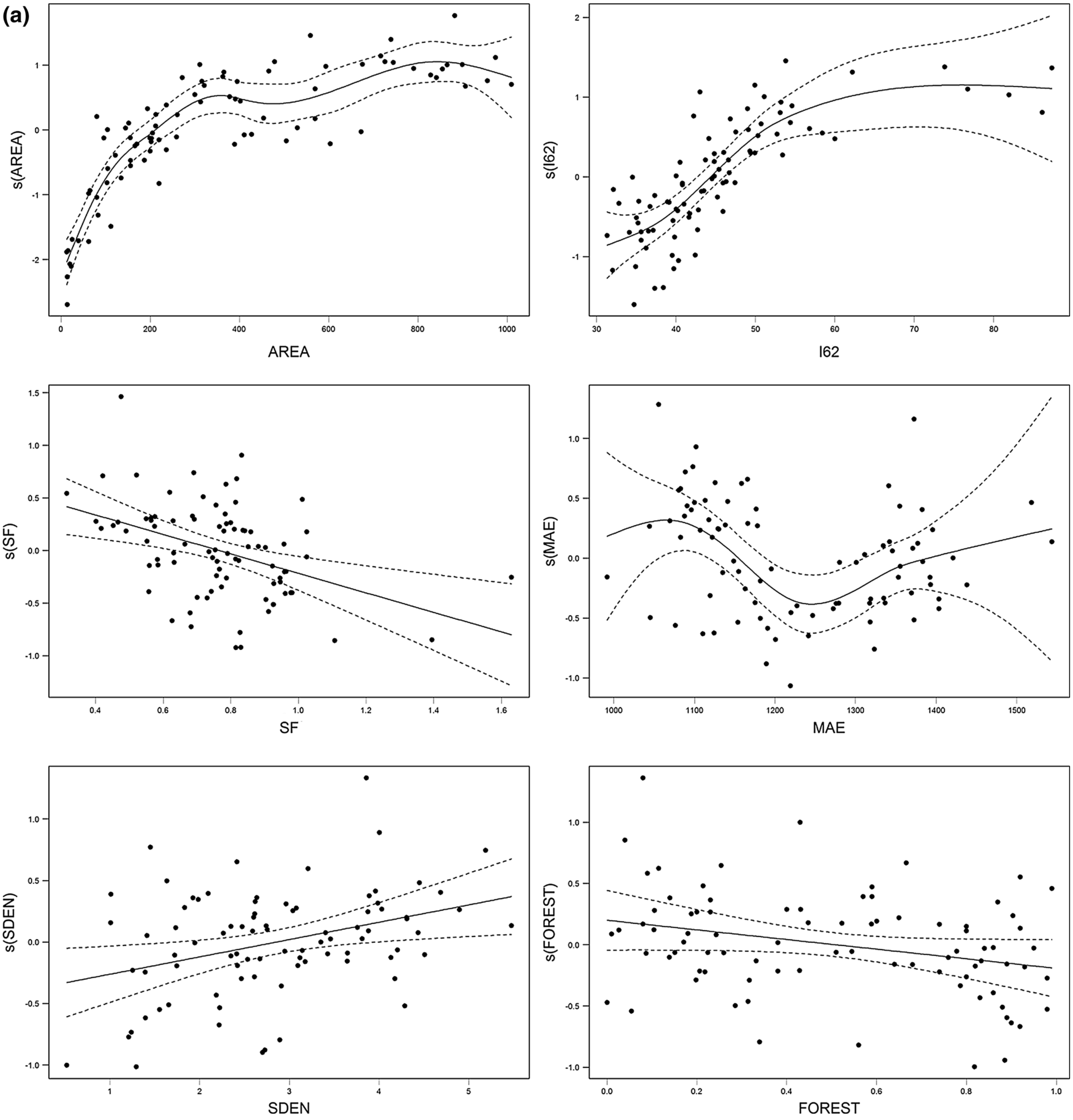

Fig. 3 Smooth functions of each predictor for **a** Q10. **b** Q50

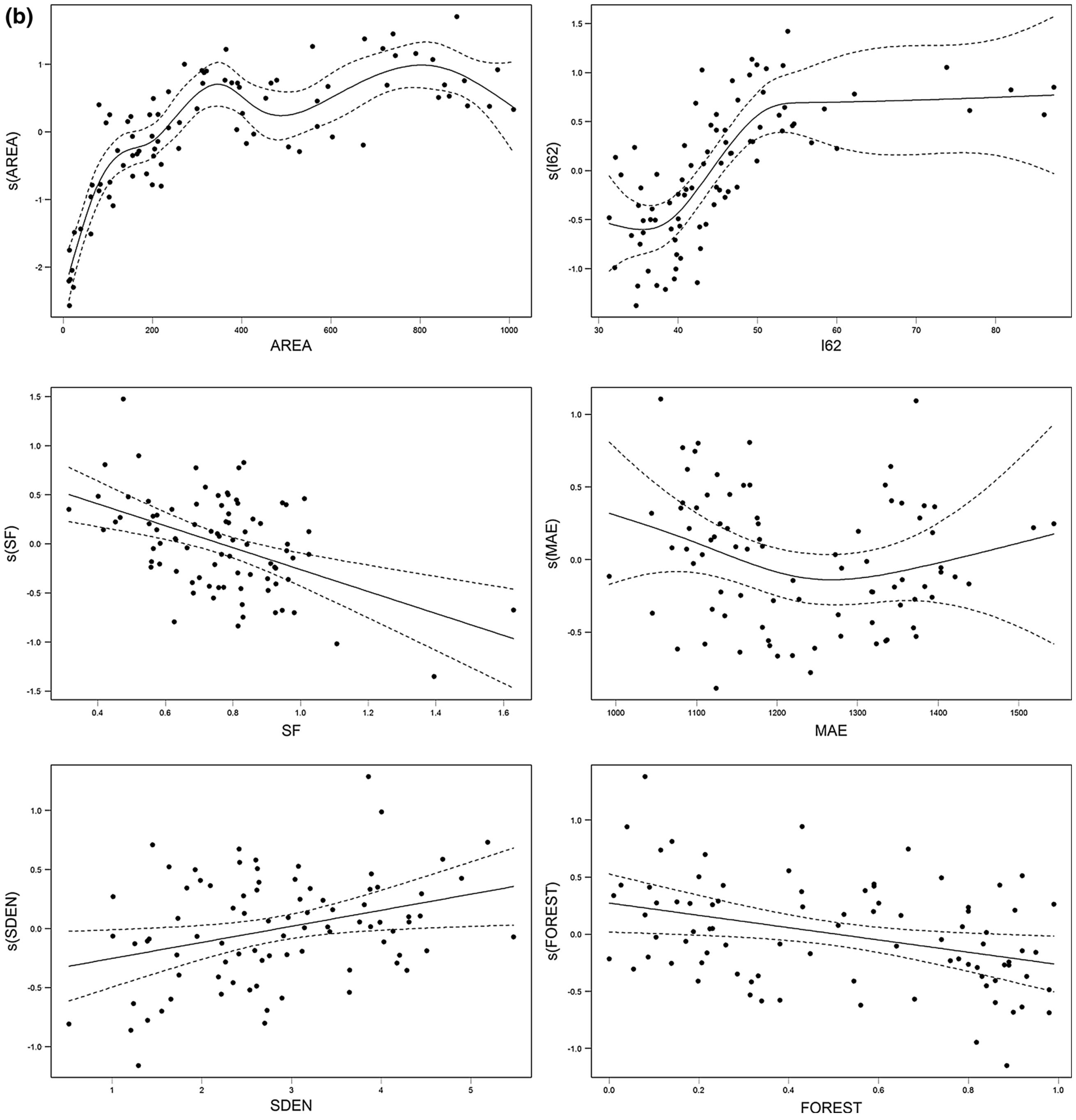

Fig. 3 continued

2.3 Canonical correlation analysis in RFFA

Canonical correlation analysis (CCA) can be used in RFFA for two sets of random variables: predictor set consisting of physio-meteorological variables $X = (X_1, X_2, \ldots, X_p)$ and dependent variable set consisting flood quantiles $Y = (Y_1, Y_2, \ldots, Y_k)$, $p \geq k$. CCA allows identification of the dominant linear modes of covariability between the sets X and Y that allows making inference about Y knowing X. Ouarda et al. (2001) provided step-by-step procedures to implement hydrologic neighborhoods with the CCA approach for both the gauged and ungauged catchments. It should be noted that, in CCA, the original dependent and independent variables data (which are available in different units of measurement) are standardized prior to the analysis to derive canonical variables that are free of scale effects. Further details on CCA can be found in Ouarda et al. (2001) and Bates et al. (1998).

2.4 Model validation

Each of the developed models is assessed by a leave-one-out (LOO) validation procedure (Haddad et al. 2013). In this procedure gauged catchments are in turn considered ungauged in RFFA. We adopted the following statistical measures compare different models:

$$\text{Coefficient of determination:} \quad \mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \tag{3}$$

$$\text{Root mean square error:} \quad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2} \tag{4}$$

Relative root mean square error:

$$\mathrm{rRMSE} = 100\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(z_i - \hat{z}_i)/z_i\right]^2} \tag{5}$$

$$\text{Mean bias: BIAS} = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i) \tag{6}$$

$$\text{Relative mean bias: rBIAS} = 100\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)/z_i \tag{7}$$

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

3 Data description

The study uses data from 85 streamflow gauging stations in New South Wales (NSW) State of Australia (Fig. 1). The selected stations fall within latitudes −28.36 to −37.37° and longitudes 146.98–153.50°. This part of Australia has moderate rainfall, with mean annual rainfalls in the range

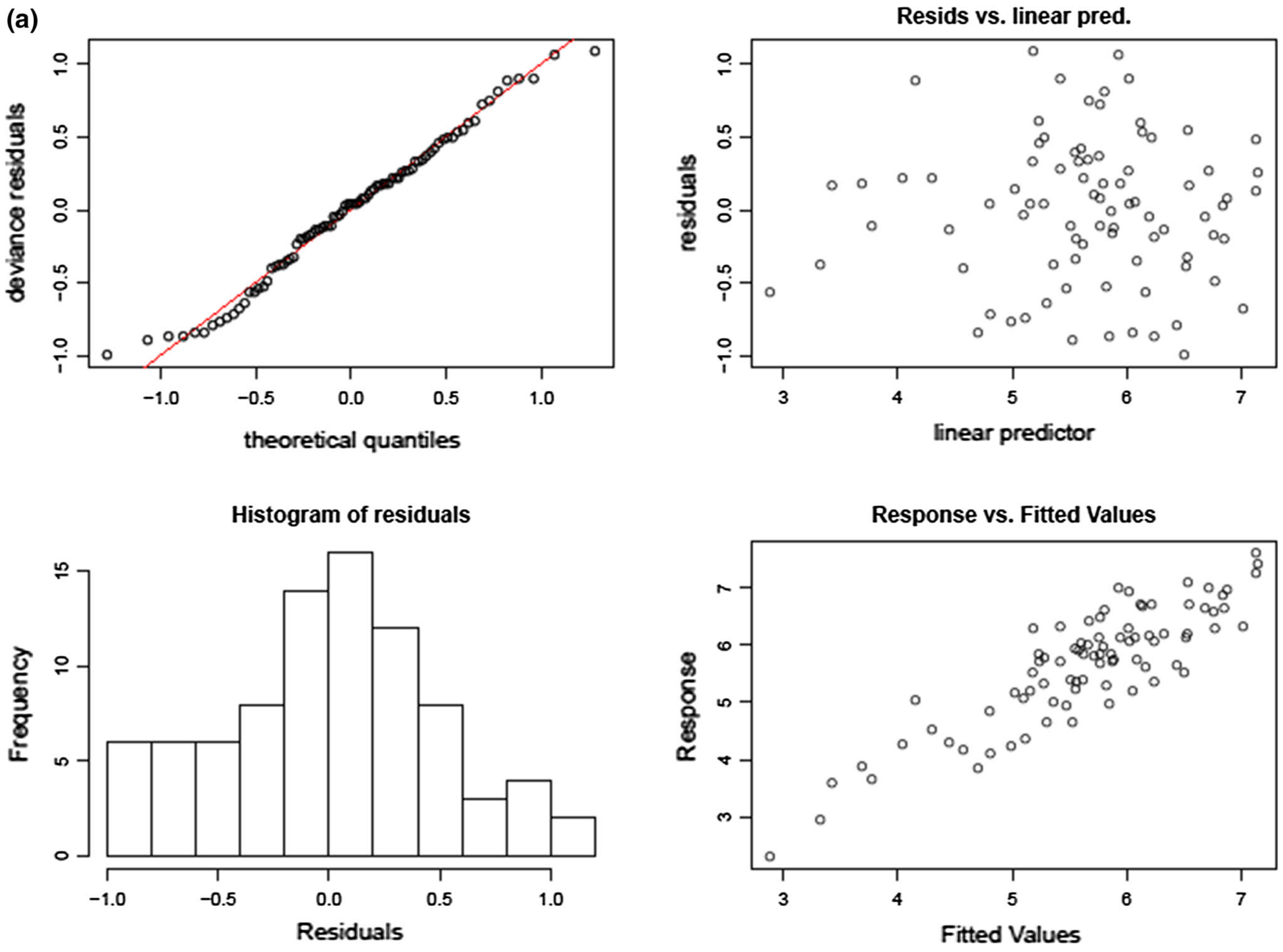

Fig. 4 Validation of the fit of **a** Q10 for the model LL (Figure generated with the package in R). **b** Q50 for the model LL (Figure generated with the package in R)

