Impacts of Outliers in Flood Frequency Analysis: A Case Study for Eastern Australia

A.S. Rahman [1,*], K. Haddad [1], A. Rahman [1]

[1]School of Computing, Engineering and Mathematics, University of Western Sydney,
Building XC, Kingswood Campus, Locked Bag1797, Penrith, New South Wales 2751, Australia

Peer Review History[1]

Abstract: *At-site flood frequency analysis is a useful technique to estimate flood quantiles if reasonably long at-site flood record is available. In Australia, FLIKE software has been proposed for at-site flood frequency analysis. The advantage of FLIKE is that, for a given application, the user can compare a number of most commonly adopted probability distributions and parameter estimation methods relatively quickly using a windows interface. The new version of FLIKE has been incorporated with the multiple Grubbs and Beck test, which can identify multiple numbers of potentially influential low flows. This paper presents a case study considering ten catchments in eastern Australia, which compares two outlier identification tests (original Grubbs and Beck and multiple Grubbs and Beck tests) and two commonly applied probability distributions (Generalized Extreme Value (GEV) and Log Pearson type 3 (LP3)). The results show that the LP3 distribution with multiple Grubbs and Beck test provides more accurate flood quantile estimates than when LP3 distribution is used with the original Grubbs and Beck test. Between these two methods, the differences in flood quantile estimates have been found to be up to 61% for the ten study catchments. It has also been found that GEV distribution with L moments and LP3 distribution with the multiple Grubbs and Beck test provide quite similar results in most of the cases; however, a difference up to 38% has been noted for flood quantiles for annual exceedance probability (AEP) of 1 in 100 for one catchment. The methodology presented in this paper can be applied to other catchments/countries.*

Keywords: *Flood, FLIKE, probability distributions, flood frequency, outlier.*

1. Introduction

Flooding is a part of the natural cycle of many ecosystems and plays an important role in maintaining ecosystem function and biodiversity (Poff et al., 1977). The multitude of environmental benefits from flooding also benefits society through continuation of ecosystem services (Bayley, 1995). However, major modification to natural processes and ecosystems through human activities has created some of the most destructive hydro-meteorological phenomena in terms of their impacts on human well-being and socioeconomic activities. There had been significant increases in the total number of natural disaster events in recent years due to climate change. Flood is the most expensive of these disasters, for example, during the period of 2010-2011, a series of floods affected various parts of Australia, with most significant damage to the state of Queensland causing damage over $5 billion.

Flood frequency analysis is the most direct method of estimating design floods, which is needed to design bridges, culverts, flood levees and other drainage infrastructure and in various water resources management tasks such as flood plain management and flood insurance studies. Griffis and Stedinger (2007) found that estimates of magnitude and frequency of floods using streamflow-gauging stations with shorter records of annual peak flow data had higher standard errors or uncertainties when compared to estimates using stream gauges with longer annual peak flow records. Flood estimation should get the maximum information from the available data, be robust with respect to the distribution model and potentially influential low flows (PILFs). However, streamflow record length at many sites is often insufficient and identification of PILFs becomes a major issue in fitting a probability distribution to available flood data.

[1] *Paper JHER0202, submitted on 28/10/2014, accepted for publication after peer review and subsequent revisions on 23/12/2014*

[*] *Corresponding author may be contacted at ayesha.rahman@uws.edu.au*

Comparing various probability distributions and parameter estimation procedure had been done in numerous occasions in the past; however, due to the limited length of observed flood data as compared to the return period of interest, flood frequency analysis is deemed to be a difficult task and often associated with controversies (Bobée et al., 1993). The selection of an 'appropriate' probability distribution and associated parameter estimation procedure is an important step in flood frequency analysis. Flood frequency analysis has been widely researched in the past (e.g. Vogel et al., 1993; Onoz and Bayazit, 1995; Bates et al., 1998; Laio, 2004; Merz et al., 2008; Meshgi and Khalili, 2009a, b; Laio et al., 2009; Ishak et al., 2010, 2011; Haddad et al., 2011, Haddad et al., 2012; Haddad and Rahman, 2012; ; Haddad et al., 2013; Rahman et al., 2013). In flood frequency analysis, a probability distribution is often selected on the basis of statistical tests or by graphical methods, and convenience plays an important role in this choice (Bobée et al., 1993). In practical applications, empirical suitability plays a much larger role in distributional choice than a priori reasoning (Cunnane, 1985; 1989).

One of the earliest studies on the search for the probability distribution of floods was done by Benson (1968). He considered two-parameter Gamma, Gumbel (EV1), Log Gumbel, Log Normal (LN2), three parameter Log Pearson Type 3 (LP3) distribution for flood data of 10 stations in various parts of the USA. The standardized average deviations were found to be high for the gamma, EV1 and Log Gumbel, but lower for the LN2 and LP3. Among these, the LP3 distribution was preferred for being in common use, and for being capable of fitting skewed data. The conclusions and recommendations of this study led to the wide-scale adoption of the LP3 distribution in the USA. In Australia, an extensive study was done for Queensland data by Kopittke et al., (1976), and another by Conway (1970) for New South Wales coastal streams. They concluded that the LP3 distribution performed the best among a number of distributions examined. Beard (1974) estimated the 1000 year flood at 300 stations in the USA with 14200 station-years of data by eight different models (LN2, gamma, EV1 with two different parameter estimates, Log Gumbel, Pearson type 3 (P3), LP3 and regional LP3). LP3 and LN2 came close to reproducing the expected 14 exceedances and were concluded to be the preferred ones. The split sample validation confirmed the superiority of these distributions.

To compare various probability distributions using the data from 172 catchments in Australia, McMahon and Srikanthan (1981) used the moment ratio diagrams. They also concluded that the LP3 was the most suitable distribution for Australia. Based on the findings of these studies, it was recommended in Australian Rainfall and Runoff (ARR) (I. E. Aust., 1987) that flood frequency analysis in Australia (I. E. Aust., 1987) should follow the footsteps of the USA i.e. to use LP3 distribution (IAWCD, 1982).

Since the publication of ARR (I. E. Aust., 1987, 2001), there have been a number of studies to compare various probability distributions (Rahman et al., 1999). For example, Nathan and Weinmann (1991) examined 53 catchments from Central Victoria, with L-moments-based goodness-of-fit test, and found that the GEV distribution was the best-fit distribution. Vogel et al. (1993) compared a number of distributions using data from 61 stations in Australia, using the L-moments ratio diagram; they found that the generalized Pareto distribution (GPA) was the best-fit distribution followed by the GEV, LN3, and LP3. Haddad and Rahman (2008) compared a number of distributions and parameter estimation procedures for 18 catchments in southeast Australia and found that the GEV distribution was the best-fit distribution for the selected catchments. In another study, Haddad and Rahman (2010) found that the two parameter distributions are preferable to Tasmania, with the lognormal appearing to be the best-fit distribution for Tasmania. As it seems an analyst might choose a different frequency model and fitting procedure for each catchment, but this could lead to inconsistencies in flood estimates across regions and among governmental agencies. National consistency in flood frequency estimates is important because these estimates are used in the allocation of resources and the implementation of the National Flood Insurance Program (Thomas, 1985; Cohn et al., 2013). For this reason, a national methodology should exhibit the characteristic of robustness, which in this context means that the analysis does not perform poorly when its assumptions are not fully satisfied.

