

## An insight into flood frequency for design floods

M. Mohssen

*Lincoln University, PO Box 84, Lincoln, New Zealand*

### Abstract

The maximum annual flood (MAF) series of peak flows have been commonly used in the literature to estimate the parameters of a statistical model in order to estimate design floods for the proper protection and management of flows during extreme events. The use of maximum annual flows restricts the size of the available sample to the number of available years of observed flows, and ignores the fact that flood events which are not the highest in some years, can be higher than the highest flows of other years. The partial duration flood (PDF) series, which includes flood events above a specified threshold, has been addressed in the literature as a more efficient alternative. A new relation between the return periods of the PDF and the MAF series have been derived based on the assumption that flood events are independent, without the need to assume that their arrival follows a Poisson process. The General Extreme Value (GEV) distributions, including the Gumbel as a special case, in addition to the Generalized Pareto (GP) distribution have been fitted and applied to the flood events of the Tokomairiro River, New Zealand. The use of the PDF series, with 115 flood events in 21 years, resulted in smoother and more homogeneous series, and produced higher design floods than the MAF series. The new derived equation for the relation between PDF and annual return periods resulted in smaller design floods for return periods up to 10 years, while design floods for higher return periods were almost the same as the values obtained from the available relation in the literature.

*Keywords: flood frequency, partial duration series, design floods, generalized extreme value, Generalized Pareto, extreme flows.*



## 1 Introduction

Estimation of floods corresponding to specified return periods (or a specified risk) is essential for the design of flood protection schemes, assessment of regions at risk of flooding, and the proper management of flooded regions. Due to high climatic variability, which drives flood events, statistical modelling has become widely applied to estimate the magnitude of floods corresponding to a specified risk [3, 10, 11]. The two main approaches to select the flood series for fitting a statistical model to the observed floods are based on the series for maximum annual flows (MAF) or partial duration series of floods (PDF). The MAF series selects the maximum flood event for each year, while the PDF series consists of all flood peaks above a specified magnitude. The MAF approach ignores the fact that the highest flows in some years can be lower than flood events in other years, and hence it excludes significant high flood events from its parameter estimation process. Moreover, the sample size for the MAF series is restricted to the available number of years of observed flows. On the contrary, the PDF series will consider all significant flood events in its parameter estimation process and its size can extend far beyond the available years of recorded flows.

Begueria [1] investigated the impact of the value of the threshold in the PDF series on the basic assumption of the model, and proposed a new approach for modelling the PDF series based on increasing the threshold value. However, this procedure was based on the assumption that the real data follows the Poisson-GP model [1,10], which could not be the case for many flood events. It is usually preferred that one should identify several models, apply and test them to choose the best, and not to be restricted to a specified model.

## 2 Statistical modelling

In this research, the focus will be on the Extreme Value and the Generalized Pareto distributions, which have been widely applied in the literature for modelling extreme flood events [4, 7, 8, 10, 12]. Mckerchar and Pearson [7] used the Gumbel distribution, which is a special case of the generalized Extreme Value, to produce regional flood frequency contours for New Zealand.

### 2.1 Models' formulation

The Extreme Value Distribution EV was derived to fit the maximum of a set of independent and identically distributed (iid) series of random variables [2, 3, 10]. The General Extreme Value Distribution GEV can be presented as:

$$F(Q) = e^{-\left[1 - \frac{k(Q-u)}{\alpha}\right]^{\frac{1}{k}}}, \quad \infty < Q < \infty \quad (1)$$

where  $Q$  is the flood flow,  $u$ ,  $\alpha$  and  $k$  are model parameters to be estimated, known as location, scale, and shape parameters, respectively. The scale



parameter  $\alpha$  is assumed to be positive. There are three limiting cases for the General Extreme Distribution, depending on the value of the shape parameter  $k$ :

(a) for  $k=0$ , the GEV reduces to the Gumbel distribution with only two parameters, the location and scale parameters. The formula for the Gumbel distribution looks different from the GEV equations and is presented as:

$$F(Q) = e^{-e^{-\frac{Q-u}{\alpha}}}, \quad -\infty < Q < \infty \quad (2)$$

The Gumbel quantile  $Q_T$  corresponding to a return period  $T$  can be estimated from (2) as:

$$Q_T = u - \alpha \ln \left[ -\ln \left( 1 - \frac{1}{T} \right) \right] \quad (3)$$

The moments of the Gumbel distribution are:

$$\text{Mean } \mu_Q = u + 0.5772 \alpha, \quad \text{Variance } \sigma_Q^2 = \frac{\pi^2 \alpha^2}{6}, \quad \text{Skewness } \gamma_Q = 1.14$$

(b) For  $k < 0$  the distribution is known as the Extreme Value Type II distribution EVII, which has a lower limit at  $u + \alpha/k$ .