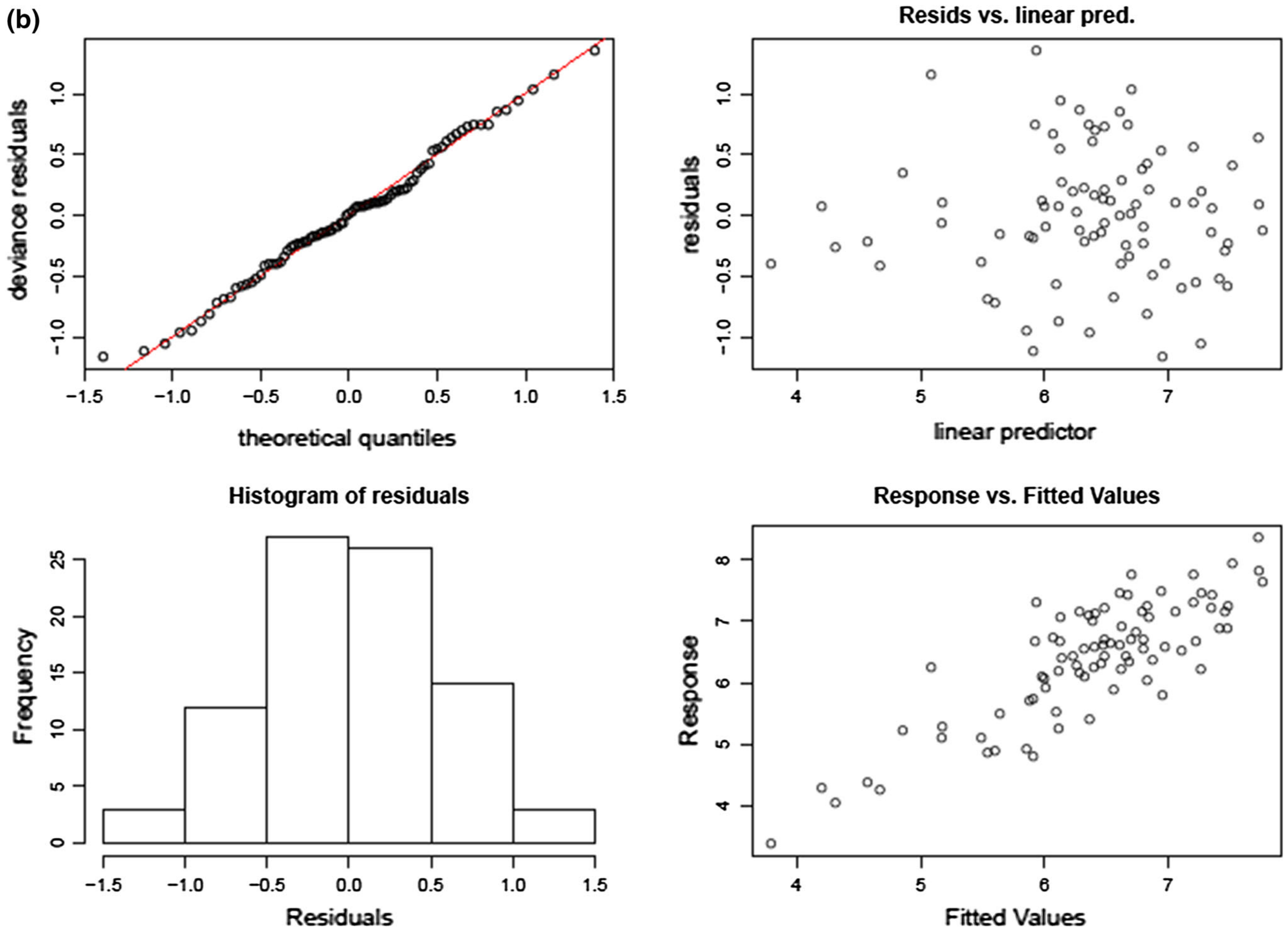

Fig. 4 continued

of 200–800 mm. The selected stations are situated on naturally flowing rivers, with no major regulations and land use changes occurring in the catchments over the period of streamflow records adopted in this study. These catchments are small to medium-sized, with catchment size in the range of 8–1010 km^2 (mean: 352 km^2). Streamflow data covers the period of 1930–2011, with some stations starting as late as 1980. The annual maximum streamflow record lengths are in the range of 25–82 years (mean 42 years).

In preparing the data set, the initially selected potential catchments were examined as detailed in Haddad et al. (2010), Ishak et al. (2013) and Rahman et al. (2015a, b): gaps in the annual maximum flood series were filled as far as could be justified (up to 3% annual maximum flood series data were in-filled by regression), outliers were detected using the multiple Grubbs-Beck test (Lamontagne et al. 2013) and error associated with rating curve extrapolation was investigated (Haddad et al. 2010) and the stations with a high rating curve extrapolation were excluded.

Flood quantiles were estimated using the FLIKE software (Kuczera 1999; Kuczera and Franks 2015) by fitting Log-Pearson Type III (LP3) distribution using Bayesian parameter estimation procedure. Flood quantiles for 2, 5, 10, 20, 50 and 100 year return periods were estimated based on the fitted LP3 distribution where censoring of the low flood values was carried out using multiple Grubbs–Beck test (Lamontagne et al. 2013). It should be noted that the LP3 distribution provided the best-fit for the study area among a number of other distributions including GEV-L moments method (Rahman et al. 2013). In this study, 10 year return period (Q_{10}) and 50 year return period (Q_{50}) quantiles are adopted to assess the applicability of GAM in Australian conditions. It is expected that other return periods will provide similar results to Q_{10} and Q_{50}.

For the selection of candidate predictor variables, the variables considered by the similar RFFA studies were initially examined. It was found that most of the previous RFFA studies included catchment area and mean annual rainfall as predictor variables (e.g. Griffis and Stedinger 2007; Shu and Ouarda 2008; Flavell 2012; Haddad and Rahman 2012). It was found that three previous Australian studies (Flavell 2012; Haddad and Rahman 2012; Rahman (2005) adopted design rainfall intensity and

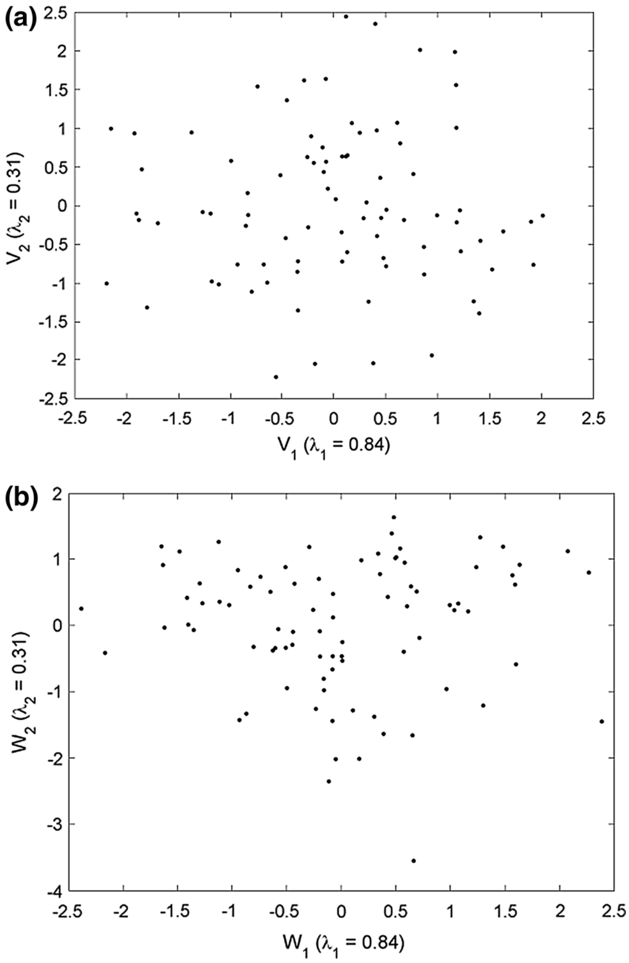

Fig. 5 Plots showing physio-meteorological canonical space (**a**) and hydrological canonical space (**b**)

evapotranspiration. Percentage of catchment covered by forest was adopted by many previous studies (e.g. Griffis and Stedinger 2007; Haddad and Rahman 2012). Main stream slope was adopted by Flavell (2012) and Griffis and Stedinger (2007)), evapotranspiration was adopted by Haddad and Rahman (2012) and Rahman (2005)) and catchment shape was adopted by Rahman et al. (2015b).

