18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Partial duration series in the annual domain

Magdy Mohssen

Lincoln University, PO Box 84, Lincoln, New Zealand, mohssenm@lincoln.ac.nz

Abstract: Flood frequency analysis has been an important tool to estimate design floods which are essential for flood protection and management. Series of maximum annual flows MAF have been the preferred choice to collect the needed sample of high flow events for parameter estimation of the identified statistical models. However, MAF series can be short as it is restricted to the available years of record, and ignores significant high flow events, as it selects the highest for each year. Partial duration series of high flows (PDF) consist of high flows above a pre-defined threshold, and it will include all significant high events above this threshold. Moreover, PDF series can be much longer than the MAF series, as more than one event can be selected per year. Two major concerns are usually connected to the PDF series: the first is the value for the threshold which will affect the size of the series, and the second is the transformation of return periods corresponding to design floods from the PDF domain to the annual domain. A new formula has been derived to transform the return periods from the PDF domain to the annual domain, and results of this new formula have been compared to the commonly used one in the literature. The new formula, which is based on the binomial process, only requires the independence of the flood events, and produces more realistic and reliable values compared to the one available in the literature, which assumes a Poisson distribution for floods above the threshold.

Hourly flow series for the Leith (46 years) and the Tokomairiro (21 years, will be referred to as Toko) Rivers in Otago, NZ have been collected from the Otago Regional Council in Dunedin. Independent flood events above several thresholds for both rivers have been identified, along with the maximum annual flows. The cumulative distribution function (CDF) of the PDF series was compared to the CDF of the MAF series for both rivers and has shown the improvement due to smoother and more consistent CDF, which will impact on the ability of any statistical model to simulate these series.

Several competitive statistical models, including the Gumbel, GEV and the GP models have been fitted and tested to the MAF series and PDF series (with different sizes) of the Leith and Toko Rivers. The GP model outperformed other models, and thus it was the choice for this study. Model application and testing have shown that fitting the GP model to the PDF series outperformed the MAF series, and resulted in much more reliable model for the simulation of the flood events of these rivers. Testing criteria was based on comparison of the historical and modelled (produced by the model) CDF and histograms, in addition to Kolmogorov Smirnov goodness of fit test, Chi^2 goodness of fit test and Filliben correlation coefficient. In general, increasing the number of flood events in the PDF series improved the goodness of fit for the fitted GP model for the Toko River, but was limited to an arrival rate (λ) of 4 for the Leith high flows. Thus, a thorough testing for several arrival rates, corresponding to different thresholds, should be carried out for the choice of the best threshold for the PDF series. Study of the impact of the size of the PDF series (bigger λ) on design floods indicated that, in general, their values will increase with increasing λ. However, the optimum value for λ does not have to be the highest, as was the case for the Leith River.

Keywords: Flood Frequency, Partial Duration Series, Design Floods, Generalized Pareto, Extreme Flows

1 INTRODUCTION

Data variability in natural phenomena is a major challenge to simulate and model. Floods or high flows from a stream or a river are the product of climatic conditions imposed upon a complex watershed system. The variability in the spatial and temporal distribution of a rainfall storm over a watershed can be immense, and is a function of meteorological and topographical characteristics, which are usually hard to simulate precisely. Add to this the wide spread variability of a watershed characteristics from land use distribution, vertical and spatial distribution of soil type, antecedent conditions before a rainfall event, and elevations and slopes. Each rainfall event could be different from other events, and has its own unique output over the watershed. However, there are common "statistical" characteristics of the main processes which produce these events. Statistical models have been widely applied in the literature to simulate flood events of a specified catchment. Historical available record of flood events is used to estimate these statistical characteristics, and a candidate from commonly used models in the literature is elected, based on its performance compared to others, to simulate the distribution of the flood events and re-produce these statistical characteristics.