An important step in flood frequency analysis is the detection of the PILFs in the flood data (Saf, 2010). PILFs are unusually small observations of flood data which depart significantly from the trend of the rest of the data. Identification and treatment of PILFs are important issues in flood frequency analysis, because such observations can have a large influence on the estimate of extreme flood quantiles. In arid regions, even when it rains, channel losses can result in annual flood peaks that are zero or nearly zero, therefore LP3 distribution cannot fit the entire flood record without censoring zero values. Furthermore, unusually small values can result in relatively poor estimates of the large flood quantiles. In frequency analyses, one often uses a probability plot to examine if the sample data is consistent with a fitted curve (Beckman and Cook, 1983; Stedinger et al., 1993), unfortunately such decisions can be relatively subjective. The Bulletin 17B explicitly notes that not dealing with this issue of PILFs would "significantly affect the [computed] statistical parameters." Both Barnett and Lewis (1994) and Beckman and Cook (1983) discussed the notion of using outlier tests to identify unusual (high or low) data points that otherwise might have

undue influence in flood frequency analysis. Using a good low-outlier identification procedure has the potential for making low-outlier identification less subjective, by providing "rejection criteria which enable significance to be assessed" (Barnett and Lewis, 1994).

A wide range of test procedures for identifying PILFs has been examined in the past (e.g. Thompson, 1935; Grubbs, 1969; Grubbs and Beck, 1972; Barnett and Lewis, 1994), including methods for dealing with the case of multiple PILFs considered here (Tietjen and Moore, 1972; Rosner, 1975, 1983; Prescott, 1975, 1978; Gentleman and Wilk, 1975; Marashinghe, 1985; Rousseeuw and Zomeren, 1990; Hadi and Simonoff, 1993; Spencer and McCuen, 1996; Rousseeuw and Leroy, 2003; Verma and Quiroz-Ruiz, 2006). Thompson (1935) provided an early criterion for the rejection of an outlier based on the ratio of the sample standard deviation and an observation's deviation from the sample mean. An alternative test was proposed by Dixon (1950, 1951), who for high outliers proposed the test statistics for second most extreme observation in either tail of the distribution. Barnett and Lewis (1994) also noted similar criteria as Dixon (1950, 1951). Grubbs (1969) and Grubbs and Beck (1972) proposed a one sided 10% significance level criteria to identify PILFs. Rosner (1975, 1983) developed a sequential two sided outlier test, based on a generalization of the Grubbs (1969) which usually detected outliers either too small or large. This procedure was found to be less computationally intensive and easy to apply in practice.

Bulletin 17B (IAWCD 1982) was the guideline for flood frequency analysis in the United States for more than 30 years. Recently, there has been an attempt to revise Bulletin 17B to include recent advances in statistical techniques and computational resources (IAWCD 2013) similar to the current revision of Australian Rainfall and Runoff. In Bulletin 17C, a new low outlier identification procedure, the multiple Grubbs-Beck (MGB) test (Lamontagne et al., 2013) has been included. The MGB test is based on significance levels computed using the new approximations developed by Cohn et al. (2013). The MGB test is a generalization of the old Bulletin 17B original Grubbs-Beck (GB) test (Grubbs 1969; Grubbs and Beck, 1972).

In this paper the original GB test and the MGB test are compared using data from Australian catchments. The Bulletin 17B's original GB test was based only on the distribution of the single smallest observation in a sample. As a result, even though multiple PILFs in flood annual maximum flood data series may exist, the original GB test rarely identifies more than a single PILF. The MGB test employs the actual distribution of the k^{th} smallest observation in a sample of n independent normal variates based upon significance levels provided by Cohn et al. (2013), and is thus suited to test for multiple PILFs.

Kuczera (1999) presented a comprehensive study on flood frequency analysis using Bayesian method and incorporated a number of probability distributions in his FLIKE software. It has several advantages including the ability to (i) incorporate prior or regional information; (ii) incorporate stage-discharge uncertainty; (iii) assess parameter uncertainty obtained from regional information; and (iv) allow for threshold values (censoring). Recently a new version of FLIKE has been released. The older version of FLIKE needed manual identification of PILFs using the original GB test. The new version of FLIKE is incorporated with the MGB test which attempts to identify multiple PILFs in the annual maximum flood series data.

The objective of this paper is to compare the performance of two probability distributions in flood frequency analysis (FFA) namely LP3 and GEV distributions with a special focus on the effects of censoring PILFs using the original GB and MGB tests. To our knowledge, the MGB test has not been applied before to Australian flood data. The reason for applying the LP3 and GEV distributions with the MGB test is that these two distributions are most commonly adopted in flood and rainfall frequency analyses in Australia.

2. Study area and data preparation

For this study, ten catchments from eastern Australia are selected from the states of New South Wales (NSW), Queensland (QLD) and Victoria (VIC) as shown in Table 1 and Figure 1. Catchment area ranges from 87 to 900 km^2 with a mean of 345.8 km^2 and median of 160 km^2. Record length ranges from 33 to 91 years with a mean of 56 years and median of 54 years. All of the stations have log-space skew values significantly different from zero. Missing data points in the annual maximum flood series were in-filled where possible by two methods. Method one involved comparing the monthly instantaneous maximum data (IMD) with monthly maximum mean daily data (MMD) at the same station. If a missing month of IMD flow corresponded to a month of very low MMD flow, then that was taken to show that the annual maximum did not occur during that missing month. Method 2 involved a simple linear regression of the annual MMD flow against the annual IMD series of the same station. It must be mentioned that the regression equations developed were used for filling gaps in the IMD record, but not to extend the overall period of record.

