Hydrological Processes
Hydrol. Process. **12**, 1685–1699 (1998)

Annual maxima and partial duration flood series analysis by parametric and non-parametric methods

Kaz Adamowski*, Geng-Chen Liang and Gilles G. Patry
Department of Civil Engineering, University of Ottawa, Ontario, Canada

Abstract:

Annual maxima (AM) and partial duration (PD) flood series are modelled by parametric and non-parametric methods. In PD analysis the number of threshold exceedances is assumed to be Poisson distributed; the peak exceedances are described by the generalized Pareto (GP) and non-parametric (NP) distributions. The generalized extreme value (GEV) and non-parametric (NP) distributions are used to describe the AM series. L-moments are employed for parameter estimation for GEV and GP distributions. Analysis of data from the provinces of Quebec and Ontario, Canada, shows that both AM and PD series can be inferred as being unimodal and bimodal, both of which can be described by the NP method. Also, this method is found not to be sensitive to the choice of threshold level; however, it was also observed that parametric methods cannot detect biomodality, give different quantile estimates for AM and PD data and PD estimates are sensitive to the selection of threshold level. Therefore, the NP method is more advantageous than the parametric methods in flood frequency analysis for both AM and PD series. © 1998 John Wiley & Sons, Ltd.

KEY WORDS floods; annual maxima; partial duration; non-parametric and parametric methods

INTRODUCTION

In flood frequency analysis it is important to extract the maximum information possible from the sample. Two basic types of extremes can be extracted from stream flow records, namely the annual maximum (AM) series and peak over threshold (POT), or partial duration (PD) series. The AM series is more commonly used; however, the PD method can be particularly useful when the period of record is short. The AM and PD approaches differ in their distribution selection and estimation of parameters and quantiles. The choice between the two series has been investigated by many researchers. Cunnane (1973), using the standard error of the quantile estimate, found that the PD series has a larger sampling variance than the AM series for a return period greater than 10 years, when the number of exceedances per year is equal to one. More recent comparison performed by Madsen *et al.* (1997) shows that the PD model is generally the preferred method for flood analysis since it is more suitable for heavy-tailed distributions, which are common in hydrological applications. However, PD models cannot be used effectively when data are bimodal as a result of more than one flood-causing event (i.e. snowmelt, rainfall) occurring in a season or within a year. Floods can have a bimodal distribution (Waylen and Woo, 1987) which requires a mixture of two-component parametric distributions. Another source of possible errors in modelling AM and PD series are the methods used in parameter estimation. However, the recently developed L-moments method (Hosking, 1990) gives almost

* Correspondence to: Professor K. Adamowski, Department of Civil Engineering, University of Ottawa, 161 rue Louis Pasteur, CP450 succ A, Ottawa, ON K1N N5, Canada

Contract grant sponsor: Natural Sciences and Engineering Research Council of Canada.
Contract grant number: G15430.

CCC 0885–6087/98/111685–15$17·50

Received 20 May 1997
Revised 8 February 1998
Accepted 27 March 1998

unbiased estimates of parameters. The purpose of this paper is to develop a new non-parametric method to describe the PD flood series, and to compare its performances against PD and AM models.

MODELLING PD AND AM SERIES

The main issues in the modelling of PD series are: the choice of the threshold, the selection of the model of the arrival rate of events larger than the threshold level and the distribution selection and parameter estimation for modelling the magnitude of floods. The threshold selection is crucial, but by no means straightforward. Based on a simulation study, Yevjevich and Taesombut (1979) found that the value of λ should be greater than 1·8. Rosbjerg (1985) concluded that when the data are independent (uncorrelated) then both AM and PD series give the same SE regardless of λ value. Considering a typical sample size (20–40 years) available for the analysis, Ashkar *et al.* (1987) found that λ could be less than 1·65. This last conclusion was further supported by Rasmussen *et al.* (1993). Birikundavyi and Rousselle (1997) analysed PD flood data from Quebec and Ontario and found the most appropriate threshold level to be equal to 1·7. However, Madsen *et al.* (1997), using a Monte Carlo simulation study comparing PD and AM series, concluded that the threshold could be between 0·4 and 15. Clearly, the issue of threshold selection requires more research.

Given that N exceedances of the threshold have been observed during M years, the occurrence of exceedances is assumed to be described by the Poisson process, with a periodicity of one year. The mean number of exceedances per year is estimated from

$$\hat{\lambda} = N/M \tag{1}$$

Hence, if n_q denotes the number of exceedances of flow q in a year, then

$$P\{n_q = k) = \lambda^k \, \mathrm{e}^{-\lambda}/k! \qquad k = 0,\ 1,\ 2, \ldots \tag{2}$$

The exceedances are assumed to be generalized Pareto (GP) distributed (Pickands, 1975).

A theoretical expression for T-year events can be derived by noting that the number of exceedances above the threshold level r in a particular year, denoted by n_r, is a Poisson-distributed random variable with a mean value given by (Rasmussen, 1994).

$$E\{n_r\} = \lambda[1 - F(r)] \tag{3}$$

Therefore, the distribution of the annual maxima flood, X is given by

$$F_a(x) = P(n_x = 0) = \exp[-E(n_x)] = \exp\{-\lambda[1 - F(x)]\} \tag{4}$$

The cumulative distribution function (CDF) of the GP distribution is given by

$$\begin{aligned} F(x) &= 1 - \left[1 - k\frac{x - x_0}{\alpha}\right]^{1/k} \qquad k \neq 0 \\ &= 1 - \exp\left[-\frac{x - x_0}{\alpha}\right] \qquad k = 0 \end{aligned} \tag{5}$$

where α is the scale, x_0 is the location and k is the shape parameter.

Insertion of the CFD of the GP distribution [Equation (5)] into Equation (4) gives a GEV distribution for AM series greater than x_0 and $k \neq 0$,

$$F_a(x) = \exp\left[-\lambda\left(1 - k\frac{x - x_0}{\alpha}\right)^{1/k}\right] = \exp\left[-\left(1 - k\frac{x - \zeta}{\beta}\right)^{1/k}\right] \qquad x \geqslant x_0 \tag{6}$$

Equation (6) reduces to a Gumbel (or EV1) distribution when $k = 0$, while for $x \geqslant x_0$, the transformed parameters ζ and β are given by

$$\beta = \alpha\lambda^{-k}$$
$$\zeta = x_0 + \frac{\alpha(1 - \lambda^{-k})}{k} \qquad \text{when } k \neq 0 \tag{7}$$

and

$$\zeta = x_0 + \alpha \ 1\text{n}(\lambda) \qquad \text{when } k = 0$$

The above is a general Poisson–Pareto parametric model that is currently used in hydrological applications with k values estimated regionally and interchangeably for PD and AM series.