The above eight predictor variables capture the flood generation and attenuation processes quite well. Catchment area largely defines the flood magnitude, i.e. the bigger the catchment area for a given rainfall, the greater is the flood peak. Rainfall (rainfall intensity and mean annual rainfall) is the main input to the rainfall-runoff process, i.e. the larger the rainfall, the higher is the flood for a given catchment area. Moreover, the higher the main stream slope the greater is the flow velocity which leads to higher flood peak and a higher stream density increases drainage efficiency of a catchment which leads a smaller catchment response time i.e. higher peak flood. The evapotranspiration is the loss component in the rainfall and runoff process. Elongated catchment shape increases the catchment response time and hence a smaller peak and increased forest cover provides a higher roughness which reduces flow velocity. Based on the above considerations, these eight predictor variables are selected in this study (summarised in Table 1).

Catchment area was measured on 1:100,000 topographic maps. Catchment shape factor was taken as the ratio of the shortest distance between catchment outlet and centroid and square root of catchment area. The main stream slope used in this study excluded the extremes of slope found at the very upstream and downstream parts of a stream; it was taken as the ratio of the difference in elevation of the stream bed at 85 and 10% of its length from the basin outlet, and 75% of the mainstream length. The slope was determined from 1:100,000 topographic maps using an opisometer to measure the stream length. The design rainfall intensity data was obtained from Australian Rainfall and Runoff (ARR) (Ball et al. 2016). The mean annual areal potential evapo-transpiration data was obtained from

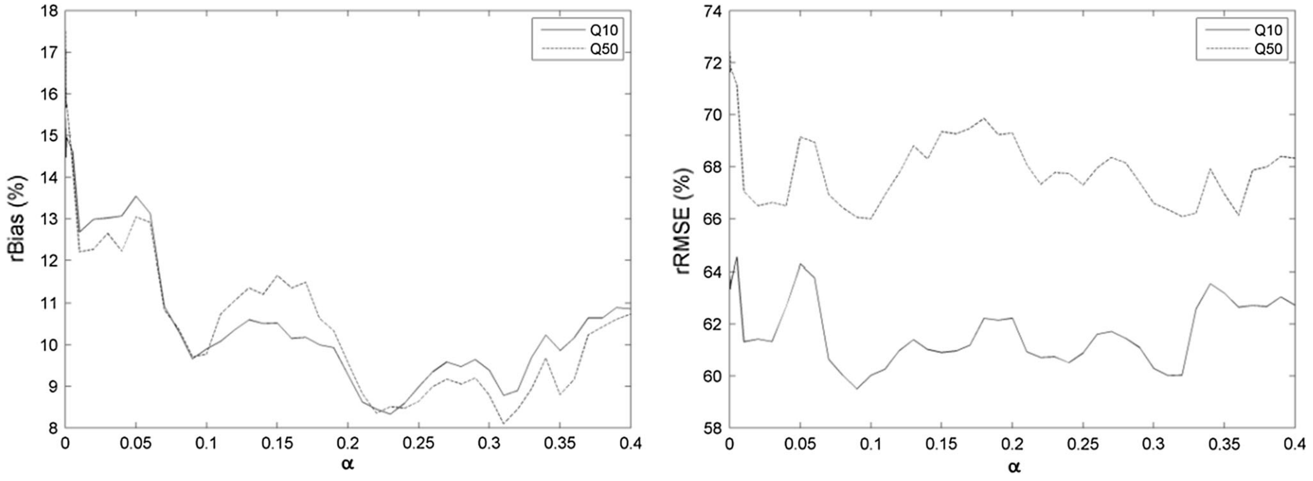

Fig. 6 rBias and rRMSE as a function of the parameter α

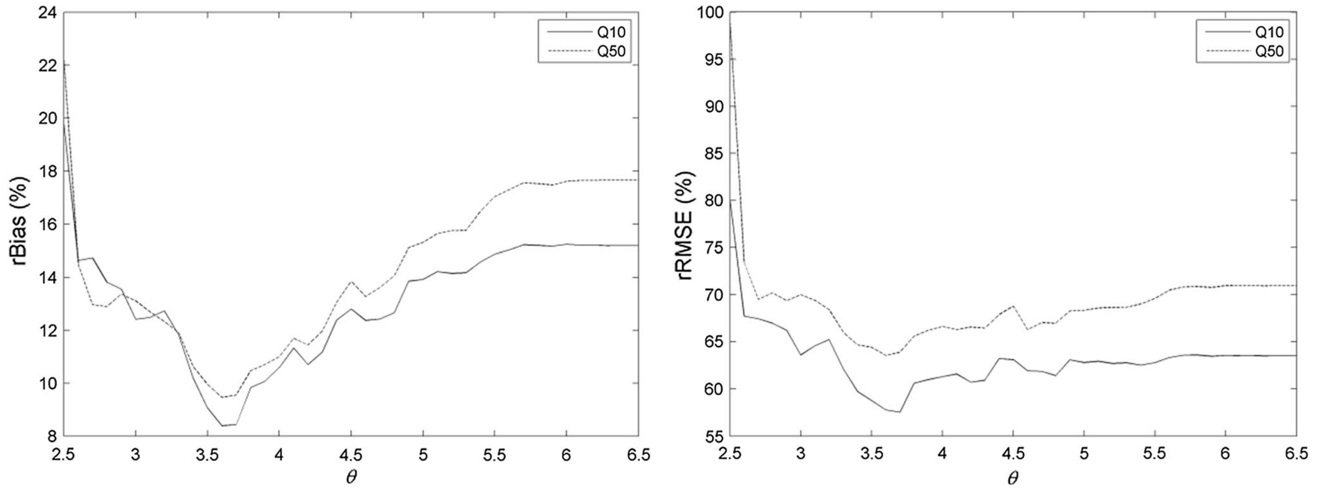

Fig. 7 rBias and rRMSE as a function of the threshold value in ROI

Table 2 Results of the cross validation with the "leave-one-out" method

	Quantiles	GAM	CCA	ROI	LL
R^2	*Q*10	**0.657**	0.550	0.616	0.596
	*Q*50	**0.576**	0.405	0.456	0.480
BIAS	*Q*10	**−28.09**	−58.67	−54.88	−36.10
	*Q*50	**−76.99**	−154.19	−129.01	−85.64
RMSE	*Q*10	**220.62**	251.37	232.32	239.69
	*Q*50	**468.13**	551.07	527.06	518.36
rBIAS (%)	*Q*10	16.87	8.78	**8.39**	15.08
	*Q*50	13.57	**8.10**	9.46	17.73
rRMSE (%)	*Q*10	69.92	60.01	**57.77**	63.32
	*Q*50	**60.34**	66.35	63.57	70.89

Bold values indicate the best statistics

the Evaporation Data CD published by the Australian Bureau of Meteorology (BOM). Similarly, the data for mean annual rainfall was extracted from the BOM CD of Mean Annual Rainfall. The area covered by forest was measured on 1:100,000 topographic maps. Stream density was taken as the sum of all the stream lengths (on a 1:100,000 topographic map) divided catchment area. The data summary of the selected eight predictor variables is presented in Table 1.