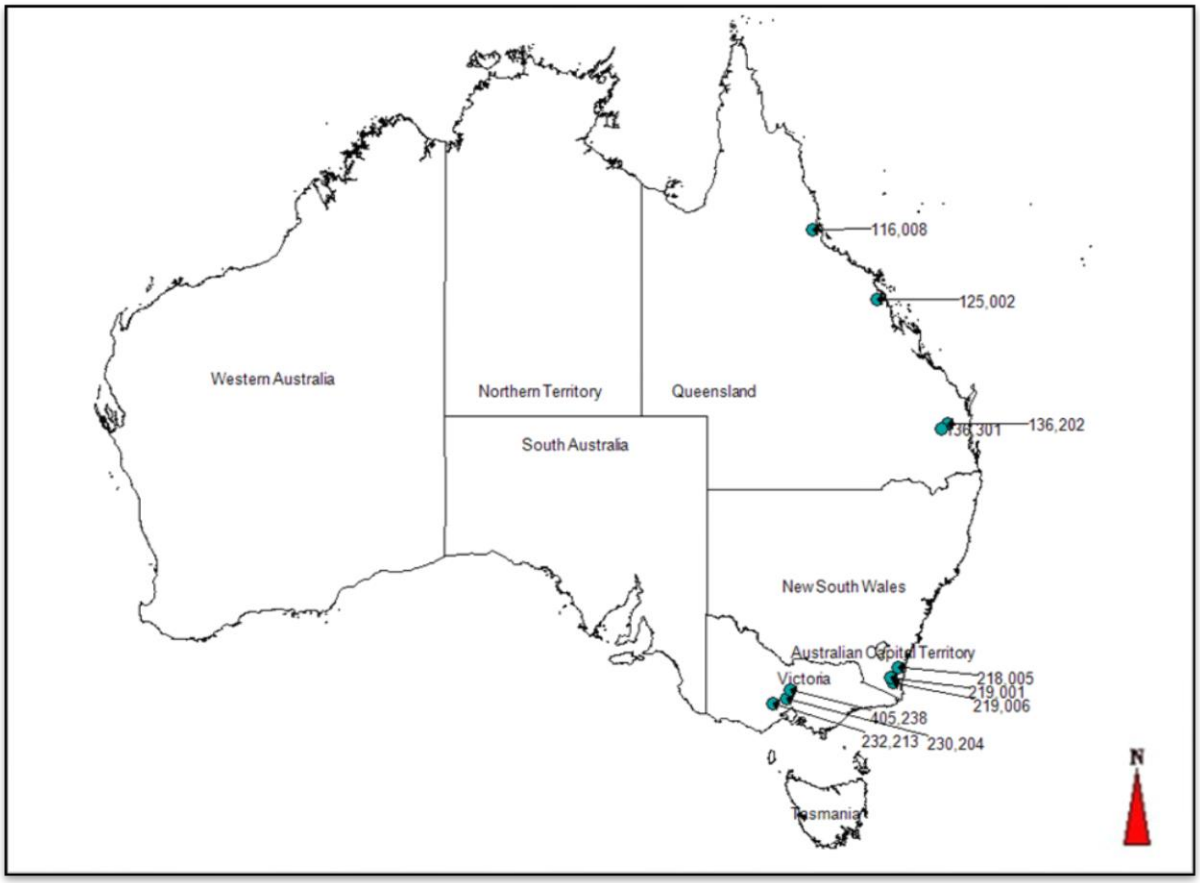

Figure 1 Selected ten catchments from eastern Australia

Table 1 Details of the selected study catchments

Station ID	Station name	River name	Catchment area (km^2)	Record length (years)	Period of record
218005	D/S Wadbilliga R Junct	Tuross	900	47	1965-2011
219001	Brown Mountain	Rutherford Ck	15	62	1949-2010
219006	Tantawangalo Mountain (Dam)	Tantawangalo Ck	87	60	1952-2011
116008	Abergowrie	Gowrie Ck	124	58	1954-2011
125002	Sarich's	Pioneer	740	51	1961-2011
136202	Litzows	Barambah Ck	681	91	1921-2011
136301	Weens Br	Stuart	512	76	1936-2011
230204	Riddells Ck	Riddells Ck	79	38	1974-2011
232213	U/S of Bungal Dam	Lal Lal Ck	157	33	1977-2011
405238	Pyalong	Mollison Ck	163	41	1972-2012

3. Methodology

The original GB test (Grubbs, 1969; Grubbs and Beck, 1972) uses the at-site logarithms of the peak-flow data to calculate a one-sided, 10% significance-level critical value for a normally distributed sample. Although more than one recorded peak flow for a stream gauge may be smaller than the Grubbs-Beck critical value, usually only one non-zero recorded peak flow is identified from the test as being a PILF. The original GB test which was recommended in Bulletin 17B (IAWCD, 1982) defines a low outlier (PILF) threshold as:

$$X_{crit} = \mu - k_n \sigma \quad (1)$$

where k_n is a one-sided, 10% significance-level critical value for an independent sample of n normal variate, and μ and σ denote the sample mean and standard deviation of the entire data set, respectively. Any observation less than X_{crit} is declared a "low outlier (PILF)" (IAWCD, 1982). As per Bulletin 17B, PILFs are omitted from the sample and the frequency curve is adjusted, using a conditional probability adjustment (IAWCD, 1982). The k_n values are tabulated in section A4 of IACWD (1982) based on Table A1 in Grubbs and Beck (1972).

Stedinger et al (1993) provide an approximation of k_n for $5 \leq n \leq 150$ (where n is sample size):

$$k_n \approx -0.9043 + 3.345\sqrt{\log_{10}(n)} - 0.4046\log_{10}(n) \quad (2)$$

The original GB test only identifies one outlier/PILF at a time from a particular data set, but there can be more numbers of PILFs available in the data. A method for statistically detecting multiple PILFs using a generalized Grubbs-Beck test has been developed (Gotvald et al., 2012). The MBG test is based on a one-sided, 10% significance-level critical value for a normally distributed sample, but the test is constructed so that groups of ordered data are examined (for example, the eight smallest values) and excluded from the dataset when the critical value is calculated. If the critical value is greater than the eighth smallest value in the example, then all eight values are considered to be PILFs according to this new method.

Here, one considers whether $\{X_{[1:n]}, X_{[2:n]}, \ldots, X_{[k:n]}\}$ are consistent with a normal distribution and the other observations in the sample by examining the statistic (Cohn et al., 2013):

$$\tilde{\omega}_{[k:n]} \equiv \left(k_{[k:n]} - \mu_k\right)/\sigma_k \quad (3)$$

where $X_{[k:n]}$ denotes the k^{th} smallest observation in the sample, and

$$\mu_k \equiv \frac{1}{n-k}\sum_{j=k+1}^{n} X_{[j:n]} \quad (4)$$

$$\sigma_k \equiv \frac{1}{n-k-1}\sum_{j=k+1}^{n}\left(X_{[j:n]} - \mu_k\right)^2 \quad (5)$$

Here the partial mean μ_k and partial variance σ_k are computed using only the observations larger than $X_{[k:n]}$ to avoid swamping.

To implement the MGB test, recommended for Bulletin 17C, the following two steps are involved: (i) starting at the median and sweeping outward towards the smallest observation, each observation is tested with a MGB test significance level α_{out}. If the k^{th} smallest observation is identified as a low outlier, the outward sweep stops and all observations less than the k^{th} smallest (i.e. j = 1, ..., k) are also identified as low outliers. (ii) An inward sweep always starts at the smallest observation and moves towards the median, with a significance level of α_{in}. If an observation $m \geq 1$ fails to be identified by the inward sweep, the inward sweep stops. The total number of low-outliers/PILFs identified by the MGB test is then the maximum of k and m – 1. The algorithm has two parameters that need to be specified (Cohn et al., 2013): (i) outward sweep significance level for each comparison, α_{out}; and (ii) inward sweep significance level for each comparison, α_{in}.

Bulletin 17B used a 10% significance test with a single outlier threshold. The new outlier detection procedure uses two multiple threshold sweeps. Those thresholds are the Cohn et al. (2013) p(k;n) function which correctly describes if the k^{th} smallest observation in a normal sample of n variates is unusual. The first outward sweep seeks to determine if there is some break in the lower half of the data that would suggest the sample is best treated as if it had a number of low outliers. The second sweep using a less severe significance level, say $p(k;n) \leq$ 10%, mimics Bulletin 17B's willingness to identify one or more of the smallest observations as low outliers so that the frequency analysis is more robust.