The CDF of GP and GEV are similar, the main difference being that the GP is bounded below at x_0 and the GEV at $e^{-\lambda}$.

Wang (1991) compared the GEV/AM model with the GP/PD model in flood analysis, and concluded that the two models perform equally well when the average number of exceedances in a year is about one ($\lambda = 1$). This is different from the value for λ suggested by Cunnane (1973) ($\lambda > 1{\cdot}65$) and Yevjevich and Taesombut (1979) ($\lambda > 1{\cdot}8$). Madsen *et al.* (1997) compared GP/PD and GEV/AM models for maximum likelihood (ML), method of moments (MOM) and probability weighted moments (PWM) and concluded that in most cases the PD/ML provides the most efficient estimator. Once the distribution is selected, then its parameters are estimated. The L-moments method gives almost unbiased estimates of parameters (Hosking, 1990). Therefore, it is adopted and used in this paper for parametric methods. However, all previous studies were based on the assumption that the distributions are known. In practical applications the actual distributions for AM and PD series are not known. Many other distributions have also gained wide recognition (e.g. log-Pearson type III), but the distribution choice is often inconclusive and even controversial.

Recognizing the problems faced when using parametric methods, an alternative non-parametric method has been introduced (Adamowsi, 1985, 1989). Its use for modelling PD series is investigated, and is compared with L-moments parametric methods in this paper.

L-moments

L-moments are linear combinations of ranked observations that do not require squaring or cubing of the observations, as do product-moment estimations. As a result they are almost unbiased estimators, and as such are very attractive in practical applications. The L-moments of X can be defined as functions of probability weighted moments (PWM). For order statistics of ranked observations $X_{(j)}$ an estimator of PWM for $i \geqslant 1$, is (Hosking, 1986).

$$b_i = n^{-1} \sum_{j=1}^{n} x_{(j)} \frac{(j-1)(j-2) \ \dots \ (j-i)}{(n-1)(n-2) \ \dots \ (n-i)} \tag{8}$$

For any distribution, the first four L-moments can be calculated from

$$\lambda_1 = b_0$$
$$\lambda_2 = 2b_1 - b_0$$
$$\lambda_3 = 6b_2 - 6b_1 + b_0 \tag{9}$$
$$\lambda_4 = 20b_3 - 30b_2 + 12b_1 - b_0$$

The L-moments can be converted to dimensionless L-moment coefficient of variation (τ_2), skewness (τ_3) and kurtosis (τ_4), as follows

$$\tau_2 = \lambda_2/\lambda_1;\ \ \tau_3 = \lambda_3/\lambda_2;\ \ \tau_4 = \lambda_4/\lambda_2 \tag{10}$$

The parameters of the GEV distribution in terms of L-moments are given by (Hosking, 1991)

$$\begin{aligned} k &\approx 7{\cdot}817740z + 2{\cdot}930462z^2 + 13{\cdot}641492z^3 + 17{\cdot}206675z^4 \\ \beta &= \lambda_2 k/[(1-2^{-k}]\Gamma(1+k)] \\ \zeta &= \lambda_1 + a[\Gamma(1+k) - 1]/k \end{aligned} \tag{11}$$

where

$$z = 2/(\tau_3 + 3) - 1\text{n}(2)/1\text{n}(3)$$

The quantile of the GEV distribution is obtained from

$$x_p = \begin{cases} \zeta + \beta\{1 - [-\ln(p)]^k\}/k, & \text{for } k \neq 0, \\ \zeta - \beta\ 1\text{n}[1 - 1\text{n}(p)], & k = 0 \end{cases} \tag{12}$$

where p is the assumed cumulative probability.

The parameters of the GP distribution in terms of L-moments are given by (Hosking, 1991)

$$k = \frac{(1 - 3\tau_3)}{(1 + \tau_3)},\quad \alpha = (1+k)(2+k)\lambda_2,\quad \mu = \lambda_1 - (2+k)\lambda_2 \tag{13}$$

where μ is the location (threshold level x_0), k is the shape and α is the scale parameter. When μ is known, parameters k and α can be estimated by

$$k = (\lambda_1 - \mu)/\lambda_2 - 2,\quad \alpha = (1+k)(\lambda_1 - \mu) \tag{14}$$

Simulation by Hosking and Wallis (1987) shows that L-moment estimation is most useful when $k \leqslant -0{\cdot}2$, because L-moment parameter and quantile estimators are then less biased than moment or maximum likelihood estimators. For $k > -0{\cdot}1$, conventional moment estimators are asymptotically more efficient than L-moment estimators.

The quantiles estimation by the GP can be computed from

$$x_p = \begin{cases} \mu + \alpha\{1 - (1-p)^k\}/k, & \text{for } k \neq 0, \\ \mu - \alpha\ \ln(1-p), & k = 0 \end{cases} \tag{15}$$

NON-PARAMETRIC ESTIMATION OF DISTRIBUTION

Current methods of flood frequency analysis are based on the assumption that the sample of flow observations come from a population with known probability density function (PDF). An *a priori* choice of a PDF (e.g. GEV) is made, and its parameters are estimated using one of several methods (e.g. moments, L-moments, maximum likelihood). Such methods are said to be 'parametric'. However, in the hydrological context, the PDF is never known, and must be assumed. Despite intensive research and legislation, no particular method has emerged as the best and most uniform across different sites.

In recent years several researchers, for example Adamowski (1989), Lall *et al.* (1993) and others, have strongly advocated the use of non-parametric density estimators. There are many theoretical and practical reasons in favour of the non-parametric method (Lall *et al.*, 1993).

The process of non-parametric density estimation is similar to that of building a histogram when a rectangular block of height $1/nh$ is added at the centre of the class interval to which a given data point belongs. The final histogram frequency distribution is the sum total of all blocks, each one of area $1/n$. In non-parametric frequency, a kernel function of area $1/n$ centred at the data point location itself is added. This kernel may be of a variety of shapes. The final non-parametric density is the sum of all kernels.