The choice of maximum annual flows MAF for a record of "n" years has been widely used in the literature to obtain a sample of the high flows for the purpose of flood frequency analysis and modelling. The MAF approach is easy to apply and, in general, guarantees that the chosen high flow events are independent. Moreover, MAF series are directly related to the commonly used concept of return periods for design floods. For instance, the highest observed flow in n years has an exceedance probability of 1/(n+1) (Weibull plotting position) with (n+1) return period, while the second highest in the MAF series will have an exceedance probability of 2/(n+1) with (n+1)/2 years return period, and so on. However, the MAF series does not really represent the process of high flows, as it will ignore high flows which are not the highest in a year, while they are higher than the other chosen highest flows of other years. Furthermore, MAF series could be short based on the available observed record, as it picks only one value for each year, while there could be so many significantly high flows observed during this record.

The partial duration series of flood events PDF has been investigated and recommended in the literature to overcome the shortcomings of the MAF series. In the PDF series, independent high flows above a threshold are selected. This usually produces much longer series than the MAF, as more than one event per year could be chosen. The PDF series does not have to include high flow events from each year as the MAF does, as flows above a threshold are selected regardless of the year in which they occurred. The PDF approach for sampling represents the population process of high flows better than the MAF, but still the MAF sampling approach is the popular one among engineers for estimating design floods. This is mainly due to the need to check for the independence of the PDF series, and the question mark in relation to the choice of the threshold or the number of events to be chosen. In addition, the exceedance probability of the PDF, and in turn the corresponding return period, will not conform to the commonly recognised ones derived from the MAF series. For example, consider a total of 2n high flow events have been selected from a record of n years. The return period of the highest flow, which would have the same value in the MAF, is (2n+1), and not (n+1) as in the MAF series. The impact of the choice of the threshold value on the basic assumption of the model, and in turn on the size of the PDF series, has been investigated by Begueria (2005), and a new approach has been proposed for modelling PDf series based on increasing the threshold value. This approach was based on the assumption that the population process follows the Poisson-GP model, which does not have to apply for each case of flood frequency study. It is usually more efficient for one to freely identify several competitive models, apply and test them, and choose the one which fits the best.

Regional flood frequency analysis has been applied to improve the quality of the estimated flood quantiles at a site based on the transfer of information from other nearby catchments (Stedinger et al 1993, McKerchar and Pearson 2001, Yue and Wang 2004), especially for sites with short record of data. Micevski and Kuczera (2009) developed a simple general procedure to combine at site and regional information such that uncertainty is minimised, and maximum utilization of available information is achieved.

This research will present and analyse a new relation to transform the return period of the PDF series into the annual domain, and will investigate the impact of the choice of the high flow threshold on the best fit of the selected model.

2 STATISTICAL MODELLING

The Generalized Extreme Value and Pareto (GEV and GP, respectively) probability distributions have been accepted and widely applied in the literature to simulate the statistical process of flood events (Connell and Pearson 2001, Todorovic 1978). The Gumbel distribution, which is a special case of the generalized Extreme Value model, has been utilised (Mckerchar and Pearson 1989) to produce regional flood frequency contours for New Zealand. In this research, several models have been applied during the identification and testing process, including the Gumbel, GEV, and the GP models. The GP model outperformed other distributions, and was the best fit for the applications of this study. Thus, only the results for the GP model will be presented in this paper.

2.1 Models' Formulation

The Generalized Pareto distribution GP was derived to fit values above a threshold, such as floods above a defined value (Hosking and Wallis 1987, Rosbjerg 1985, Stedinger et al. 1993). For the case when the shape parameter k< 0, the GP distribution has a lower limit at u, while for the case when k>0, it has an upper limit at u + α/k, in addition to the lower limit at u. The GP is described by the following formula:

$$F(Q) = 1 - \left[1 - \frac{k(Q-u)}{\alpha} \right]^{\frac{1}{k}} \qquad \text{for } k>0 \text{ and } u \le Q \le u + \alpha/k \tag{1}$$

Where Q is the high flow, while u, α and k are model parameters (location, scale, and shape parameters, respectively) to be estimated. The GP quantile, corresponding to a return period T is:

$$Q_T = u + \frac{\alpha}{k}\left\{1 - (\frac{1}{T})^k\right\} \tag{2}$$

The moments for the GP distribution are as follows: Mean is $\mu_Q = u + \frac{\alpha}{1+k}$, Variance is $\sigma_Q{}^2 = \frac{\alpha^2}{\left[(1+k)^2(1+2k)\right]}$, and the skewness (exists for k> -0.33) is $\gamma_Q = \frac{2(1-k)(1+2k)^{\frac{1}{2}}}{(1+3k)}$.