A reasonable concern is that a flood record could contain more than one low outlier and the additional outliers can cause the original GB test statistic to fail to recognize the smallest observation as an outlier (by inflating the sample mean and variance). This effect is known as masking (Tietjen and Moore, 1972). Inward sweep tests are particularly susceptible to masking (Branett and Lewis, 1994); therefore an

outward sweep is desirable to avoid the masking problem (Spencer and McCuen, 1996). Rosner (1983) used a two-sided outward sweep. McCuen and Ayuub (1996) recommended an outward sweep with their test for multiple outliers when fitting a LP3 distribution.

The GEV distribution is a family of continuous probability distributions developed within extreme value theory to combine the Gumble, Frèhet and Weibull families also known as type I, II and III extreme value distributions. Here GEV distribution is used to compare the results with the LP3 distribution. The LP3 distribution has been used for several decades to model annual maximum flood series. Estimation of the parameters of the distribution using a method of moment (MOM) estimator in log space was suggested by Beard (1962); this method was used presumably for computational ease. The only complication was the need for frequency factors to compute quantile estimates given the sample moments of the logs of the data. The needed frequency factors were tabulated in Benson (1968) and in Bulletin 17B. Kirby (1972) provided an excellent approximation. Currently, these can be computed directly with built-in functions in many software packages, including Excel and MATLAB.

To fit the LP3 distribution, it is required to calculate the mean, standard deviation and skew coefficient of the logarithms of the annual maximum flood data. Estimate of the p percent annual exceedance probability (AEP) flood is computed by inserting the three statistics of the frequency distribution into the equation:

$$\log Q_p = \overline{X} + K_p S \tag{6}$$

where Q_P is the p-percent AEP flood or flood quantile; $\overline{X}$ the mean of the logarithms of the annual peak flows; k_p is the frequency factor that depends on the skew coefficient and AEP and can be obtained from Bulletin 17B and S is the standard deviation of the logarithms of the annual peak flows.

The mean, standard deviation and skew coefficient can be estimated from the available sample data (recorded annual-peak flows), but a skew coefficient calculated from small samples tends to be an unreliable estimator of the population skew coefficient. Accordingly, the guidelines in Bulletin 17B (IAWCD, 1982) indicates that the skew coefficient calculated from at-site sample data (station skew) needs to be weighted with a generalized or regional skew determined from an analysis of selected long-term stream gauges in the study region. The value of the skew coefficient used in equation 6 is the weighted skew that is based on station skew and regional skew. However, Australian Rainfall and Runoff 1987 did not adopt the weighted skew for application in Australia (I. E. Aust., 1987).

4. Results

For five stations (218005, 219001, 219006, 116008, 136301), the original GB test did not find any PILF but the MGB test found 24, 26, 27, 29, 31 PILFs, respectively and for the remaining five stations (125002, 136202, 230204, 232213, 405238), the original GB test found only one PILF for each of them but the MGB test found 26, 46, 17, 17, 21 PILFs, respectively. These results show a remarkable difference between the results by the two methods of outlier detection.

Table 3 presents the log space skews of the original annual maximum (AM) flood data set and after removing the PILFs using the original GB test and the MGB test. This table shows that application of MGB results in a greater reduction in log space skew than the original GB test. This is likely to affect the quantile estimation by LP3 distribution as skew plays an important role in the fitting of LP3 distribution.

Figures 2 and 3 show how does application of GB and MGB tests affect the fitting of a probability distribution to the AM flood series for Station 136202. The application of GB test did not identify any PILF for Station 136202; however, the application of MGB test identifies 46 PILFs, the fitting of LP3 distribution is remarkably better in Figure 3 (where MGB test is applied) than in Figure 2 (where GB test is applied). In another example, Figures 4, 5 and 6 show the effects of PILFs on fitting a probability distribution to the AM flood series for Station 405238 where the application of GB test identified only one PILF; however, the MGB test identified 21 PILFs. Figure 6 shows a better fit of the LP3 distribution to the AM flood data series (where MGB test is applied) than in Figure 5 (where GB test is applied).

Table 2 Number of PILFs identified by the MGB test and original GB test

Station ID	Number (%) of PILFs identified by MGB test	Number of PILFs identified by original GB test
218005	24 (51.06%)	None
219001	26 (41.94%)	None
219006	27 (45.00%)	None
116008	29 (50.00%)	None
125002	26(50.98%)	1
136202	46 (50.55%)	1
136301	31 (40.79%)	None
230204	17 (44.74%)	1
232213	17 (51.52%)	1
405238	21 (51.23%)	1

Table 3 Skew without removing any PILFs, PILFs removed by original GB test and PILFs removed by MGB test, respectively

Station ID	Skew (no PILFs removed)	Skew (PILFs removed by original GB test)	Skew (PILFs removed by MGB test)
218005	-0.375	-0.375	-0.553
219001	-0.525	-0.525	-0.098
219006	-0.514	-0.514	0.367
116008	-0.310	-0.310	-0.079
125002	-0.901	-0.757	0.095
136202	-1.434	-1.059	0.076
136301	-0.738	-0.738	0.477
230204	-0.671	-0.373	0.079
232213	-1.244	-1.197	-0.531
405238	-1.220	-1.151	0.285

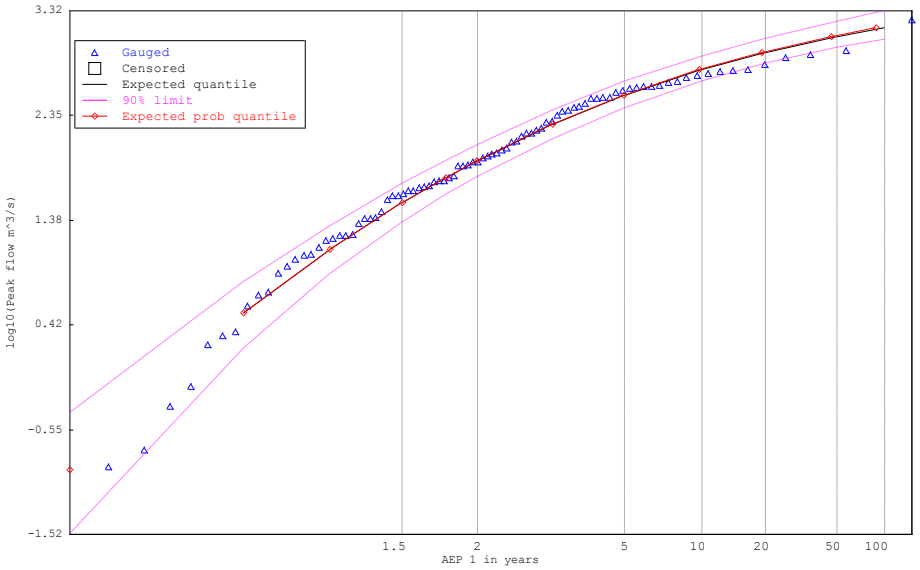

Figure 2 Flood frequency curve for Station 136202 using LP3 distribution (there was one PILF as per original GB test)