The non-parametric kernel density estimate $f_n(x)$ from a sample $\{x_1, x_2, \ldots, x_n\}$ of size n is given by (Adamowski, 1985)

$$f_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right) \tag{16}$$

where $K()$ is a kernel function, itself a probability density function, and h is a bandwidth or smoothing factor, which is to be estimated from the data. In hydrological applications, the two most commonly used estimators are constant kernel estimator (CKE), whose bandwidth is constant throughout the data points, and variable kernel estimator (VKE), whose bandwidth depends on the interpoint distance between x_j and its kth nearest neighbour amongst all the data points. It has been shown (Adamowski, 1989) that there is no apparent advantage of the VKE over the CKE. Therefore, in this study, CKE is used.

The choice of the kernel has a relatively minor effect on the resulting density function. In this study, the optimal Epanechnikov kernel (Adamowski and Feluch, 1990) is used and is given by

$$K(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{for } -1 \leqslant t \leqslant 1 \\ 0 & \text{elsewhere} \end{cases} \tag{17}$$

The calculation of h is crucial to the performance of the estimator. Several means of using the data to yield an objective choice of h have been proposed. The optimal values of h in terms of integrated mean square error (IMSE) is obtained from the least-square cross-validation (LSCV) method. Using an IMSE criterion, Labatiuk and Adamowski (1987) found that the various numerical algorithms for computing h perform similarly and are all close to the optimal value predicted by theory, which is expressed by

$$h = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \frac{(x_i - x_j)}{5^{1/2} n(n - 10/3)} \tag{18}$$

Using an optimal kernel [Equation (17)] and an expression for the optimal value of h [Equation (18)], it is possible to estimate the density function by Equation (16).

In hydrological applications, it often happens that some elements in the sample are far apart (i.e. outliers), which might result in a discontinuity of the density function giving an increased bias at such boundaries. In order to remedy this problem an unbiased boundary kernel is used in this paper. Let $q = (x - x_{\min})/h$, then for the left boundary (i.e. $0 \leqslant q \leqslant 1$), the boundary kernel is given by (Muller, 1991)

$$K(q, z) = 6(1 + z)(q - z)\frac{1}{(1 + q)^3}\left[1 + 5\left(\frac{1 - q}{1 + q}\right)^2 + 10\frac{1 - q}{(1 + q)^2}z\right] \tag{19}$$

where $z = (x - x_i)/h$. Similarly, for the right boundary, assume q $(x_{\max} - x)/h$.

NUMERICAL ANALYSIS

The data

Numerical analysis was performed using data from eight gauging sites (Table I) from the Ontario and Quebec provinces of Canada. The two provinces cover a large territory encompassing various climatic

Table I. Hydrometric stations studied and mean number of exceedances/year (m) at selected threshold level (x_0) for the PD model

Site name and no.	Drainage area (km^2)	Years of record	Threshold level x_0 (m^3/s)	Number of exceedances/ year (m) at x_0	Province
Nottawasaga near Baxter (02ED003)	1180	42	26·3	4	Ontario
Saugeen near Port Elgin (02FC001)	3960	76	11·3	4	Ontario
Ausable near Springbank (02FF002)	865	45	17·7	7	Ontario
Pinewood near Pinewood (05PC011)	461	39	6·99	3·5	Ontario
Hall pres d'east Hereford (02OE018)	218	42	23·3	4	Quebec
Beaurivage a Sainte-Etienne (02PJ007)	709	62	44·3	5	Quebec
Mistassini en amon de la Mistassibi (02RD003)	9713	27	294	3	Quebec
Chamouchouane a la Chute a Michel (02RF001)	1 5333	27	467	2	Quebec

regimes. As a consequence, flooding is caused by different mechanisms in different parts of these two provinces (Gingras *et al*., 1994). One extreme is in southern Ontario, which experiences hot summers and short winters. Owing to a number of thaws in winter, as well as the occurrence of numerous thunderstorms during warm weather months, the annual maximum flood can take place during any month of the year. The other extreme is in northern Quebec, which experiences short summers and long, cold winters, which lead to large snowpack accumulations. There, the prolonged melt period leads to a large spring flood peak which is almost always the annual maximum flood. At the northernmost tip of Quebec, the spring flood peak will even occur in early July in some years.

Other parts of Ontario and Quebec experience snowmelt and rainfall floods in a proportion that is essentially dependent on their latitude and proximity to the Atlantic Ocean. This is because a south to north increased gradient in snowpack exists. The different timing of floods in various parts of the two provinces has led to varied shapes of the probability density function of the annual maximum flood series. Consequently, Ontario and Quebec were divided into nine homogeneous regions of roughly similar climate and in which floods were generated by similar mechanisms (Gingras *et al*., 1994). While all parts of Ontario and Quebec will experience both snowmelt and rainfall — floods in a partial duration (PD) series, the month of occurrence of these floods, as well as the number of snowmelt floods, will vary between the various parts of the provinces.

Threshold level for PD series

The PD model requires the abstraction of N independent events from a record length of M years ($N > M$), all of which exceed a threshold of discharge of magnitude x_0. A variety of approaches are available for threshold specification; perhaps the most widely used method is based upon the assumption that the number of values exceeding the threshold each year is considered to be a random variable with a Poisson distribution with parameter λ (Cunnane, 1979)

$$P(\text{m peaks} > x_0 \text{ in a year}) = P_m = \mathrm{e}^{-\lambda}\lambda^m/m! \qquad (20)$$

where P_m is the probability of having m peaks over the threshold in a year. In this study, selection of threshold level is made in terms of the variance-to-mean ratio, as proposed by Cunmane (1979). It is based on the idea that if the time of occurrences of peaks follows a Poisson process, then the ratio of the mean number of peaks occurring in each year to its variance should be close to or equal to unity.