4 Results and discussion

4.1 GAM model

In the GAM, predictor variables are selected based on a backward stepwise procedure for Q_{10} as this is more accurate than Q_{50} quantile estimates from at-site flood frequency analysis (e.g. Q_{50} is associated with a higher degree of sampling variability). The predictor variables selected for Q_{10} are then used for Q_{50}. Six predictor variables are found to be statistically significant based on the results of regression analysis: AREA, I6,2, SF, MAE, SDEN and FOREST. In building the prediction equation in GAM, the 'Gaussian family' is adopted with 'identity' link function as this is the most common approach. The log-transformed response variable [e.g. $\ln(Q_{10})$] is used in model building. Figure 2a, b show the GAM model fitting results for Q_{10} and Q_{50}, respectively. It can be seen in these figures that the residuals for Q_{10} model follow more closely a normal distribution than that of Q_{50}. The plot of observed values (i.e. flood quantiles obtained from at-site flood frequency analysis) closely match with the response values (i.e. predicted by the GAM model). Figure 3a, b show that smoothing functions of the predictor variables for Q_{10} and Q_{50} GAM models, revealing that the degree of non-linearity in predictor variables AREA, I6,2 and MAE are quite high as compared with the variables SF, SDEN and FOREST.

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

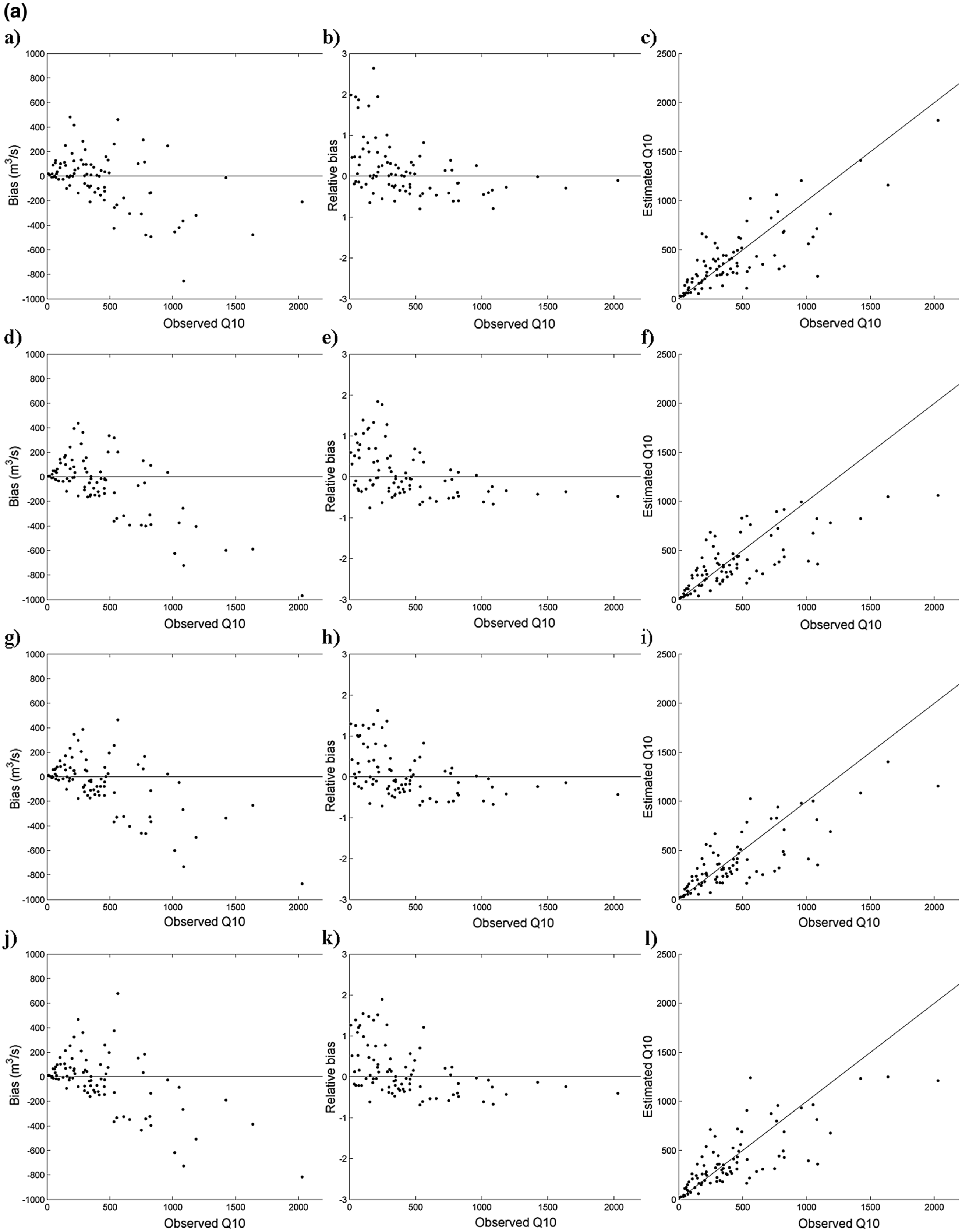

Fig. 8 Graphs of the bias, the relative bias and the estimated versus observed values for **A** Q10 and the models GAM (**a**–**c**), CCA (**d**–**f**), ROI (**g**–**i**) and LL (**j**–**l**). **B** Q50 and the models GAM (**a**–**c**), CCA (**d**–**f**), ROI (**g**–**i**) and LL (**j**–**l**)