2.2 Parameters Estimation

There are a variety of methods to estimate the parameters of a statistical model. Among these tentative approaches are the method of moments MOM, maximum likelihood method ML, least squares LS, and the L moments (Benjamin and Cornell 1970, Yue and Wang 2004). The most efficient approach for parameter estimation for a specified model should be applied. The efficiency of an estimator $\hat{\theta}$ of a parameter θ is measured by its mean square error: $E\left[(\hat{\theta}-\theta)^2\right] = \sigma_{\hat{\theta}}{}^2 + \left(E\left[\hat{\theta}\right]-\theta\right)^2$, which equals the variance of the estimator added to the square of the bias.

The method of moments can produce efficient estimators for some models, such as the Normal distribution and the autoregressive time series models, but it fails to produce efficient parameter for many other models. The method of maximum likelihood is accepted to provide efficient estimators [Benjamin and Cornell 1970, Chow et al. 1988). However, in many cases, it is hard to derive simple forms for the ML estimators, and numerical techniques have to be applied to estimate them. The method of L-moments, which is based on the probability weighted moments, provides sufficiently efficient estimators and usually it is simple to derive their formulae (McKerchar and Pearson 1989, Stedinger et al. 1993, Van Montfort and Van Putten 2002). Estimators of the L-moments for a sample are linear combinations of the ranked observations. In the case of flood events, this procedure would be preferable as it is related to the probability of exceedance of the flood event. In addition, L-moments estimators are almost unbiased and are not significantly affected

by outliers [Rosbjerg 1985, Stedinger et al. 1993). The following presents the first and second L-moments for the GP distribution and for higher order L-moments refer to Stedinger et al. (1993):

$$\lambda_1 = u + \frac{\alpha}{1+k}, \quad \lambda_2 = \frac{\alpha}{(1+k)(2+k)}$$

L-moments from the sample are substituted in their corresponding formulas for the GP model to obtain a system of simultaneous equations which are solved to obtain the model's parameters.

3 RETURN PERIODS OF THE PDF IN THE ANNUAL DOMAIN

3.1 Commonly Used Formula

For the case of the MAF series, Eq. 2 will produce design floods corresponding to the commonly used return period T_A. However, this is not the case with the PDS, where these design floods will correspond to T_P, which is different from T_A. Consider a case where the size of record of observed flows is n years, while N high flow events (where $N \geq n$) have been selected. The ratio $\lambda = N/n$, where $\lambda \geq 1$, is called the arrival rate. The arrival rate λ is the average flood events per year in a PDF series. Stedinger et al. (1993, p 18.38) stated that the application of partial duration series for $\lambda > 1.65$, assuming Poisson arrival with the exponential model for exceedance probability, should produce more reliable estimates for the quantiles of the design floods. Stedinger et al. (1993), assuming a Poisson distribution for floods between the threshold value Q_0 and a flood value Q, where $Q \geq Q_0$, in one year period, presented the following relation between T_A and T_P:

$$T_P = -\frac{\lambda}{Ln(1-\frac{1}{T_A})} \qquad (3)$$

However, this equation is based on the assumption that the arrival rate follows a Poisson distribution, which might not be the case for many flood events.