Figure 1 shows the plot of ratio of the observed variance to observed mean number of exceedances per year for the site 02ED003 (Nottawasaga near Baxter, used as an illustration site through out this paper), together with the 5% significant level for a chi-square test of significance. Examining Figure 1, together with the variance-to-mean ratio plots on the remaining sites, it can be concluded that none of the sites follows a

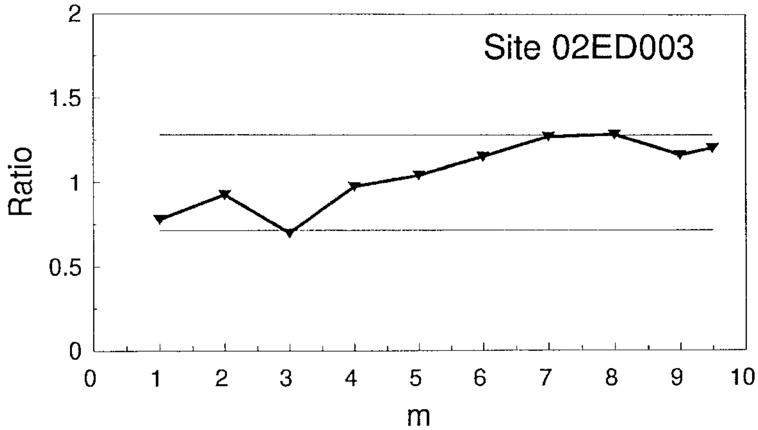

Figure 1. Ratio of observed variance to observed mean of number of exceedances per year over a threshold (thick line) and 90% confidence interval (thin line) for illustration site 02ED003

Poisson distribution exactly. Any pronounced departure in the plot was removed by raising the value of x_0, or lowering the value of m (mean number of exceedances per year) so as to obtain a ratio close to one. In theory, once the lowest Poisson-admissible threshold level produces independent peaks, then for any other higher threshold levels, the corresponding arrival rates of exceedances will follow a Poisson process too. This lowest Poisson-admissible threshold level will produce the highest number of exceedances, and hence the most reliable parameter estimates can be expected. The mean number selected for each site based on the ratio is presented in Table I.

The assumption that the distribution of the number of values exceeding the chosen threshold level is Poisson distributed is checked by comparing the observed against the predicted values. Figure 2 shows that the Poisson distribution assumption is valid for 02ED003 for m equal to 4. Similar conclusions were obtained for other sites.

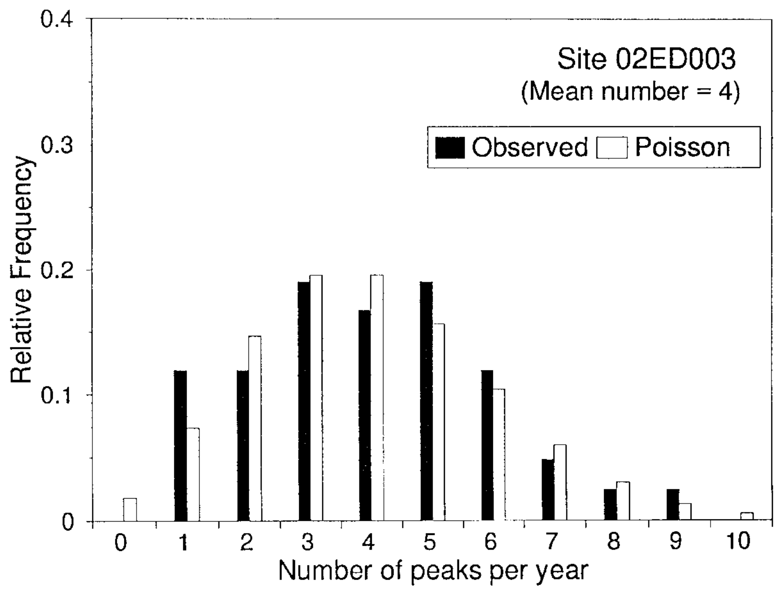

Figure 2. Comparison of the observed and predicted (by Poisson process) distribution of the occurrences/year of the threshold exceedances at site 02ED003

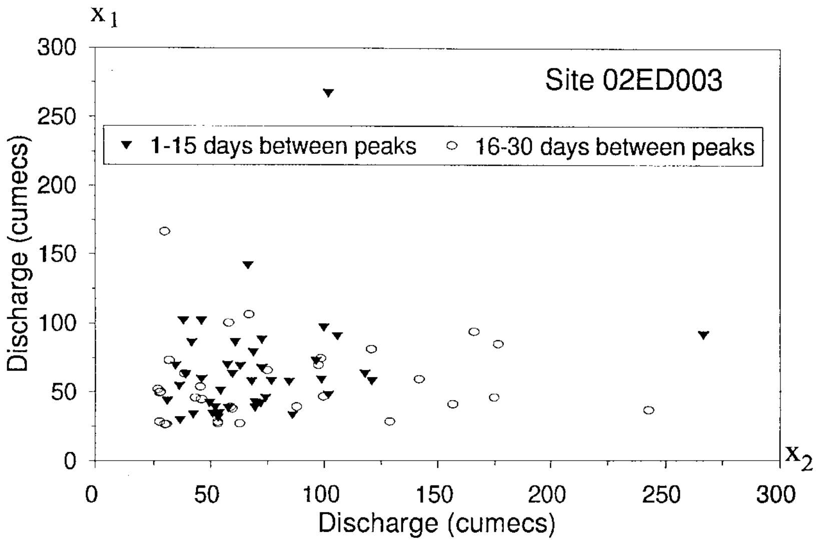

Figure 3. Magnitudes of successive peaks over a threshold occurring within 1–15 and 16–30 days of one another at site 02ED003 showing the PD series independence

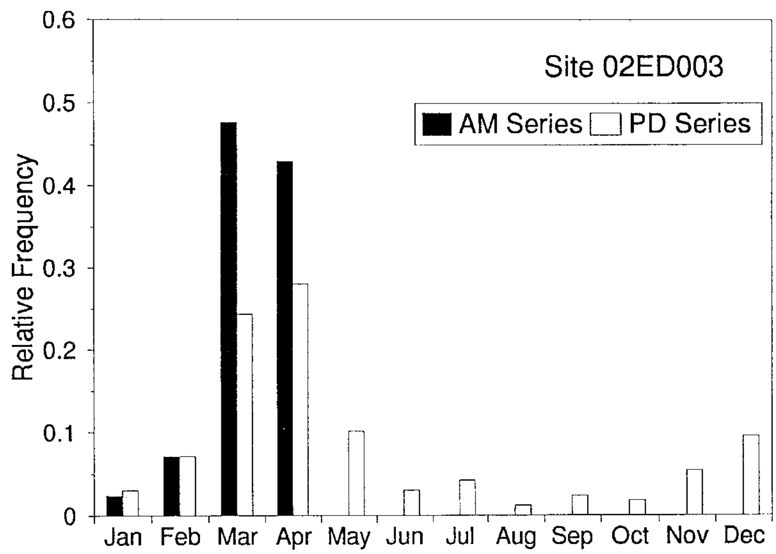

Figure 4. Comparison of relative frequencies of flood occurrences in each month for the AM series and PD series at site 02ED003

Test of independence

Statistical flood frequency analysis assumes that floods are independent random events. The validity of this assumption is tested using a scatter diagram showing successive peaks over a threshold occurring within specified days (subjectively separated, e.g. from 1 to 15 days, and 16 to 30 days) of one another. From the results presented in Figure 3, it can be concluded that successive peaks are independent.