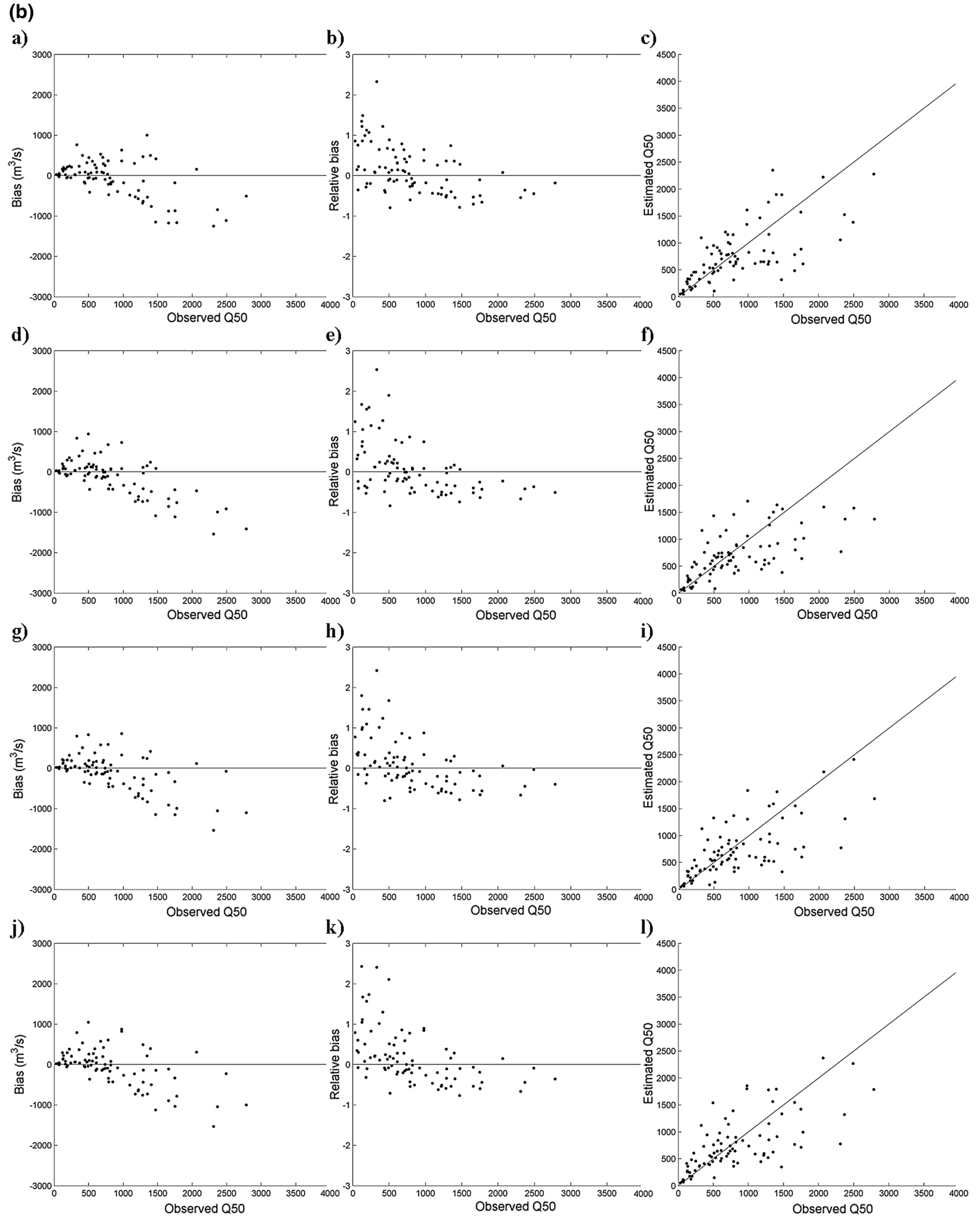

Fig. 8 continued

$$\ln(\mathrm{Q}) = \alpha + \mathrm{s}(\mathrm{AREA}) + s(\mathrm{I}_{6,2}) + s(\mathrm{SF}) + s(\mathrm{MAE}) + s(\mathrm{SDEN}) + s(\mathrm{FOREST}) \quad (8)$$

4.2 Log-linear model (LL)

In developing the log-linear (LL) models, both the response variable (Q_{10} or Q_{50}) and predictor variables are log-transformed. Predictor variables are selected using a backward selection procedure. In the LL model, only four predictors are found to be statistically significant: AREA, $\mathrm{I}_{6,2}$, SF and SDEN; in contrast, for the GAM model, there are two additional predictors as can be seen in Eq. 8.

Figure 4a, b show the LL model fitting results for Q_{10} and Q_{50}, respectively. Here, Q_{50} model residuals follow more closely a normal distribution than that of the Q_{10} model. The plots of response and fitted values for the LL model (Fig. 4a, b) show a higher degree of scatter as compared with that of the GAM model (Fig. 2a, b). The general form of the LL model is given by:

$$\ln(\mathrm{Q}) = b_0 + b_1 \ln(\mathrm{AREA}) + b_2 \ln(\mathrm{I}_{6,2}) + b_3 \ln(\mathrm{SF}) + b_4 \ln(\mathrm{SDEN}) \quad (9)$$

4.3 Canonical correlation analysis (CCA)

A neighbourhood is defined for the target site with the CCA method. The LL model is used for hydrological information transfer. The important predictor variables identified in the LL model are used for both the CCA and the hydrological information transfer. In CCA, physio-meteorological variables are AREA, $\mathrm{I}_{6,2}$, SF, SDEN and the hydrological variables are Q_{10} and Q_{50}. All the variables are normalized with the Box-Cox transformation. Figure 5 shows the physio-meteorological canonical and hydrological canonical space obtained from the analysis. In the optimization of the parameter α controlling the neighbourhood size, no improvement is found in the absolute error indices (BIAS and RMSE) with the neighbourhood approach, and hence optimal parameter is found with relative error indices (rBIAS and rRMSE), as shown in Fig. 6. The optimal parameter in CCA is found to be 0.32.

Regression model for hydrological information transfer in CCA is given by:

$$\ln(\mathrm{Q}) = b_0 + b_1 \ln(\mathrm{AREA}) + b_2 \ln(\mathrm{I}_{6,2}) + b_3 ln(rmSF) + b_4 ln(\mathrm{SDEN}) \quad (10)$$

4.4 Region-of-influence (ROI) approach

In the ROI, regions are delineated in the space of the following physio-meteorological attributes: AREA, $\mathrm{I}_{6,2}$, SF and SDEN (which are same as the predictors in the LL model).

The regression model for hydrological information transfer is given by:

$$\ln(\mathrm{Q}) = b_0 + b_1 \ln(\mathrm{AREA}) + b_2 \ln(\mathrm{I}_{6,2}) + b_3 \ln(\mathrm{SF}) + b_4 \ln(\mathrm{SDEN}) \quad (11)$$

The Euclidian distance between sites i and j is given by:

$$D_{ij} = \left[\sum_{k=1}^{p} (X_{k,i} - X_{k,j})^2\right]^{1/2} \quad (12)$$

where p is the number of attributes considered, and $X_{k,i}$ and $X_{k,j}$ are the standardized values of the k-th attribute at sites i and j respectively. A threshold value θ is defined for which all stations with a distance inferior to the target site are included in the region. No improvement is found in the absolute error indices (BIAS and RMSE) with the ROI approach, and hence the optimal threshold is found with relative indices (rBIAS and rRMSE) (Fig. 7). The optimal Euclidian distance is found to be 3.6.

4.5 Comparison of methods with leave-one-out (LOO) validation

The performances of the four methods adopted in this study (GAM, CCA, ROI and LL) are compared using LOO validation i.e. each of the selected 85 catchments is removed in building the model, and then the developed model is applied to the removed catchment to predict flood quantiles (Q_{10} or Q_{50}). Five model performance statistics are computed for each of the four models using Eqs. 3–7, and the results are summarised in Table 2.

It can be seen from Table 2 that GAM outperforms the CCA, ROI and LL models in terms of R^2, BIAS and RMSE values. However, with respect to rBIAS value, ROI and CCA perform better for Q_{10} and Q_{50}, respectively. ROI and GAM perform better with respect to rRMSE for Q_{10} and Q_{50}, respectively. Figure 8a, b show the plots of the performance statistics for the four RFFA methods for Q_{10} and Q_{50}, respectively. In Fig. 8a, it can be seen that in relation to rBIAS, the GAM model performs better, for medium to large catchments, than for smaller catchments. Indeed, all four methods generally perform better for larger catchments than smaller ones. The plots of predicted and observed quantiles (i.e. quantiles obtained by at-site flood frequency analysis) (Fig. 8a, b) show that GAM estimates generally provide the best match with the observed quantiles. Overall, GAM shows the best performance in LOO validation among the four RFFA methods compared in this study.

It should be noted that ROI approach delivers a relatively homogeneous group of stations and hence generally outperforms fixed region approach. Here, GAM (with a fixed region approach) has outperformed the ROI approach, which clearly highlights the strength of the GAM. Combining ROI and GAM could lead to the flexibility of the GAM as functional fitting and reducing the covariates, because of the greater degree of homogeneity of the ROI regions.

4.6 Predictor variables

In this study, a total of eight predictor variables were selected as described in Sect. 3. Four of these predictor variables (AREA, $I_{6,2}$, SF and SDEN) are found to be significant in the LL model; however, GAM model has selected two additional predictor variables (MAE and FOREST). Two predictor variables (S10,85 and MAE) have not been selected by any of the RFFA methods considered here. Among all the selected variables, AREA and $I_{6,2}$ have been found to be the most influential ones (based on standardised regression coefficients), followed by SF and SDEN. It should be noted that Australian Rainfall and Runoff recommended regional flood estimation model contained only three predictor variables (AREA, $I_{6,2}$ and SF) (Rahman et al. 2015b). The selection of predictor variables is generally governed by adopted optimisation criterion and correlation structure of the predictor variable set. Generally, a model with the smallest number of predictor variables is preferred given the model accuracy is not compromised.