3.2 Derivation of a New Formula

Assuming that flood events of the PDF series are independent, the probability of getting exactly one event every year, out of λ events, for the case when λ is integer, can be derived from the Binomial distribution as follows:

$$p_A = \binom{\lambda}{1}(p_P)(1-p_P)^{\lambda-1} \qquad (4)$$

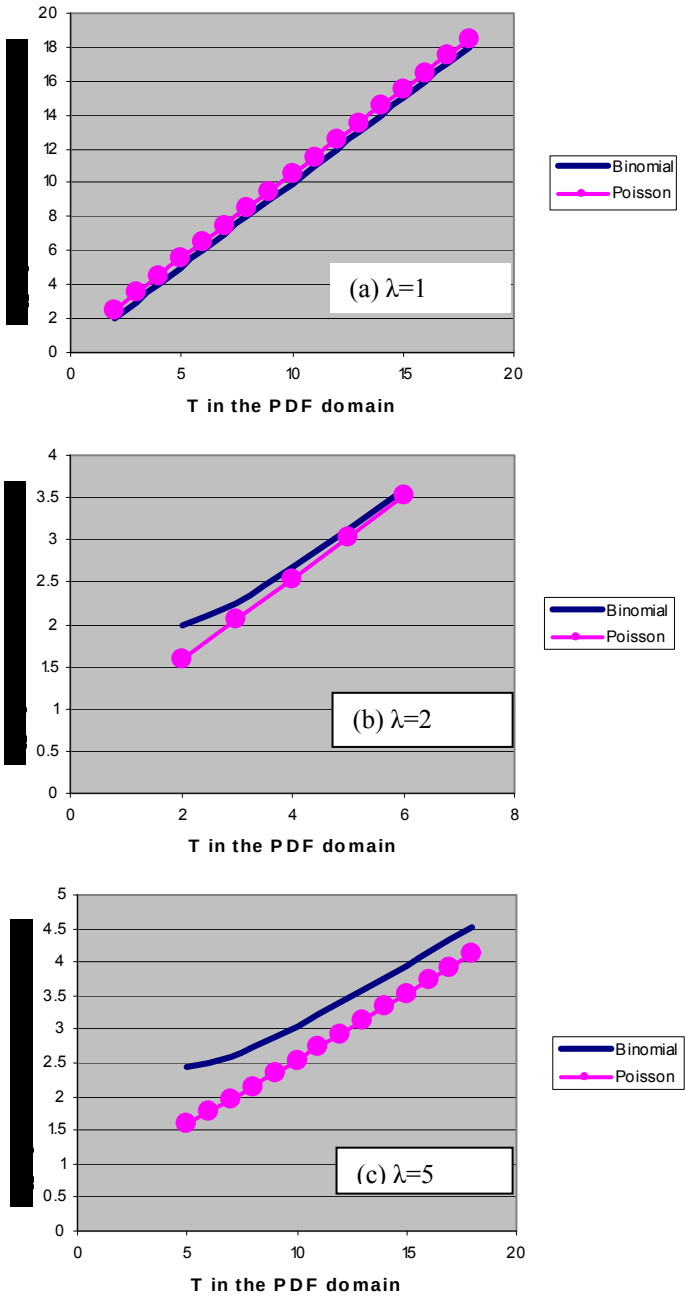


Figure 1. Return Periods of PDF series in the Annual Domain for: (a) λ=1, (b) λ=2, and (c) λ=5

Where p_A is the annual exceedance probability = $1/T_A$, and p_P is the exceedance probability for the PDF series $=1/T_P$. In general, the return period of design floods obtained from the PDF series, in years (similar to the one obtained from the MAF series), can be obtained from the PDF return periods as follows:

$$\frac{1}{T_A} = \lambda\left(\frac{1}{T_P}\right)\left(1-\frac{1}{T_P}\right)^{\lambda-1} \qquad (5)$$

Equation (5) does not require the average arrival rate of flood events to follow the Poisson distribution, and should be applicable to any value of λ. However, this equation applies only for values of $T_A > 1$ and, in turn, is only applicable to $T_p \geq \lambda$.