Data screening for flood-generating mechanisms

The occurrence of AM and PD series on a monthly basis is shown in Figure 4 for the site 02ED003. It is observed that AM floods tend to occur more frequently during spring, when snowmelt may be the

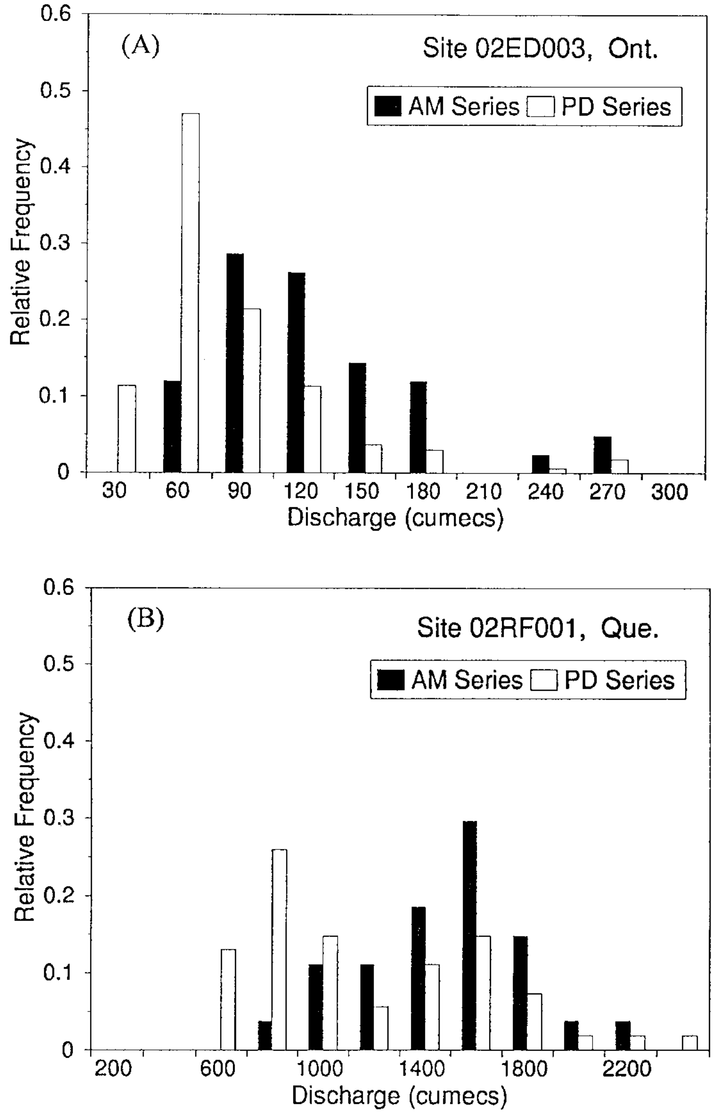

Figure 5. Comparison of relative frequencies of flood occurrences for the AM series and PD series at sites 02ED003 (A) and 02RF001 (B)

causative factor. However, PD floods can occur at any time of the year, indicating a mixture of causative flood factors.

Figures 5A and 5B present relative frequencies for the AM and PD series at sites 02ED003 and 02RF001, respectively. In general, both AM and PD series can exhibit unimodal or multimodal density function shapes depending on the flood-generating mechanisms. Figure 5A shows that site 02ED003 has a unimodal density function for both AM and PD series. However, for site 02RF001, as shown in Figure 5B, unimodal distribution is observed for the AM series, but the PD series is bimodal. It has also been observed that at some other sites PD series exhibit unimodal while AM series exhibit bimodal distributions. Such unimodal and bimodal shapes can be detected and described by the non-parametric method.

Probability density functions were also examined for data obtained from splitting the AM and PD series into three different seasons, i.e. snowmelt season (March to May), rain season (June to November) and

Table II. L-moments for site 02ED003 (Nottawasaga near Baxter) used for illustration

Model	L-moments				Minimum peak (m^3/s)
	L-location λ_1	L-scale λ_2	L-skewness τ_3	L-kurtosis τ_4	
Annual maximum	110·7	28·39	0·236	0·183	41·1
Annual exceedance	124·1	24·15	0·435	0·199	81·0
Partial duration	66·24	20·89	0·367	0·210	26·3

winter thaw season (December to February). Such division by mechanisms is not exact, since the end of the snowmelt season may vary with location. It was observed that for the AM series, most flood peaks occur in the snowmelt season, with the exception of station 02FF002. However, for the PD series, a significant number of floods may occur in any season. It was also noticed that the bimodal distribution of the PD series at station 02RF001 (as shown in Figure 5B) is composed of two unimodal distributions, one corresponding to the snowmelt season, the other to the rainfall season.

L-moments analysis

The L-moments for the Nottawasaga River near Baxter (02ED003) are presented in Table II. L-skewness indicates that the PD series exhibit larger skewness than the AM series. This is similar for the eight sites. There appears to be a relationship between bimodality and L-kurtosis. In general, bimodal data series exhibit lower L-kurtosis values. For the commonly used unimodal, two- or three-parameter frequency distributions in hydrology, the ability to represent the kurtosis of the observed data is an important aspect in the selection of an appropriate distribution. However, this may mislead in the correct choice of a distribution when a data set exhibits multimodality.