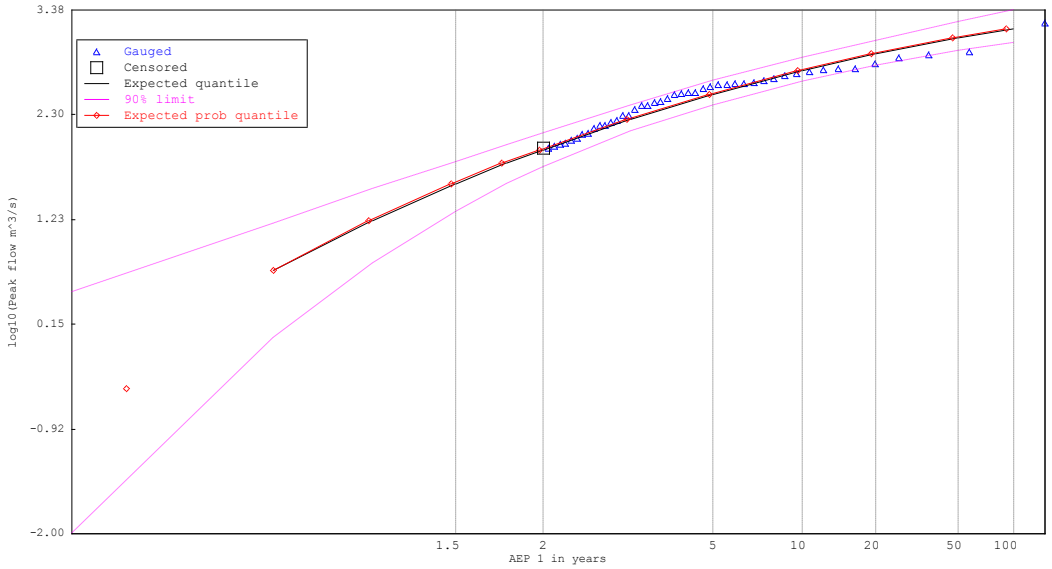

Figure 3 Flood frequency curve for Station 136202 using LP3 distribution (46 PILFs censored as per MGB test)

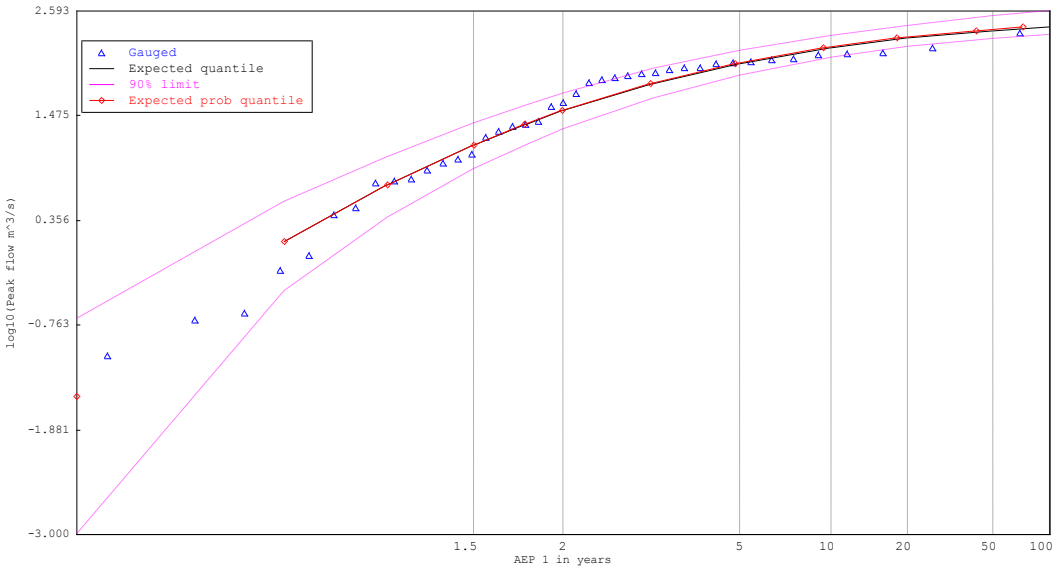

Figure 4 Flood frequency curve for Station 405238 using LP3 distribution (no PILF censored)

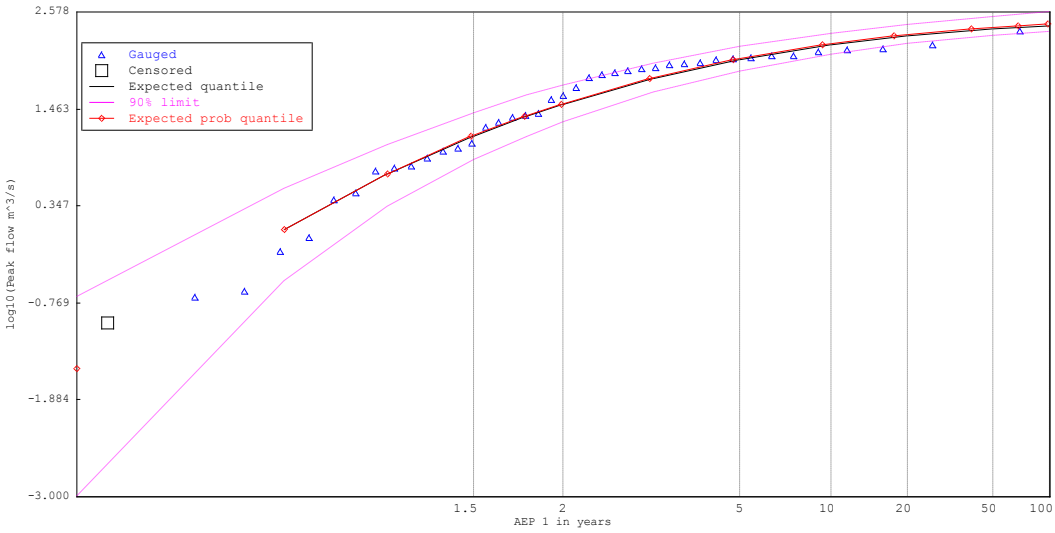

Figure 5 Flood frequency curve for Station 405238 using LP3 distribution (one PILF censored as per original GB test)

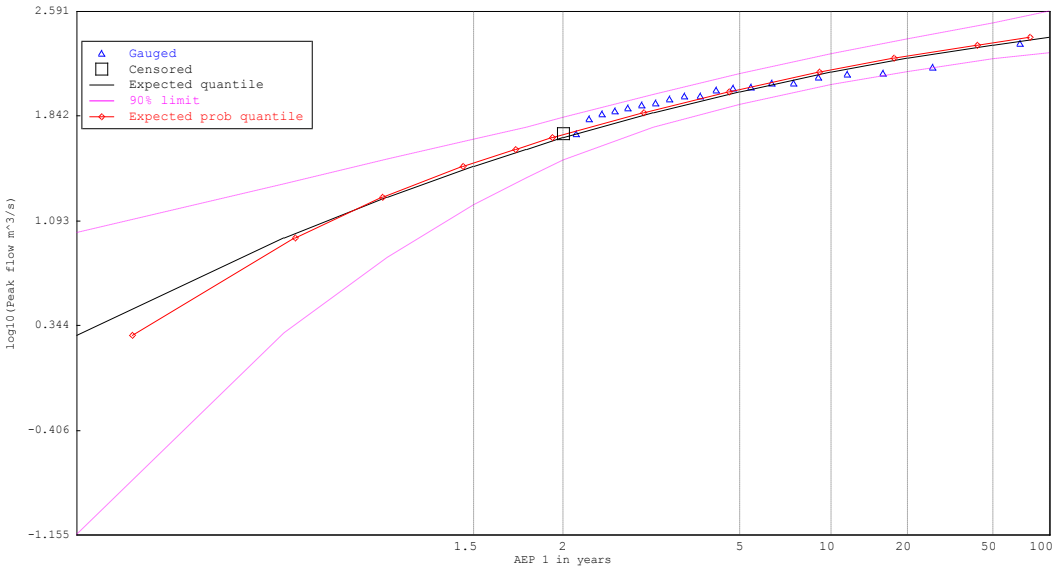

Figure 6 Flood frequency curve for Station 405238 using LP3 distribution (21 PILFs censored as per MGB test)