For λ=1, Eq. 5 produces Ta=Tp, which of course is expected as the number of high flow events is the same as the number of years, the same case for the MAF series. However, the use of the Eq. 3, which is commonly applied in the literature, yields return periods which are 0.5 years higher than the correct ones, as shown in Fig. (1.a). Figures (1.b) and (1.c) show that the new derived equation produces significantly higher return periods in the annual domain at the lower range, compared to the commonly used one in the literature of Eq. 3. Thus, the newly derived equation for transforming the return period from the PDF domain to the annual domain would produce lower design floods, compared to the commonly practised approach of Eq. (3). Moreover, Fig. 1 shows that the Binomial model approaches the Poisson model for higher return period, as expected (Benjamin & Cornell, 1970, p.240).

4 APPLICATIONS TO THE LEITH AND THE TOKOMAIRIRO HIGH FLOW EVENTS

The Leith River goes through Dunedin city in NZ, and has a catchment area of 45 km^2, and its available record of hourly flows extends from 1963 until 2008 (46 years). A record of 21 years during the period 1982 to 2002 of hourly observed flows for the Tokomairiro River at its West Branch site, in Milton south of Dunedin, NZ has also been considered in this research. The catchment area upstream of the gauge site on the West Branch of the Tokomairiro River is 69.55 km^2.

Figure 2 shows the cumulative probability distributions for observed floods of the Tokomairiro River and the Leith River by using both the maximum annual high flows, and the highest 115 independent high flow events for the tokomairiro River, and the highest 138 flood events for the Leith River. The figure shows the smoothness and consistency added to the historical sample due to the use of the PDF series, and not ignoring significant high flow events from the historical sample, which will be used for parameter estimation in the modelling process.

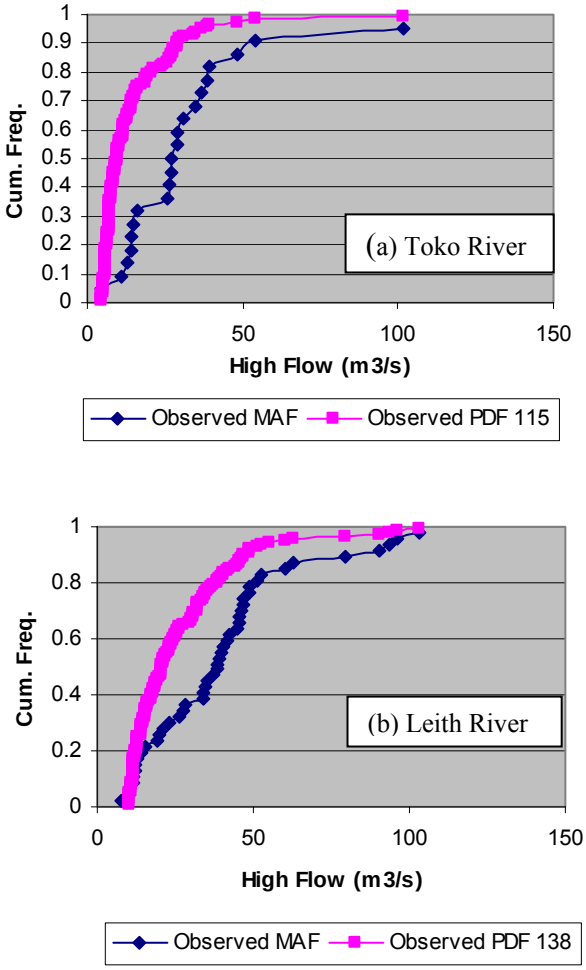


Figure 2. Cumulative Distributions for the observed MAF and PDF Series: (a) Toko River, (b) Leith River

Figure 3 shows the cumulative frequency for the historical and the fitted GP model, for the MAF series and the PDF series (λ=5 for the Toko River, while. λ=4 for the Leith River). The choice of the presented case