5 Summary and conclusions

Hydrological processes are generally non-linear. However, most of the regional flood frequency analysis (RFFA) methods assume linearity; in this regard, log-linear model is one of the most widely used RFFA model worldwide. In the log-linear model, a log linear relationship between the dependent and predictor variables is assumed, which however may not be satisfied in many applications and generally does not capture the complexity of flood generation processes involved. The application of more sophisticated non-linear methods such as the generalized additive model (GAM) has increased in many fields of science and engineering in recent years to model complex processes. However, the application of GAM has hardly been made in RFFA problems except for one or two instances.

This paper develops a GAM-based RFFA model and compares with three other alternative RFFA models/approaches (log-linear model, canonical correlation analysis and region-of-influence approach). The data from 85 New South Wales catchments in Australia is used in this study. It has been found that some of the most important predictor variables in RFFA such as catchment area and design rainfall intensity are better described by non-linear functions such as thin plate regression splines, allowing a more realistic understanding of the true relationship between the dependent and predictor variables. Based on the leave-one-out validation, it has been found that GAM-based RFFA model generally outperform the other three RFFA models. GAM is found to be performing better even without the neighbourhood/region-of-influence approach. The results of this study reveal that GAM is a viable modelling option in RFFA that is easy to implement, and which generally requires a reduced number of assumptions. The finding of this study is expected to encourage other researchers worldwide to apply GAM in RFFA studies.

Acknowledgements The authors acknowledge Australian Rainfall and Runoff (ARR) Revision team for sharing some of the data used in this study, Australian Bureau of Meteorology for providing design rainfall data and NSW Office of Water for providing streamflow data and ARR Project 5 team Dr Khaled Haddad, Dr Md Mahmudul Haque, Professor George Kuczera, Mr Erwin Weinmann, Dr William Weeks, Mr Mark Babister and Assoc Prof James Ball for their input in preparation of the database used in this study.

References

Alobaidi MH, Marpu PR, Ouarda TBMJ, Chebana F (2015) Regional frequency analysis at ungauged sites using a two-stage resampling generalized ensemble framework. Adv Water Resour 84:103–111. doi:10.1016/j.advwatres.2015.07.019

Asquith WH, Herrmann GR, Cleveland TG (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. J Hydrol Eng 18:1331–1348

Aziz K, Rahman A, Fang G, Shreshtha S (2014) Application of artificial neural networks in regional flood frequency analysis: a case study for Australia. Stoch Environ Res Risk Assess 28(3):541–554

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

Aziz K, Haque MM, Rahman A, Shamseldin AY, Shoaib M (2016) Flood estimation in ungauged catchments: application of artificial intelligence based methods for Eastern Australia. Stoch Environ Res Risk Assess. doi:10.1007/s00477-016-1272-0

Ball J, Babister M, Nathan R, Weeks W, Weinmann E, Retallick M, Testoni I (2016) Australian rainfall and runoff: a guide to flood estimation. Commonwealth of Australia, Canberra

Bates BC, Rahman A, Mein RG, Weinmann PE (1998) Climatic and physical factors that influence the homogeneity of regional floods in south-eastern Australia. Water Resour Res 34(12):3369–3381

Bayentin L, Adlouni SE, Ouarda T, Doyon B, Chebana F (2010) Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989–2006 in Quebec, Canada. Int J Health Geogr 9:5

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. Adv Meteorol 2012:1–16

Blöschl G, Sivapalan M, Wagener T, Viglione A, Savenije H (2013) Runoff prediction in ungauged basins: synthesis across processes, places and scales. Cambridge University Press, New York

Burn DH (1990) An appraisal of the "region of influence" approach to flood frequency analysis. Hydrol Sci J 35(2):149–165

Castellarin A, Burn DH, Braith A (2008) Homogeneity testing: how homogenous do heterogeneous cross-correlated regions seem. J Hydrol 360(1–4):67–96

Chebana F, Ouarda TBMJ (2007) Multivariate L-moment homogeneity test. Water Resour Res 43:W08406. doi:10.1029/2006WR005639,1-14

Chebana F, Ouarda TBMJ (2009) Index flood—based multivariate regional frequency analysis. Water Resour Res 45:W10435. doi:10.1029/2008WR007490

Chebana F, Charron C, Ouarda TBMJ, Martel B (2014) Regional frequency analysis at ungauged sites with the generalized additive model. J Hydrometeorol 15:2418–2428

Clifford S, Low Choy S, Hussein T, Mengersen K, Morawska L (2011) Using the generalised additive model to model the particle number count of ultrafine particles. Atmos Environ 45:5934–5945

Cunderlik JM, Burn DH (2006) Site-focused nonparameteric test of regional homogeneity based on flood regime. J Hydrol 318(1–4):276–291

Dawson CW, Abrahart RJ, Shamseldin AY, Wilby RL (2006) Flood estimation at ungauged sites using artificial neural networks. J Hydrol 319:391–409

Durocher M, Chebana F, Ouarda TBMJ (2015) A nonlinear approach to regional flood frequency analysis using projection pursuit regression. J Hydrometeorol 16(4):1561–1574. doi:10.1175/jhm-d-14-0227.1

Durocher M, Chebana F, Ouarda TBMJ (2016) On the prediction of extreme flood quantiles at ungauged locations with spatial copula. J Hydrol 533:523–532

Eng K, Milly PCD, Tasker GD (2007) Flood regionalization: a hybrid geographic and predictor-variable region-of-influence regression method. J Hydrol Eng 12(6):585–591

Fill HD, Stedinger JR (1995) Homogeneity tests based upon Gumbel distribution and a critical appraisal of Dalrymple's test. J Hydrol 166(1–2):81–105

Flavell D (2012) Design flood estimation in Western Australia. Aust J Water Resour 16(1):1–20

Galiano SGG, Gimenez PO, Giraldo-Osorio JD (2015) Assessing nonstationary spatial patterns of extreme droughts from long-term high-resolution observational dataset on a semiarid basin (Spain). Water 7:5458–5473

Girard C, Ouarda TBMJ, Bobée B (2004) Étude du biais dans le modèle log-linéaire d'estimation régionale—study of the bias in the log-linear regional estimation model. Can J Civ Eng 31:1–8

Griffis VW, Stedinger JR (2007) The use of GLS regression in regional hydrologic analyses. J Hydrol 204:82–95

Guan BT, Hsu HW, Wey TH, Tsao LS (2009) Modeling monthly mean temperatures for the mountain regions of Taiwan by generalized additive models. Agric For Meteorol 149:281–290

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. J Hydrol 430–431(2012):142–161

Haddad K, Rahman A, Weinmann PE, Kuczera G, Ball JE (2010) Streamflow data preparation for regional flood frequency analysis: lessons from south-east Australia. Aust J Water Resour 14(1):17–32

Haddad K, Rahman A, Stedinger JR (2012) Regional flood frequency analysis using Bayesian generalized least squares: a comparison between quantile and parameter regression techniques. Hydrol Process 26:1008–1021

Haddad K, Rahman A, Zaman M, Shrestha S (2013) Applicability of Monte Carlo cross validation technique for model development and validation using generalised least squares regression. J Hydrol 482:119–128

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

Hastie T, Tibshirani R (1986) Generalized additive models. Stat Sci 1:297–310

Hosking JRM, Wallis JR (1993) Some statistics useful in regional frequency analysis. Water Resour Res 29(2):271–281