Figure 7 shows the fitting of the GEV distribution to Station 219001 without removing any PILFs and Figure 8 shows the fitting of LP3 distribution after removing 26 PILFs. From these two plots it is evident that LP3 distribution (Figure 8) shows a better fit to the AM flood series than the GEV distribution.

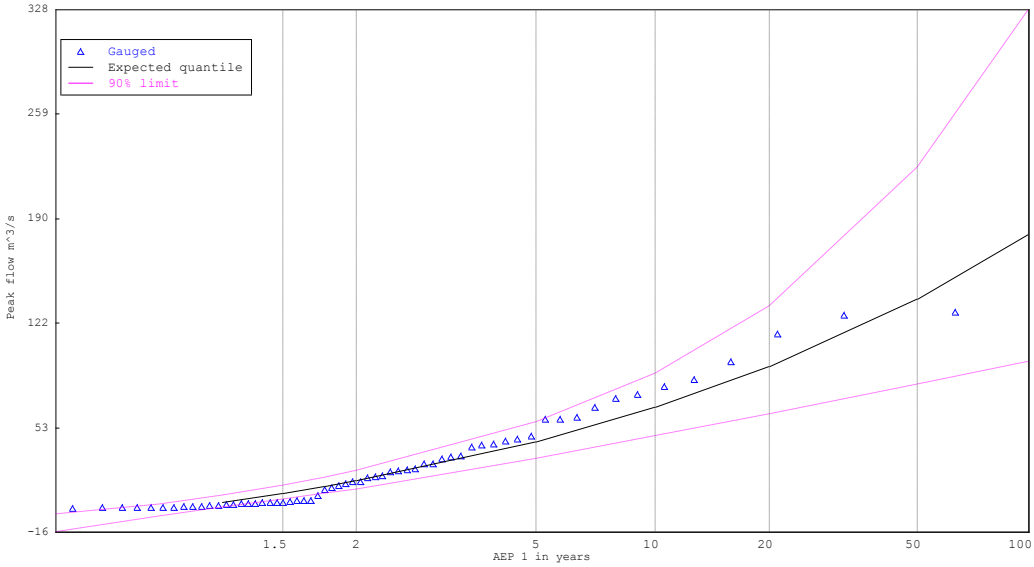

Figure 7 Fitting of the GEV distribution to the AM flood data for Station 219001 (No PILF censored)

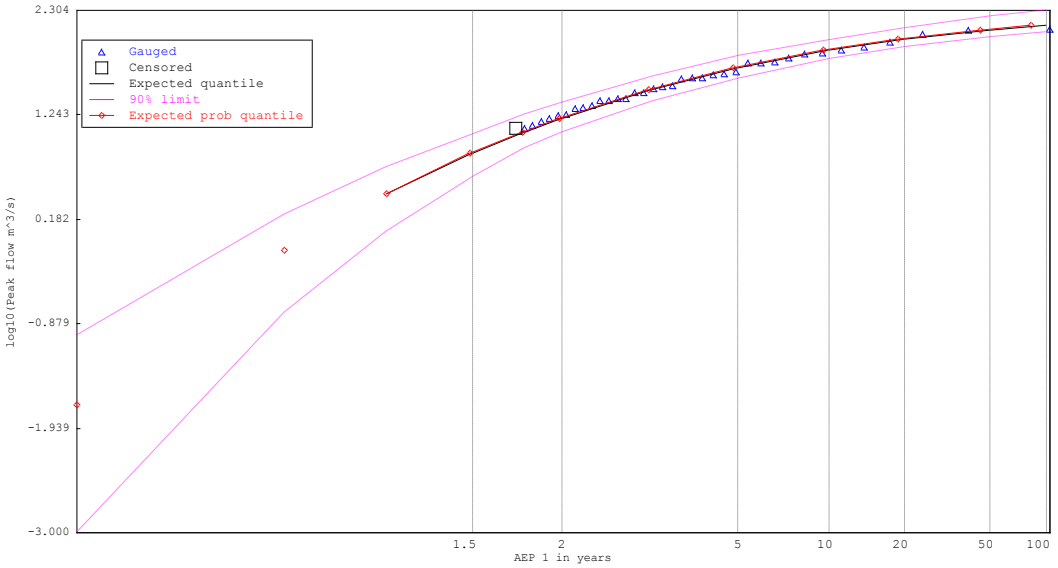

Figure 8 Fitting of the LP3 distribution to the AM flood data for Station 219001 (26 PILFs censored by MGB test)

Table 4 shows the flood quantile estimates using LP3 distribution where the PILFs are identified and censored by the original GB test and MGB test for AEPs of 1 in 10, 1 in 20, 1 in 50 and 1 in 100. It is found that there are notable differences between the two methods where flood quantiles show a variation in the range of -61% to 28%. Table 5 shows the variation between the flood quantiles estimated by two methods: LP3 with MGB test and GEV with L moments. It is found that for Station 218005 GEV distribution underestimates 1 in 10 AEP flood quantile by 6.68%, but for AEPs of 1 in 20, 1 in 50 and 1 in 100, GEV overestimates the flood quantiles by 4.4%, 24.1% and 38%, respectively. For other 9 stations the variations between the GEV and LP3 estimated quantiles are mixed i.e. a combination of over- and under-estimation by 0.24% to 26.7%. These results highlight the expected differences in flood quantile estimates between the LP3 and GEV distributions in eastern Australia.