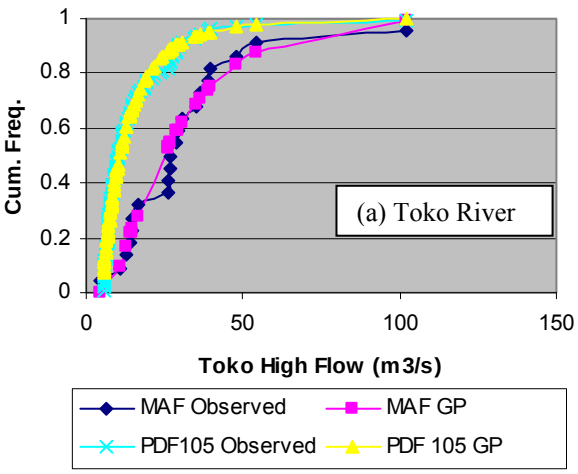
for λ was based on the results for different sample sizes for the PDF series, and the best case was presented, as shown in Table 1. Figure 3 confirms that the PDF series, with more flow events, and accounting for all high flow events above the specified threshold outperformed the MAF series in the best fit for the selected GP model. In addition, Table 1 shows that the goodness of fit for the GP model is improved with increasing the size of the PDF series, but to a limit. For the Leith flows, it was λ=4 which produced the best fit GP model, and not λ=5. In fact, the fitted GP model failed the chi^2 goodness of fit test for λ=5 in the Leith flows case.

This result sheds an important light on the limitation of the size (or the threshold flow) for the PDF series, and that it is not a matter that one selects as many events as he could. The choice of λ=4 was based on the Filliben correlation coefficient, however one can argue that the GP model performed better at λ=3 based on the value of the Chi^2 calculated statistic. Despite the fact that most of the fitted GP models passed the imposed 3 tests, one should choose the model with the best goodness of fit. The value of the design flood corresponding to a certain return period increased, in general, with the increase of the size of the PDF series, as shown in Fig. 4. This emphasizes the need to properly select the threshold for the PDF series based on a thorough analysis of the goodness of fit for the identified model, otherwise an overestimation of the design flood could result due to a lower chosen threshold in the PDF series.

5 CONCLUSIONS

This research study has highlighted the need to implement the use of partial duration series for flood frequency analysis, instead of the commonly used approach of selecting maximum annual flows. The newly developed formula for transforming the return period of the partial duration series to the annual domain was shown to be more reliable and efficient than the currently available one in the literature. The new formula will produce smaller design floods for lower return periods, but for higher return periods, both methods will essentially produce the same design floods.

Applications to high flow events for the Tokomairiro River and the Leith River in Otago, NZ have supported the applicability of the PDF series approach, which outperformed the corresponding MAF series. The choice of the threshold for the choice of high flows for the PDF series requires a thorough testing for several cases, and the one with the best fit for the identified model should be implemented. In general, design floods increase with the increase of the size of the PDF series (decreasing the threshold).



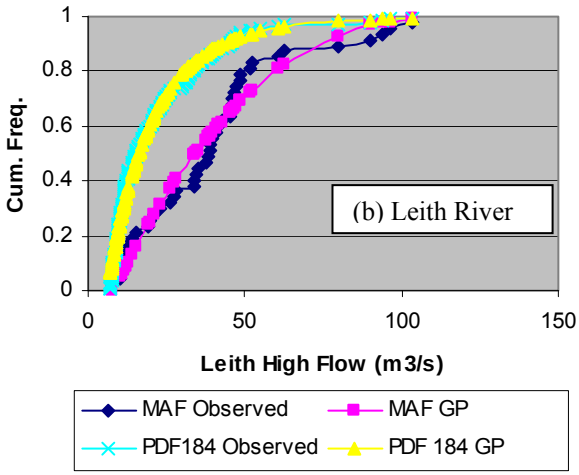


Figure 3. Cumulative Distributions for the Observed and Modelled MAF and PDF Series: (a) Toko River, (b) Leith River

Table 1. Model Testing for the GP model Fitted to the MAF and PDF Series

	Toko River High Flows			Leith River High Flows		
High Flow Series	Smirnov Kolmogorov	Chi^2	Filliben	Smirnov Kolmogorov	Chi^2	Filliben
MAF	0.144	2.710	0.941	0.118	9.217	0.975
PDF: λ=1	0.095	0.532	0.961	0.070	6.819	0.977
PDF: λ=2	0.081	1.646	0.945	0.053	6.773	0.985
PDF: λ=3	0.059	1.533	0.962	0.060	6.885	0.991
PDF: λ=4	0.069	2.184	0.972	0.060	8.676	0.993
PDF: λ=5	0.064	3.133	0.979	0.071	10.730	0.992