IE Aust (1987) Australian rainfall and runoff—a guide to flood estimation. Engineers Australia, Canberra

Ishak E, Rahman A, Westra S, Sharma A, Kuczera G (2013) Evaluating the non-stationarity of Australian annual maximum floods. J Hydrol 494:134–145

Kauermann G, Opsomer JD (2003) Local likelihood estimation in generalized additive models. Scand J Stat 30:317–337

Kim D, Cho H, Onof C et al (2016) Let-it-rain: a web application for stochastic point rainfall generation at ungaged basins and its applicability in runoff and flood modelling. Stoch Environ Res Risk Assess. doi:10.1007/s00477-016-1234-6

Kovalchuk SV, Krikunov AV, Knyazkov KV et al (2016) Classification issues within ensemble-based simulation: application to surge floods forecasting. Stoch Environ Res Risk Assess. doi:10.1007/s00477-016-1324-5

Kuczera G (1999) Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference. Water Resour Res 35(5):1551–1557

Kuczera G, Franks S (2015) At-site flood frequency analysis. In: Ball J (ed) Australian rainfall and runoff. Engineers Australia. http://book.arr.org.au/

Lamontagne JR, Stedinger JR, Cohn TA, Barth NA (2013) Robust national flood frequency guidelines: what is an outlier? In: World environmental and water resources congress, pp 2454–2466

Leitte AM, Petrescu C, Franck U, Richter M, Suciu O, Ionovici R, Herbarth O, Schlink U (2009) Respiratory health, effects of ambient air pollution and its modification by air humidity in Drobeta–Turnu Severin, Romania. Sci Total Environ 407:4004–4011

Merz R, Blöschl G (2005) Flood frequency regionalisation—spatial proximity vs. catchment attributes. J Hydrol 302:283–306

Micevski T, Kuczera G (2009) Combining site and regional flood information using a Bayesian Monte Carlo approach. Water Resour Res 45:W04405. doi:10.1029/2008WR007173

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. Hydrol Process 29(2):250–260

Morlini I (2006) On multicollinearity and concurvity in some nonlinear multivariate models. Stat Methods Appl 15:3–26

Morton R, Henderson BL (2008) Estimation of nonlinear trends in water quality: an improved approach using generalized additive models. Water Resour Res 44:W07420. doi:10.1029/2007WR006191

Motevalli A, Vafakhah M (2016) Flood hazard mapping using synthesis hydraulic and geomorphic properties at watershed scale. Stoch Environ Res Risk Assess 30:1889. doi:10.1007/s00477-016-1305-8

Ouali D, Chebana F, Ouarda TBMJ (2015) Non-linear canonical correlation analysis in regional frequency analysis. Stoch Environ Res Risk Assess. doi:10.1007/s00477-015-1092-7

Ouali D, Chebana F, Ouarda TBMJ (2016) Quantile regression in regional frequency analysis: a better exploitation of the available information. J Hydrometeorol. doi:10.1175/JHM-D-15-0187.1

Ouarda TBMJ (2013) Regional hydrological frequency analysis. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of environmetrics. Wiley, New York

Ouarda TBMJ, Shu C (2009) Regional low-flow frequency analysis using single and ensemble artificial neural networks. Water Resour Res 45:W11428. doi:10.1029/2008WR007196

Ouarda TBMJ, Girard C, Cavadias GS, Bobée B (2001) Regional flood frequency estimation with canonical correlation analysis. J Hydrol 254:157–173

Ouarda TBMJ, Cunderlik JM, St-Hilaire A, Barbet M, Bruneau P, Bobée B (2006) Data-based comparison of seasonality-based regional flood frequency methods. J Hydrol 330:329–339

Ouarda TBMJ, St-Hilaire A, Bobée B (2008a) Synthèse des développements récents en analyse régionale des extrêmes hydrologiques/A review of recent developments in regional frequency analysis of hydrological extremes. Revue des sciences de l'eau/J Water Sci 21:219–232

Ouarda TBMJ, Ba KM, Diaz-Delgado C, Carsteanu A, Chokmani K, Gingras H, Quentin E, Trujillo E, Bobée B (2008b) Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. J Hydrol 348:40–58

Ouarda TBMJ, Charron C, Marpu PR, 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 J Sel Top Appl Earth Obs Remote Sens 9(4):1553–1566. doi:10.1109/JSTARS.2016.2522764

Pandey GR, Nguyen VTV (1999) A comparative study of regression based methods in regional flood frequency analysis. J Hydrol 225:92–101

Rahman A (2005) A quantile regression technique to estimate design floods for ungauged catchments in south-east Australia. Aust J Water Resour 9(1):81–89

Rahman A, Bates BC, Mein RG, Weinmann PE (1999) Regional flood frequency analysis for ungauged basins in south-eastern Australia. Aust J Water Resour 3(2):199–207

Rahman SA, Rahman A, Zaman M, Haddad K, Ashan A, Imteaz MA (2013) A study on selection of probability distributions for at-site flood frequency analysis in Australia. Nat Hazards 69:1803–1813

Rahman A, Haddad K, Haque M, Kuczera G, Weinmann PE (2015a) Australian rainfall and runoff project 5: regional flood methods: stage 3 report, technical report, No. P5/S3/025, Engineers Australia, Water Engineering, 134 pp

Rahman A, Haddad K, Kuczera G, Weinmann PE (2015b) Regional flood methods. In: Ball JE (ed) Australian rainfall & runoff, chapter 3, Book 3. Engineers Australia. http://book.arr.org.au/

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 Infect Dis 9:190

Shortridge JE, Guikema SD, Zaitchik BF (2015) Empirical streamflow simulation for water resource management in data-scarce seasonal watersheds. Hydrol Earth Syst Sci Discuss 12:11083–11127

Shu C, Ouarda TBJM (2007) Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resour Res 43:W07438

Shu C, Ouarda TBMJ (2008) Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. J Hydrol 349:31–43

Stedinger JR, Tasker GD (1985) Regional hydrologic analysis. 1. Ordinary, weighted, and generalised least squares compared. Water Resour Res 21(9):1421–1432

Tisseuil C, Vrac M, Lek S, Wade AJ (2010) Statistical downscaling of river flows. J Hydrol 385:279–291

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. Environ Health 8:3

Wang Y, Jianzhu L, Feng P, Hu R (2015) A time-dependent drought index for non-stationarity precipitation series. Water Resour Manage 29(15):5631–5647

Wazneh H, Chebana F, Ouarda TBMJ (2013) Depth-based regional index-flood model. Water Resour Res 49(12):7957–7972. doi:10.1002/2013wr013523

Wen L, Rogers K, Saintilan N, Ling J (2011) The influences of climate and hydrology on population dynamics of waterbirds in the lower Murrumbidgee River floodplains in Southeast Australia: implications for environmental water management. Ecol Model 222:154–163

Westra S, Fowler HJ, Evans JP, Alexander LV, Berg P, Johnson F, Kendon EJ (2014) Future changes to the intensity and frequency of short-duration extreme rainfall. Rev Geophys 52:522–555. doi:10.1002/2014RG000464

Wood SN (2003) Thin plate regression splines. J R Stat Soc Ser B Stat Methodol 65:95–114

Wood SN (2006) Generalized additive models: an introduction with R. Chapman and Hall/CRC Press, Florida

Wood SN (2008) Fast stable direct fitting and smoothness selection for generalized additive models. J R Stat Soc Ser B Stat Methodol 70:495–518

Wood SN, Augustin NH (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol Model 157:157–177