Table 4 Estimated flood quantiles and percentage difference between two sets of quantiles: LP3 with MGB test and LP3 with original GB test

	Estimated quantiles (m^3/s) using LP3 distribution (PILFs removed by MGB test)				Estimated quantiles (m^3/s) using LP3 distribution (PILFs removed by original GB test (% difference between LP3 with MGB test and LP3 with original GB test)			
Station ID	AEPs (1 in *Y*)							
	10	20	50	100	10	20	50	100
218005	1422.74	1780.04	2063.82	2183.63	1072.7 (24.60%)	1704.52 (4.24%)	2690.8 (-30.38%)	3519.78 (-61.19%)
219001	79.68	103.3	128.01	142.16	77.55 (2.67%)	104.48 (-1.14%)	135.83 (-6.11%)	155.83 (-9.62%)
219006	184.37	279.38	427.86	555.27	133.51 (27.59%)	228.33 (18.27%)	402.83 (5.85%)	576.06 (-3.74%)
116008	920.88	1136.21	1342.61	1451.81	886.91 (3.69%)	1141.43 (-0.46%)	1454.31 (-8.32%)	1670.81 (-15.08%)
125002	3512.61	4198.86	4887.37	5276.62	2879.16 (18.03%)	3808.77 (9.29%)	4866.33 (0.43%)	5528.19 (-4.77%)
136202	565.08	836.87	1221.68	1518.39	586.6 (-3.81%)	854.9 (-2.15%)	1207.98 (1.12%)	1459.38 (3.89%)
136301	267.64	371.74	520.36	638.94	272.64 (-1.87%)	370.26 (0.40%)	498.4 (4.22%)	592.06 (7.34%)
230204	46.02	60.56	76.41	85.86	34.92 (24.12%)	53.71 (11.31%)	81.32 (-6.43%)	103.21 (-20.21%)
232213	26.7	32.38	38.36	41.9	22.3 (16.48%)	28.38 (12.35%)	34.89 (9.05%)	38.75 (7.52%)
405238	143.15	180.04	224.71	255.35	159.86 (-11.67%)	201.32 (11.82%)	240.6 (-7.07%)	260.91 (-2.18%)

Table 5 Comparison of flood quantiles by LP3 with MGB test and GEV with L moments

	Flood quantiles by GEV-L moments (m^3/s)				Flood quantiles by LP3 with MGB test (m^3/s) (% difference between LP3 with MGB test and GEV with L moments)			
Station ID	AEPs (1 in *Y*)							
	10	20	50	100	10	20	50	100
218005	1333.6	1861.3	2719.2	3522.4	1422.74 (-6.68%)	1780.04 (4.37%)	2063.8 (24.10%)	2183.6 (38.01%)
219001	66.3	93.1	137.3	179.5	79.68 (-20.18%)	103.3 (-10.96%)	128.01 (6.77%)	142.16 (20.80%)
219006	151	220.6	344.1	469.8	184.3 (-22.10%)	279.3 (-26.65%)	427.8 (-24.34%)	555.27 (-18.19%)
116008	819.1	1074.6	1457.8	1789.5	920.88 (-12.43%)	1136.2 (-5.73%)	1342.6 (7.90%)	1451.8 (18.87%)
125002	3263.1	4188.7	5488.6	6544.1	3512.61 (-7.65%)	4198.8 (-0.24%)	4887.3 (10.95%)	5276.6 (19.37%)
136202	453.3	662.5	1033.8	1411.9	565.08 (-24.66%)	836.8 (-26.32%)	1221.6 (-18.1%)	1518.39 (-7.54%)
136301	235.1	325.5	474.9	617	267.64 (-13.84%)	371.7 (-14.21%)	520.36 (-9.57%)	638.94 (-3.56%)
230204	37.5	52.4	76.9	100	46.0 (-22.72%)	60.56 (-15.57%)	76.41 (0.64%)	85.86 (14.14%)
232213	24.5	30.3	38	43.9	26.7 (-8.98%)	32.38 (-6.86%)	38.36 (-0.95%)	41.9 (4.56%)
405238	134.9	171.4	221.4	261	143.15 (-6.12%)	180.04 (-5.04%)	224.71 (-1.50%)	255.35 (2.16%)

5. Conclusion

This paper presents a case study using ten catchments from eastern Australia which evaluates two outlier tests being the original Grubbs and Beck (GB) test and multiple Grubbs and Beck (MGB) test. Two most commonly adopted probability distributions i.e. the GEV and LP3 have been adopted in the flood frequency analysis. For five stations, the original GB test did not detect any potentially influential low flows (PILFs); however, for these stations MGB test detected 40% to 50% of the annual maximum flood peaks as PILFs. For the remaining five stations, the original GB test identified one PILF from each station and the MGB test identified 45% to 50% as PILFs. Between these two methods, the differences in flood quantile estimates have been found to be up to 61% for the ten study catchments. It has also been found that GEV distribution (with L moments) and LP3 distribution with the multiple Grubbs and Beck test provide similar results in most of the cases; however, a difference up to 38% has been found for flood quantiles for AEP of 1 in 100 for one catchment.

6. Acknowledgements

Authors would like to acknowledge Australian Rainfall and Runoff revision Project 5 team for providing the data and FLIKE software for this study and Engineers Australia and Geosciences Australia for providing financial support for this project and Professor George Kuczera, Mr Erwin Weinmann, Associate Professor James Ball, Mr Mark Babister and Dr William Weeks for their support and input to this study.

References

Barnett V, Lewis T (1994). Outliers in Statistical Data, John Wiley, New York.

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

Bayley PB (1995). Understanding large river-floodplain ecosystems, *Bioscience*, 45, 153-162.

Beard LR (1962). Statistical methods in hydrology, Civil works investigation project CW-151, U.S. Army Corps of Engineers, Sacramento, California.

Beckman RJ, Cook RD (1983). Outlier ... ...s", *Technometrics*, 25(2), 119-149.

Benson MA (1968). Uniform flood-frequency estimating methods for federal agencies, *Water Resources Research*, 4(5), 891–908.

Bobée B, Cavidas G, Ashkar F, Bernier J, Rasmussen P (1993). Towards a systematic approach to comparing distributions used in flood frequency analysis, *Journal of Hydrology*, 142, 121–136.

Cohn TA, England JF, Berenbroc CE, Mason RR, Stedinger JR, Lamontagne JR (2013). A generalized Grubbs-Beck test statistic for detecting multiple potentially influential outliers in flood series, *Water Resources Research*, 49, 5047–5058.

Conway KM (1970). Flood frequency analysis of some NSW coastal rivers, Thesis (M. Eng. Sc.), University of New South Wales, Australia.

Cunnane C (1985). Factors affecting choice of distribution for flood series, *Hydrological Sciences Journal*, 30, 25–36.

Cunnane C (1989). Statistical distributions for flood frequency analysis, in Proc. Operational hydrological Report No. 5/33, World Meteorological Organization (WMO), Geneva, Switzerland.

Dixon WJ (1950). Analysis of extreme values, *The Annals of Mathematical Statistics*, 21(1), 488–506.

Dixon WJ (1951). Ratios involving extreme values, *The Annals of Mathematical Statistics*, 22(1), 68–78.

Gentleman J, Wilk M (1975). Detecting PILFs. II. Supplementing the direct analysis of residuals, *Biometrics*, 31(2), 387–410.

Gotvald AJ, Barth NA, Veilleux AG, Parrett C (2012). Methods for determining magnitude and frequency of floods in California, based on data through water year 2006, U.S. Geological Survey, Reston, Virginia.

Griffis V, Stedinger JR (2007). The LP3 distribution and its application in flood frequency analysis, 2. Parameter estimation methods, *Journal of Hydrologic Engineering*, 12(5), 492–500.

Grubbs FE (1969). Procedures for Detecting Outlying Observations in Samples, *Technometrics*, 11(1), 1-21.