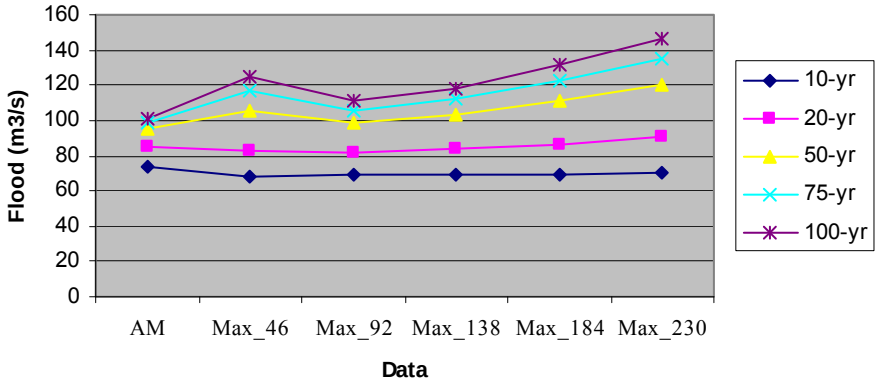


Figure 4. Leith Design Floods corresponding to different sizes of PDF series

6 REFERENCES

Begueria, S. (2005). Uncertainties in partial duration series modelling of extremes related to the choice of the threshold value, *J. of Hydrology, Amsterdam*, 303(1/4), 215-230.

Benjamin, J.R. and C.A. Cornell (1970). Probability, Statistics, and Decision for Civil Engineers, McGRAW-HILL: New York, 380-.403.

Chow, V.T., D.R. Maidment and L.W. Mays. (1988). Applied Hydrology, McGRAW-HILL: New York, 371-376.

Connell, R.J. and C.P. Pearson (2001). Two-component extreme value distribution applied to Canterbury annual maximum flood peaks, *Journal of Hydrology (NZ)*, 40(2), 105-127.

Freeman, H. (1963). Introduction to Statistical Inference, Addison-Wesley, Massachussets.

Hosking, J.R.M. and J.R. Wallis (1987). Parameter and Quantile Estimation for the Generalized Pareto Distribution, *Technometrics*, 29(3), 339-349.

McKerchar, A.I and C.P. Pearson (1989). Flood Frequency in New Zealand, Hydrologic Centre, Christchurch, Publication No.20, 26-30.

McKerchar, A.I and C.P. Pearson (2001). Comparison of a regional method for estimating design floods with two rainfall-based methods, *Journal of Hydrology (NZ)*, 40(2), 129-138.

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

Pearson, C.P. and T. Davies (1997). Stochastic methods (Chapter 5), Floods and Droughts: the New Zealand Experience, eds. P..Mosley & C.P. Pearson, New Zealand Hydrological Society, New Zealand, 65-87.

Rosbjerg, D. (1985). Estimation in Partial Duration Series with Independent and Dependent Peak Values, *J. of Hydrology*, 76(1), 183-196.

Stedinger, J.R., R.M. Vogel, R.M. and E. Foufoula-Georgiou (1993). Frequency Analysis of Extreme Events (Chapter 18), Handbook of Hydrology, ed. D.R. Maidment, McGRAW-HILL: New York, 18.1-18.58.

Todorovic, P. (1978). Stochastic Models of Floods, *Water Resources Research*, 14(2), 345-356.

Van Montfort, M.A.J. & B. van Putten (2002). A comment on modelling extremes: Links between Multi-Component Extreme Value and General Extreme Value distributions, *Journal of Hydrology (NZ)*, 41(2), 197-202.

Yue, S. and C. Wang (2004). Determination of regional probability distributions of Canadian flood flows using L-moments, *Journal of Hydrology (NZ)*, 43(1), 59-73.