(c) For  $k > 0$  the distribution is known as the Extreme Value Type III distribution EVIII, which has an upper limit at  $u + \alpha/k$ .

The quantile for the GEV (type II or III) is:

$$Q_T = u + \frac{\alpha}{k} \left\{ 1 - \left[ -\ln \left( 1 - \frac{1}{T} \right) \right]^k \right\} \quad (4)$$

The moments for the GEV types II and III distributions are as follows:

Mean:  $\mu_Q = u + \left( \frac{\alpha}{k} \right) [1 - \Gamma(1+k)]$ , and the Variance (exists for  $k > 0.5$ ):

$$\sigma_Q^2 = \left( \frac{\alpha}{k} \right)^2 \left\{ \Gamma(1+2k) - [\Gamma(1+k)]^2 \right\}$$

The Generalized Pareto distribution (GP) was derived to fit values above a threshold, such as floods above a defined value [5, 10]. 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  $u + \alpha/k$ , in addition to the lower limit at  $u$ . The GP is described by the following formula:

$$F(Q) = 1 - \left[ 1 - \frac{k(Q-u)}{\alpha} \right]^{\frac{1}{k}} \quad \text{for } k > 0 \text{ and } u \leq Q \leq u + \alpha/k \quad (5)$$

The GP quantile, corresponding to a return period T is:

$$Q_T = u + \frac{\alpha}{k} \left\{ 1 - \left( \frac{1}{T} \right)^k \right\} \quad (6)$$

The moments for the GP distribution are as follows:

$$\text{Mean } \mu_Q = u + \frac{\alpha}{1+k}$$

$$\text{Variance } \sigma_Q^2 = \frac{\alpha^2}{[(1+k)^2(1+2k)]}$$

$$\text{Skewness (exists for } k > -0.33) \gamma_Q = \frac{2(1-k)(1+2k)^{\frac{1}{2}}}{(1+3k)}$$

## 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 [2, 10]. The most efficient approach for parameter estimation for a specified model should be applied. The efficiency of an estimator  $\hat{\theta}$  of a parameter  $\theta$  is measured by its mean square error:

$$E[(\hat{\theta} - \theta)^2] = \sigma_{\hat{\theta}}^2 + (E[\hat{\theta}] - \theta)^2,$$

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 [2, 10]. 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 [7,10,13]. 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 exceedence of the flood event. In addition, L-moments estimators are almost unbiased and are not significantly affected by outliers, [10, 13].

The following presents the first and second L-moments for the Gumbel, GEV and GP distributions, and for higher order L-moments refer to Stedinger et al. [11]:

$$\text{Gumbel: } \lambda_1 = u + 0.5772\alpha, \quad \lambda_2 = \alpha \ln 2$$

$$\text{GEV: } \lambda_1 = u + \frac{\alpha}{k} [1 - \Gamma(1+k)], \quad \lambda_2 = \frac{\alpha}{k} (1 - 2^{-2k}) \Gamma(1+k)$$

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

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

### 2.3 Relation between the return period of the MAF and the PDF

The return period  $T$  corresponding to a flood obtained by using the quantile equations 3, 4 and 6, with the MAF series, will correspond to an annual return period (in years). However, using the PDF series, with flood events ( $N_F$ ) more than the available years of record ( $N_Y$ ), will yield a different return period from the commonly known annual one. Assume that the commonly known annual return period is  $T_A$ , while the one corresponding to the PDF series is  $T_P$ . The ratio between flood events used in the PDF series and the available years of record, which is the same as the flood events used in the MAF series, is called

the arrival rate and can be estimated from  $\beta = \frac{N_F}{N_Y}$ , where  $\beta \geq 1$ . The arrival

rate  $\beta$  is the average flood events per year in a PDF series. For  $\beta > 1.65$ , the application of partial duration series, assuming Poisson arrival with the exponential model for exceedance probability, should produce more reliable estimates for the quantiles of design floods corresponding to specified return periods, [9–11]. Stedinger et al. [11], 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  $T_P$ :

$$T_P = -\frac{1}{\ln(1 - \frac{1}{T_A})} \quad (7)$$

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.