The computed L-moments ratio for the eight sites under study are plotted in Figure 6, together with the theoretical L-moments ratios of GEV (with Gumbel distribution as a special case) and GP (with the exponential distribution as a special case) distributions for the AM series. Figure 7 shows the computed

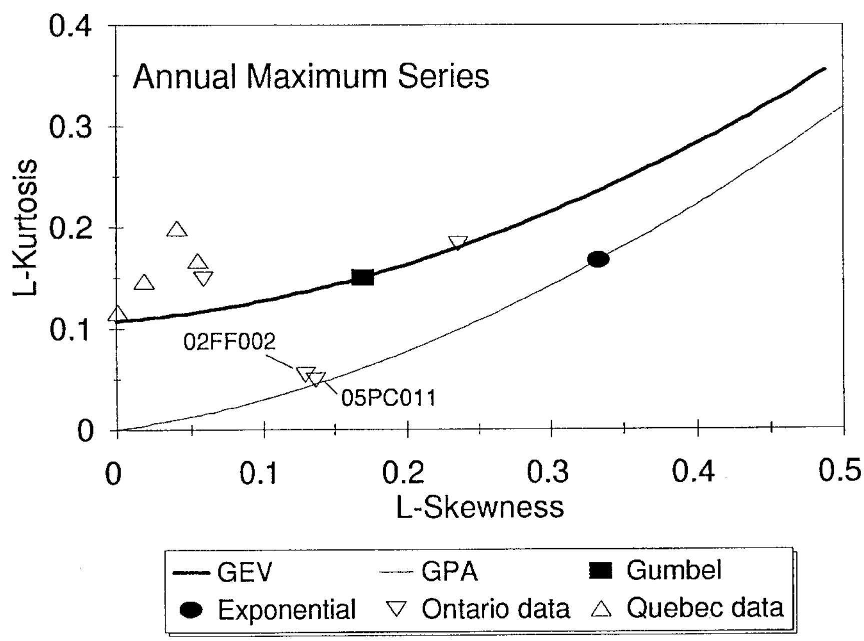

Figure 6. L-moment ratios diagram for AM series

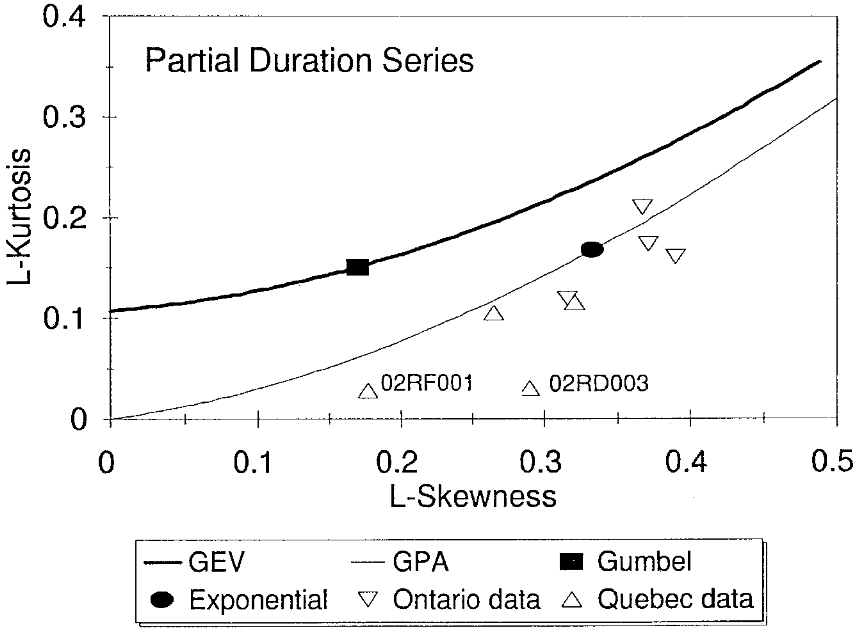

Figure 7. L-moment ratios diagram for PD series

L-moments ratio for the eight sites and those of GEV and GP for the PD series. Examination of these two figures shows that, amongst the four distributions, the GEV distribution is the best for describing the AM series while the GP distribution is the best for describing the PD series. However, while the L-moments ratio diagrams provide only guidelines in choosing a distribution, they do not guarantee choosing a correct one. In fact, no methods in present hydrological practice guarantee a correct choice of distribution for use in flood frequency analysis. For example, the observed L-moments ratios of sites 02FF002 and 05PC001 for the AM series fall right in the GP distribution range (see Figure 6) because of their low L-kurtosis values. However, by examining their corresponding density functions it was observed that the low-kurtosis values are caused by the bimodality of the AM series at these two sites.

Estimation of parameters

The estimated parameters of the GEV distribution [Equation (6)] and the bandwidth for the non-parametric approach [Equation (16)] are given in Table III for the AM series at site 02ED003. Table IV gives the bandwidth and the parameters for GP [Equation (5)] distribution for both the PD model and the AE model, where the AE model is a special case of the PD model with mean peak occurrences per year equalling one. For the PD model, the parameter k of the GP distribution is close to zero for the site 02ED003, indicating that an exponential distribution is applicable. As $k < 0$, the GEV and GP distributions are not bounded at the upper end. Flood estimation with the NP is upper bounded at $(x_{\max} + h)$, where $x_{\max}$ is the maximum observed flow in the data series.

Table III. Parameters of the generalized extreme value (GEV) and non-parametric distributions (NP) for the annual maximum (AM) model at site 02ED003 used for illustration

Model	GEV parameters [Equation (6)]			NP bandwidth [Equation (6)]
	ζ	β	k	h
Annual maximum	85·3	37	−0·0998	30·1

Table IV. Parameters of the generalized Pareto (GP) and non-parametric distributions (NP) for the annual exceedance (AE) and partial duration (PD) model at site 02ED003

Model	GEV parameters [Equation (5)]			NP bandwidth [Equation (16)]
	x_0	α	k	h
Annual exceedance	80·95	34·01	−0·2133	25·61
Partial duration	25·98	37·3	−0·0733	21·19

Table V. Flood quantiles (m^3/s) estimated by GEV, GP and NP with AM and PD models (stations 02ED003 and 02RF001 used for illustration)

Site no.	Return period (years)	AM model		AE model		PD model	
		GEV	NP	GP	NP	GP	NP
02ED003	50	262	259	288	263	268	263
	100	301	270	347	271	307	270
	200	344	279	414	277	348	275
02RF001	50	2240	2387	2300	2392	2390	2406
	100	2340	2473	2360	2470	2480	2491
	200	2430	2526	2410	2518	2610	2556

AM, annual maximum model; AE, annual exceedance model; PD, partial duration model; GEV, generalized extreme value; GP, generalized Pareto; NP, non-parametric

Theoretically, the parameters k for GEV/AM and GP/PD, as well as for GP/AE, should be equal in value under the Poisson distribution arrival rate assumption. However, Tables III and IV do not support this. Perhaps such discrepancies can be attributed to the fact that these distributions do not describe exactly all of the data under study, or, possibly that such differences are due to sampling errors in the data.