Grubbs FE, Beck G (1972). Extension of sample sizes and percentage points for significance tests of outlying observations, *Technometrics*, 4(14), 847–853.

Haddad K, Rahman A (2008). Investigation on at-site flood frequency analysis in south-east Australia, IEM Journal, *The Journal of The Institution of Engineers, Malaysia*, 69(3), 59-64.

Haddad K, Rahman A (2010). Selection of the best fit flood frequency distribution and parameter estimation procedure: a case study for Tasmania in Australia, *Stochastic Environmental Research and Risk Assessment*, 25, 415–428.

Haddad K, Rahman A, Kuczera G (2011). Comparison of ordinary and generalised least squares regression models in regional flood frequency analysis: a case study for New South Wales, *Australian Journal of Water Resources*, 15(2), 59–70.

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

Haddad K, Rahman A (2012). Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework—quantile regression versus parameter regression technique, *Journal of Hydrology*, 430–431, 142–161.

Haddad K, Rahman A, Zaman M, Shrestha S (2013). Applicability of Monte Carlo cross validation technique for model development and validation in hydrologic regression analysis using ordinary and generalised least squares regression, *Journal of Hydrology*, 482, 119–128.

Hadi A, Simonoff J (1993). Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, 88(424), 1264– 1272.

Interagency Advisory Committee on Water Data (IAWCD) (1982). Guidelines for Determining Flood Flow Frequency: Bulletin 17-B, Hydrology Subcommittee, Washington DC.

Institution of Engineers Australia (I. E. Aust.) (1987). Australian Rainfall and Runoff: A Guide to Flood Estimation, Canberra.

Institution of Engineers Australia (I. E. Aust.) (2001). Australian Rainfall and Runoff: A Guide to Flood Estimation, Canberra.

Interagency Advisory Committee on Water Data (IAWCD) (2013). Robust National Flood Frequency Guidelines: What is an Outlier? Bulletin 17C, IAWCD, USA.

Ishak EH, Rahman A, Westra S, Sharma A, Kuczera G (2010). Preliminary analysis of trends in Australian flood data, in Proc.World Environmental and Water Resources Congress, American Society of Civil Engineers (ASCE), Providence, Rhode Island, USA, 120-124.

Ishak EH, Haddad K, Zaman M, Rahman A (2011). Scaling property of regional floods in New South Wales Australia, *Natural Hazards*, 58, 1155–1167.

Kirby W (1972). Computer-oriented Wilson–Hilferty transformation that preserves the first three moments and the lower bound of the Pearson type 3 distribution, *Water Resources Research*, 8(5), 125l–l254.

Kopittke RA, Stewart BJ, Tickle KS (1976). Frequency analysis of flood data in Queensland, in Proc. Hydrological Symposium, Institution of Engineers Australia, National Conference, Publication No. 76/2, 2-24.

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

Lamontagne JR, Stedinger JR, Cohn TA, Barth NA (2013). Robust national flood frequency guidelines: What is an outlier? In Proc. World Environmental and Water Resources Congress, ASCE.

Laio F (2004). Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters, *Water Resources Research*, 40:W09308. doi:10.1029/2004WR003204.

Laio F, Di Baldassarre G, Montanari A (2009). Model selection techniques for the frequency analysis of hydrological extremes, *Water Resources Research*, 45:W07416. doi:10.1029/2007/WR006666.

Marasinghe M (1985). A multistage procedure for detecting several PILFs in linear regression, *Technometrics*, 27(4), 395–399.

McCuen RH, Ayyub BM (1996). Probability, statistics, and reliability for engineers and scientists, A Chapman & Hall book, USA.

McMahon TA, Srikanthan R (1981). Log Pearson III distribution is it applicable to flood frequency analysis of Australian streams? *Journal of Hydrology*, 52. 1-2, 139-147.

Merz R, Bloschil G, Humer G (2008). National flood discharge mapping in Austria, *Natural Hazards*, 46, 53–72.

Meshgi A, Khalili D (2009a). Comprehensive evaluation of regional flood frequency analysis by L-abd LHmoments. 1. A re-visit to regional homogeneity, *Stochastic Environmental Research and Risk Assessment*, 23, 119–135.

Meshgi A, Khalili D (2009b). Comprehensive evaluation of regional flood frequency analysis by L-and LHmoments. II. Devewlopment of LH-moments parameters for the generalized Pareto and generalizedlogistic distributions, *Stochastic Environmental Research and Risk Assessment*, 23, 137–152.

Nathan RJ, Weinmann PE (1991). Application of at-site and regional flood frequency analyses, in Proc. International Hydrology Water Resources Symposium, Perth, 769-774.

Onoz B, Bayazit M (1995). Best-fit distribution of largest available flood samples, *Journal of Hydrology*, 167(1-4), 195-208.

Poff NL, Allan JD, Bain MB, Karr JR, Prestegaard KL, Richter BD, Sparks RE and Stromberg JC (1997). The natural flow regime: a paradigm for river conservation an restoration, *Bioscience*, 47, 769-784.

Prescott P (1975). An approximate test for PILFs in linear models, *Technometrics*, 17(1), 129–132.

Prescott P (1978). Examination of the behaviour of tests for outliers when more than one outlier is present, *Applied Statistics*, 27, 10–25.

Rahman A, Weinmann PE, Mein RG (1999). At-site flood frequency analysis: LP3-product moment, GEV-L moment and GEV-LH moment procedures compared, in Proc. Hydrology and Water Resource Symposium, Brisbane, 715–720.

Rahman AS, 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, *Natural Hazards*, 69, 1803-1813.

Rosner B (1975). On the detection of many outliers, *Technometrics*, 17(2), 221–227.

Rosner B (1983). Percentage points for a generalized ESD many outlier procedure, *Technometrics*, 25(2), 165–172.

Rousseeuw P, Zomeren BV (1990). Unmasking multivariate PILFs and leverage points, *Journal of American Statistical Association*, 85(411), 633–639.

Rousseeuw P, Leroy A (2003). Robust Regression and Outlier Detection, John Wiley, Hoboken, N. J.

Saf B (2010). Assessment of the effects of discordant sites on regional flood frequency analysis, *Journal of Hydrology*, 380 (3–4), 362–375.

Spencer C, McCuen R (1996). Detection of PILFs in Pearson type III data, *Journal of Hydrologic Engineering*, 1, 2–10.

Stedinger JR, Vogel RM, Foufoula-Georgiou E (1993). Frequency Analysis of Extreme Events, Chapter 18, Handbook of Hydrology, McGraw Hill, New York.

Thomas WO (1985). A uniform technique for flood frequency analysis, *Journal of Water Resources Planning and Management*, 111(3), 321–337.

Thompson W (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation, *The Annals of Mathematical Statistics*, 6, 214–219.

Tietjen G, Moore R (1972). Some Grubbs-type statistics for the detection of several outliers, *Technometrics*, 14(3), 583–597.

Verma S, Quiroz-Ruiz A (2006). Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering, *Revista Mexicana de Ciencias Geológicas*, 23(2), 133–161.

Vogel RM, McMahon TA, Chiew FHS (1993). Flood flow frequency model selection in Australia, *Journal of Hydrology*, 146, 421-449.