Another formula, which only requires the independence of flood events will be derived and presented herein. Assuming that flood events of the PDF series are independent, the probability of getting exactly one event every year, out of  $\beta$  events, for the case when  $\beta$  is integer, can be derived from the Binomial distribution as follows:

$$p_A = \binom{\beta}{1} (p_P)(1 - p_P)^{\beta-1} \quad (8)$$

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:

$$\frac{1}{T_A} = \beta \left( \frac{1}{T_P} \right) \left( 1 - \frac{1}{T_P} \right)^{\beta-1} \quad (9)$$

Equation (9) does not require the average arrival rate of flood events to follow the Poisson distribution, and should be applicable to any value of  $\beta$ .

### 3 Applications to the Tokomairiro flood events

The Tokomairiro River at its West Branch site, in Milton south of Dunedin, New Zealand has been selected for this study. The record period for this site extends during the period 1962 to 2002. Gauging of flows at this site was carried out during low and high flows, which would incur higher reliability in its rating curve, and in turn in the estimated flows from the river stage. The catchment area upstream of the gauge site on the West Branch is 69.55 km<sup>2</sup>, while the whole catchment of the Tokomairiro River, including the East Branch, up to its mouth at the Pacific Ocean is 393.9 km<sup>2</sup>.

Figure 1 shows the cumulative probability distributions for observed floods of the Tokomairiro River by using both the maximum annual 21 floods and the highest 115 independent flood events in the available 21 years of record. The figure shows the inconsistency and sudden changes in the maximum annual series, which is expected to have its impact on the quality and accuracy of the fitted statistical model. To the contrary, the plot of the highest 115 flood events is smoother and shows consistency which is expected in the population process from which these flood events are produced. It should be noted that the same flood event will have different cumulative probability in the PDF from the MAF due to its calculation based on 115 events instead of 21 in the latter approach.

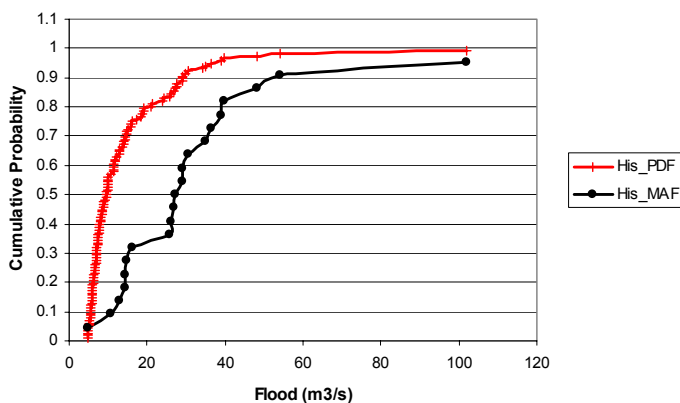


Figure 1: Cumulative distributions for the observed MAF and PDF series.

Figure 2 and Table 1 present the flood histogram and the results for model testing for the 3 models fitted to the MAF series. Despite that the Kolmogorov Smirnov and the  $\chi^2$  tests accept the three models, the histogram and the Philliben correlation coefficient strongly recommend the GEV distribution. Moreover, the calculated statistic for the  $\chi^2$  test for the GEV distribution is significantly lower than its corresponding values for the Gumbel and GP distributions.

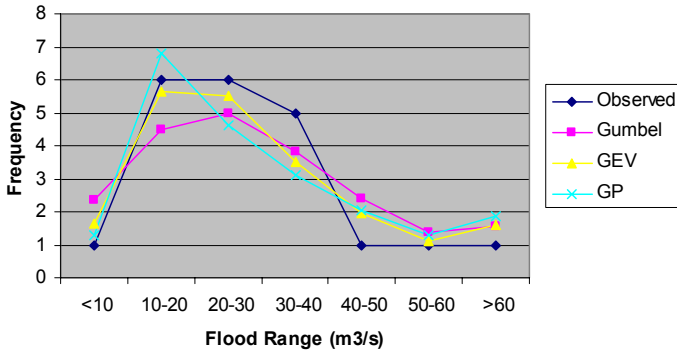


Figure 2: Floods histograms for models fitted to the MAF series.