Figure 8A and 8B shows the estimated density function by the non-parametric method for both AM and PD series for sites 02ED003 and 02RF001, respectively. The density function at site 02ED003 is unimodal for both the AM and PD series. For site 02RF001, the density function for the AM series is unimodal, while it is bimodal for the PD series. This result is consistent with that observed in Figure 5. Such bimodality cannot be easily detected or described by parametric methods. The detection of bimodality or unimodality in this paper is visual, and therefore subjective.

Flood quantiles estimation

Flood quantiles corresponding to assumed 50-, 100- and 200-year return periods were estimated from each fitted distribution for AM and PD models for each site. Table V presents the results for sites 02ED003 and 02RF001.

Examination of Table V shows that parametric distribution results are very sensitive to the skewness of data, especially for higher return periods. For example, site 02RF001 has a small skewness (L-skewness $< 0{\cdot}177$), and the flood quantiles estimated by both parametric and non-parametric methods are quite close to each other. However, the differences are large for site 02ED003 which has a large skewness value ($> 0{\cdot}236$). In addition, the sensitivity of a distribution to the skewness of observed data is distribution dependent. The GP distribution has a higher ability than GEV to accommodate high skewness and long tail data, which is usually true for the PD data series.

For almost all the sites, the parametric estimations of flood quantiles are significantly different for GEV/AM, GP/AE and GP/PD. For example, for the site 02RF001, the flood quantile estimated by GEV/AM is 2430 (m^3/s), while the corresponding estimation by the GP/PD is 2610 (m^3/s). The corresponding results for the non-parametric models, however, are more consistent. For the non-parametric method, the estimations obtained from both AM and PD series are very close (2526 and 2556, respectively).

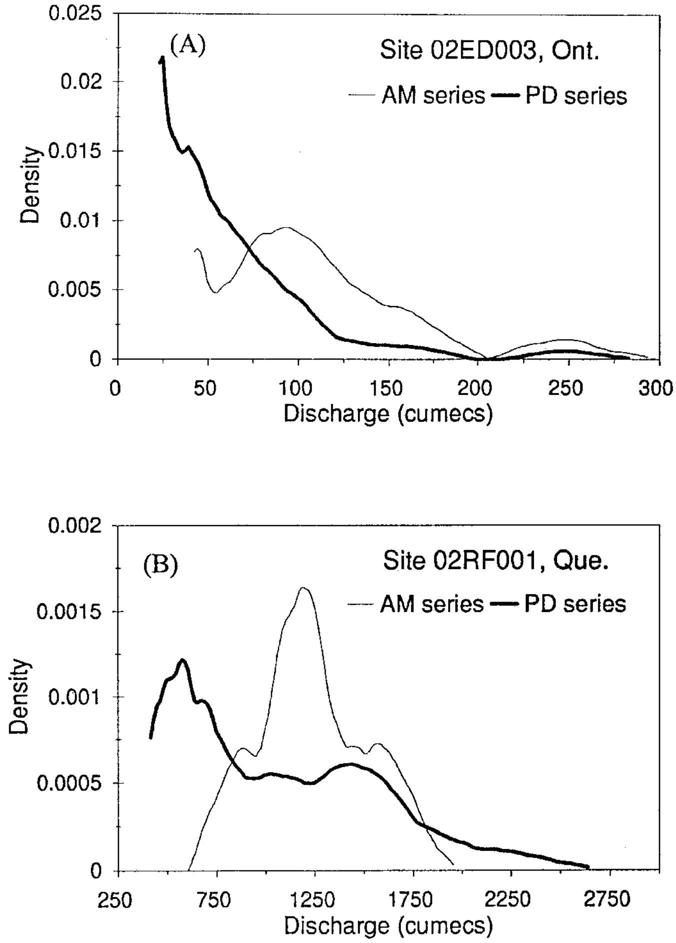

Figure 8. Probability density functions estimated by the non-parametric method for both the AM and PD series at sites 02ED003 (A) and 02RF001 (B)

For the PD series analysis, the parametric method is very sensitive to the choice of threshold level. For example, the 200-year flood quantile estimate at the site 02ED003 by the GP distribution from the PD series ($m = 4$) is 348 (m^3/s), while from the AE series ($m = 1$) the corresponding value is 414 (m^3/s). However, the corresponding non-parametric estimates are 275 and 277 for the PD and AE series, respectively.

Figure 9 shows a comparison of distribution fitting of the AM data at site 02ED003 by the GEV and non-parametric methods. Figure 10 shows the comparison between the results obtained from the GP and non-parametric methods for fitting the PD series. Clearly, the methods give different results, particularly at the upper tail of the distribution. The non-parametric method approximates the observed data very closely.

CONCLUSIONS

The analysis of annual maxima (AM) and partial duration (PD) series by parametric and non-parametric methods revealed the following.

1. Both the AM and PD series can be bimodal, which can be described by the NP method. Currently used parametric distributions, such as GEV and GP, are unimodal and cannot describe bimodal data, unless a mixture model is used.

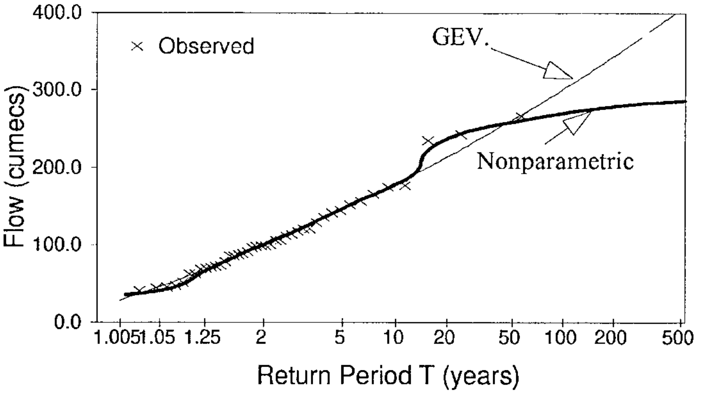

Figure 9. Comparison of flood quantiles estimated by GEV and non-parametric distributions for the AM series for site 02ED003