Table 1: Model testing for models fitted to the MAF series.

Test	Model	Gumbel	GEV	GM
Kolmogorov Smirnov	Calculated Value	0.12	0.13	0.14
	Tabulated Value	0.284	0.284	0.284
$\chi^2$	Model	Accepted	Accepted	Accepted
	Calculated Value	2.97	1.67	2.71
	Tabulated Value	9.5	7.82	7.82
Filliben Correlation Coefficient	Model	Accepted	Accepted	Accepted
		0.93	0.95	0.94

Figure 3 and Table 2 present the same above mentioned results, but for the PDF series containing 115 independent flood events observed during the record period of 21 years. The histogram shown in Fig. 3 recommends the GP model, which is further supported by the testing results of Table 2. It is noted that the GEV was the favourable distribution given the MAF series, while the GP has become the preferred model based on the PDS of 115 flood events. Moreover, the Gumbel distribution has been rejected in the PDS modelling process, while it

was accepted, still not preferred, when the MAF series was considered in the modelling process.

Equation 9, to estimate the design return period in years for design floods of PDF series, yielded smaller values for the design floods corresponding to return periods of up to 10 years compared to the use of Eq. 8. However, this difference was small, and has become insignificant when higher return periods were considered. For instance, the 5 year flood event from a GP model was 39.23 m<sup>3</sup>/s by using Eq. 9, while it was 40.66 m<sup>3</sup>/s if Eq. 8 was used instead.

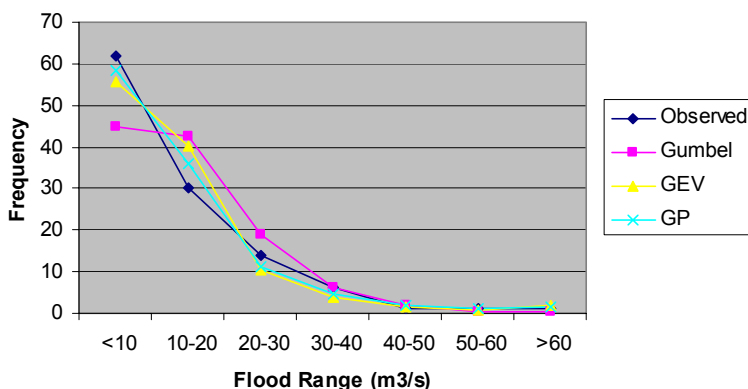


Figure 3: Flood histograms for models fitted to the PDS series of 115 events.

Table 2: Model testing for models fitted to the PDS series of 115 events.

Test	Model	Gumbel	GEV	GM
Kolmogorov Smirnov	Calculated Value	0.172108	8.60E-02	5.39E-02
	Tabulated Value	0.126821	0.126821	0.126821
	Model	Rejected	Accepted	Accepted
Chi <sup>2</sup>	Calculated Value	14.83837	6.361572	3.307905
	Tabulated Value	9.5	7.82	7.82
	Model	Rejected	Accepted	Accepted
Filliben Correlation Coefficient		0.898605	0.984218	0.980359

Figure 4 presents design flood events for the 50 year return period corresponding to the MAF series, and 6 alternatives for selecting the PDF series, by considering the 21, 42, 63, 84, 105 and 115 highest flood events. The use of PDF with 115 highest flood events to fit a GP model yielded a quite higher 50 year flood event (98 m<sup>3</sup>/s) compared with the use of the MAF series with only 21 values to fit the same model (85.4 m<sup>3</sup>/s), as shown in Fig. 4. In general, the



magnitudes of design flood events were higher with increasing the size of the flood events sample in the PDF series.

## 4 Conclusions

This research stresses the need to implement the use of partial duration series for the estimation of design floods. Partial duration series produced more homogeneous and smoother series, which is expected to result in a more robust model fitting. The choice of PDS with 115 flood events, much more than the available 21 years of observed record, not only resulted in a higher design floods, but also resulted in the choice of a different model to fit the flood series. In addition, the derived equation to transform return periods of PDS to their corresponding annual return periods resulted in smaller values for the design floods compared to the relation available in the literature. However, more research is still needed to identify the lower threshold above which independent flood events are selected.

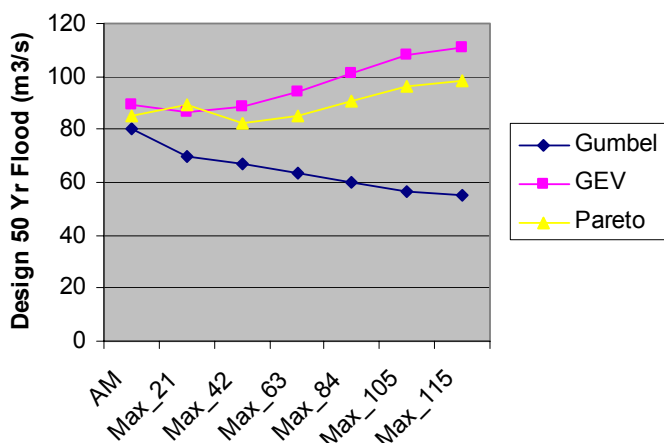


Figure 4: 50 year design floods for models fitted to the MAF and PDS series.

## References

- [1] Begueria, S., Uncertainties in partial duration series modelling of extremes related to the choice of the threshold value, *J. of Hydrology*, Amsterdam, 303(1/4), pp 215–230, 2005.
- [2] Benjamin, J.R. & Cornell, C.A., *Probability, Statistics, and Decision for Civil Engineers*, McGraw-HILL: New York, pp. 380–403, 1970.
- [3] Chow, V.T., Maidment, D.R. & Mays, L.W., *Applied Hydrology*, McGraw-Hill: New York, pp. 371.376, 1988.



- [4] Connell, R.J. & Pearson, C.P., Two-component extreme value distribution applied to Canterbury annual maximum flood peaks, *Journal of Hydrology (NZ)*, 40(2), pp 105–127, 2001.
- [5] Freeman, H., *Introduction to Statistical Inference*, Addison-Wesley, Massachusetts, 1963.
- [6] Hosking, J.R.M. & Wallis, J.R., Parameter and Quantile Estimation for the Generalized Pareto Distribution, *Technometrics*, 29(3), pp. 339–349, 1987.
- [7] McKerchar, A.I & Pearson, C.P., Flood Frequency in New Zealand, Hydrologic Centre, Christchurch, Publication No.20, pp 26–30, 1989.
- [8] McKerchar, A.I & Pearson, C.P., Comparison of a regional method for estimating design floods with two rainfall-based methods, *Journal of Hydrology (NZ)*, 40(2), pp 129–138, 2001.
- [9] Pearson, C.P. & Davies, T., Stochastic methods (Chapter 5), *Floods and Droughts: the New Zealand Experience*, eds. P.M. Mosley & C.P. Pearson, New Zealand Hydrological Society, New Zealand, pp. 65–87, 1997.
- [10] Rosbjerg, D., Estimation in Partial Duration Series with Independent and Dependent Peak Values, *J. of Hydrology*, 76(1), pp. 183–196, 1985.
- [11] Stedinger, J.R., Vogel, R.M. & Foufoula-Georgiou, E., Frequency Analysis of Extreme Events (Chapter 18), *Handbook of Hydrology*, ed. D.R. Maidment, McGraw-Hill: New York, pp. 18.1–18.58, 1993.
- [12] Todorovic, P., Stochastic Models of Floods, *Water Resources Research*, 14(2), pp. 345–356, 1978.
- [13] Van Montfort, M.A.J. & van Putten, B., A comment on modelling extremes: Links between Multi-Component Extreme Value and General Extreme Value distributions, *Journal of Hydrology (NZ)*, 41(2), pp 197–202, 2002.
- [14] Yue, S. & Wang, C., Determination of regional probability distributions of Canadian flood flows using L-moments, *Journal of Hydrology (NZ)*, 43(1), pp 59–73, 2004.