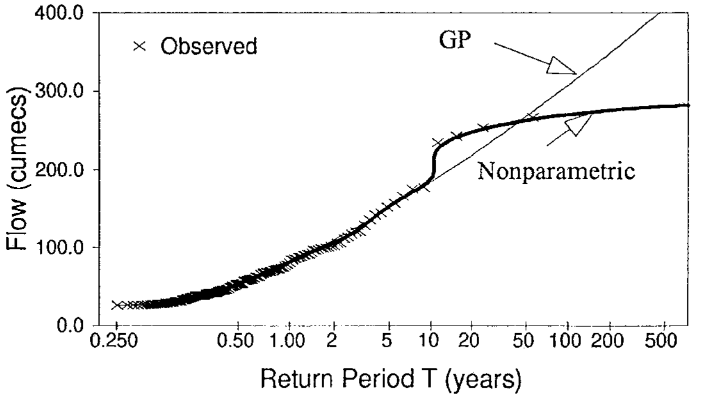

Figure 10. Comparison of flood quantiles estimated by GP and non-parametric distributions for the PD series for site 02ED003

2. Parametric distributions, when applied to both AM and PD data, provide quantile estimates that differ significantly from each other, and are sensitive to threshold level choice. However, the NP method is not sensitive to the data type and choice of threshold level, and provides quantile estimates that do not differ very significantly when applied to both AM and PD data.
3. Unlike parametric methods, which assume a certain distribution for the observed data, the non-parametric approach has no such restriction. Because of its simplicity and accuracy, the non-parametric method is therefore equally applicable to both AM and PD series flood frequency analysis.
4. In summary, it can be concluded that the non-parametric method: (a) is superior to the parametric method because it is based on a more accurate fit of the AM and PD series to the data; (b) lacks sensitivity to decisions about threshold levels selection, series definition and choice of probability distributions; and (c) is more in tune with the often encountered bimodal nature of flood series.

ACKNOWLEDGEMENTS

Financial support provided by the Natural Sciences and Engineering Research Council of Canada (Strategic Grant No. G15430) is gratefully acknowledged. Dr D. Gingras and Dr PJ Pilon are thanked for stimulating discussions.

REFERENCES

Adamowski, K. 1985. 'Nonparametric kernel estimation of flood frequencies', *Wat. Resour. Res.*, **21**, 1585–1590.
Adamowski, K. 1989. 'A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies', *J. Hydrol.*, **108**, 295–308.
Adamowski, K., and Feluch, W. 1990. 'Nonparametric flood-frequency analysis with historical information', *J. Hydraul. Eng.*, **116**, 1035–1047.
Ashkar, F., El-Jabi, N., and Bobee, B. 1987. 'On the choice between annual flood series and peaks over threshold series in flood frequency analysis', in Feldman, A. D. (ed.), *Proc. Eng. Hydrol., Virginia, August 3–7, 1987*. ASCE, New York. pp. 276–280.
Birikundavyi, S., and Rousselle, J. 1997. 'Use of partial duration series for single-station and regional analysis of floods', *J. Hydrol. Eng., ASCE*, **2**, 68–75.
Cunnane, C. 1973. 'A particular comparison of annual maximum and partial duration series methods of flood frequency prediction', *J. Hydrol.*, **18**, 257–271.
Cunnane, C. 1979. 'A note on the Poisson assumption in partial duration series models', *Wat. Resour. Res.*, **15**, 489–494.
Gingras, D., Adamowski, K., and Pilon, P. 1994. 'Regional flood equations for the provinces of Ontario and Quebec', *Wat. Resour. Bull.*, **30**, 55–67.
Hosking, J. R. M. 1986. *The Theory of Probability Weighted Moments*, Research Report RC 12210. IBM Research, Yorktown Heights.
Hosking, J. R. M. 1990. 'L-moments: analysis and estimation of distribution using linear combinations of order statistics', *J. R. Statist. Soc. B.*, **52**, 105–124.
Hosking, J. R. M. 1991. *Fortran Routines for Use with the Method of L-Moments, Version 2*, Research Report RC 17097. IBM Research, Yorktown Heights.
Hosking, J. R. M., and Wallis, J. R. 1987. 'Parameter and quantile estimation for the generalized Pareto distribution', *Technometrics*, **29**, 339–349.
Labatiuk, C., and Adamowski, K. 1987. 'Application of nonparametric density estimation to computation of flood magnitude/frequency', in MacNeill, I. B. and Umphrey, G. J. (eds), *Stochastic Hydrology*. D. Reidel, Dordrecht, pp. 161–180.
Lall, U., Moon, Y.-I. and Bosworth, K. 1993. 'Kernel flood frequency estimators: bandwidth selection and kernel choice', *Wat. Resour. Res.*, **29**, 1003–1015.
Madsen, H., Rasmussen, P. R., and Rosbjerg, D. 1997. 'Comparison of annual maximum series and partial duration series methods for modelling extreme hydrologic events. 1. At-site modelling', *Wat. Resour. Res.*, **33**, 747–757.
Muller, H. G. 1991. 'Smooth optimum kernel estimators near endpoints', *Biometrika*, **78**, 521–530.
Pickands, J. 1975. 'Statistical inference using extreme order statistics', *Ann. Stat.*, **3**, 119–130.
Rasmussen, P. F. 1994. 'A regional index flood procedure based on partial duration series', *INRS-Eau Report*. Institut National de la Recherche Scientifique (INRS), Université du Québec, Québec, Canada, GIV 4C7.
Rasmussen, P. F., Ashkar, F., Rosbjerg, D., and Bobee, B. 1993. 'The POT method for flood estimation: a review', in *Proc. Stochastic and Statistical Methods in Hydrology and Environmental Engineering, Waterloo, June 21–23, 1993*. University of Waterloo.
Rosbjerg, D. 1985. 'Estimation in partial duration series with independent and dependent peak values', *J. Hydrol.*, **76**, 183–195.
Wang, Q. J. 1991. 'The POT model described by the generalized Pareto distribution with Poisson arrival rate', *J. Hydrol.*, **129**, 263–1286.
Waylen, P., and Woo, M. K. 1987. 'Prediction of annual floods generated by mixed processes', *Wat. Resour. Res.*, **18**, 1283–1286.
Yevjevich, V., and Taesombut, V. 1979. 'Information on flood peaks in daily flow series', in McMean, E. A., Hipel, K. W., and Unny, T. E. (eds), *Input for Risk Analysis in Water Systems*. Water Resources Publications, Fort Collins. pp. 171–192.

