

Estimating the Upper Tail of Flood Frequency Distributions

JAMES A. SMITH

Interstate Commission on the Potomac River Basin, Rockville, Maryland

Procedures for estimating recurrence intervals of extreme floods are developed. Estimation procedures proposed in this paper differ from standard procedures in that only the largest 10–20% of flood peaks are explicitly used to estimate flood quantiles. Quantile estimation procedures are developed for both annual peak and seasonal flood frequency distributions. The underlying model of flood peaks is a marked point process $\{T_j^i, Z_j^i\}$, where T_j^i represents time of occurrence of the j th flood during year i and the mark Z_j^i represents magnitude of the flood peak. Results from extreme value theory are used to parameterize the upper tail of flood peak distributions. Quantile estimation procedures are applied to the 92-year record of flood peaks from the Potomac River. Results suggest that Potomac flood peaks are bounded above. The estimated upper bound is only 20% larger than the flood of record.

1. INTRODUCTION

The classical approach to flood frequency analysis is to treat the sequence of instantaneous annual peak discharges over a period of n years, Y_1, \dots, Y_n , as independent and identically distributed (IID) random variables with distribution function F . We are primarily concerned with estimating quantiles of F ,

$$Q(\alpha) = \inf \{x: F(x) \geq \alpha\} \quad \alpha \in [0, 1] \quad (1)$$

where α is typically very close to 1; the value of the 100-year flood, for example, is given by $Q(0.99)$. In the classical framework quantile estimators are obtained after first estimating all of the parameters of F . If, for example,

$$F(x) = 1 - \exp \left\{ - \left(\frac{x - \mu}{\sigma} \right)^k \right\} \quad (2)$$

is the Weibull distribution, the maximum likelihood estimator of $Q(\alpha)$ is given by

$$\hat{Q}(\alpha) = \hat{\mu} + \hat{\sigma} [-\log(1 - \alpha)]^{k^{-1}} \quad (3)$$

where $\hat{\mu}$, $\hat{\sigma}$, and \hat{k} are maximum likelihood estimators of the parameters of F obtained from Y_1, \dots, Y_n .

The procedures we present for estimating quantiles of flood frequency distributions are motivated by *DuMouchel's* [1983] suggestion to "let the tails speak for themselves." In practice, this suggestion means that only the upper order statistics should be used to estimate the upper tail of a distribution. Similar approaches have been proposed for flood frequency analysis. *Prescott and Walden* [1983] propose that parameters of a specified flood frequency distribution F should be estimated by censored maximum likelihood with censoring applied so as to retain the upper order statistics of the annual peak sample. Procedures presented in this paper for flood frequency analysis are closely related to censored maximum likelihood methods; there are, however, fundamental differences.

The quantile estimation procedures proposed in sections 3 and 5 differ from censored maximum likelihood methods in that we do not specify a parametric form of the annual peak distribution F . Instead we use results from extreme value

theory to specify a parametric form for the "tail distribution,"

$$F_u(y) = P\{Y_i - u \leq y | Y_i > u\} \quad (4)$$

Our procedures will use the largest 10–20% of observations to estimate the tail distribution F_u .

Connections between extreme value theory and flood frequency analysis have been close. In *Statistics of Extremes*, *Gumbel* [1958] suggests that annual flood peaks, by virtue of their representation as the maxima of numerous loosely connected events, should follow one of the extreme value distributions. Extreme value theory is not, however, used in this paper to specify the annual peak distribution, but rather to specify the tail distribution (4). The main result we use is due to *Pickands* [1975]. He shows that under certain assumptions on F (see section 2) the tail distribution (4), for sufficiently large u , can be accurately approximated by a generalized Pareto distribution. Utility of this result for problems of flood frequency analysis is suggested by *Smith* [1984]. *Pickand's* result forms the basis of quantile estimation procedures presented in sections 3 and 5.

Bryson [1974] proposes that different types of upper tail behavior can be distinguished from the conditional mean exceedance (cme) function:

$$M(u) = E[Y_i - u | Y_i > u] \quad (5)$$

The conditional mean exceedance $M(u)$ is the average amount by which an annual peak exceeds a threshold u given that it is larger than u . It follows from *Pickand's* theorem that there are three possible types of upper tail behavior that the cme function can exhibit (1) an unbounded "thick-tailed" distribution has cme function that is approximately linearly increasing in the upper tail, (2) an unbounded "thin-tailed" distribution has constant cme function in the upper tail, and (3) a bounded distribution has cme function that is approximately linearly decreasing in the upper tail. Figure 1 shows the estimated cme function for instantaneous annual peaks of the Potomac River (1895–1986). Note that beyond 200,000 cubic feet per second (cfs) (1 cfs = 0.283 m³/s) (there are 9 larger floods) the cme function appears to be linearly decreasing suggesting that Potomac flood peaks are bounded; we return to this topic in section 3.

Two quantile estimation problems are considered in this paper. The topic of section 3 is quantile estimation for the upper tail of the annual peak distribution. In section 5, seasonally varying quantile estimators are developed. Interest in

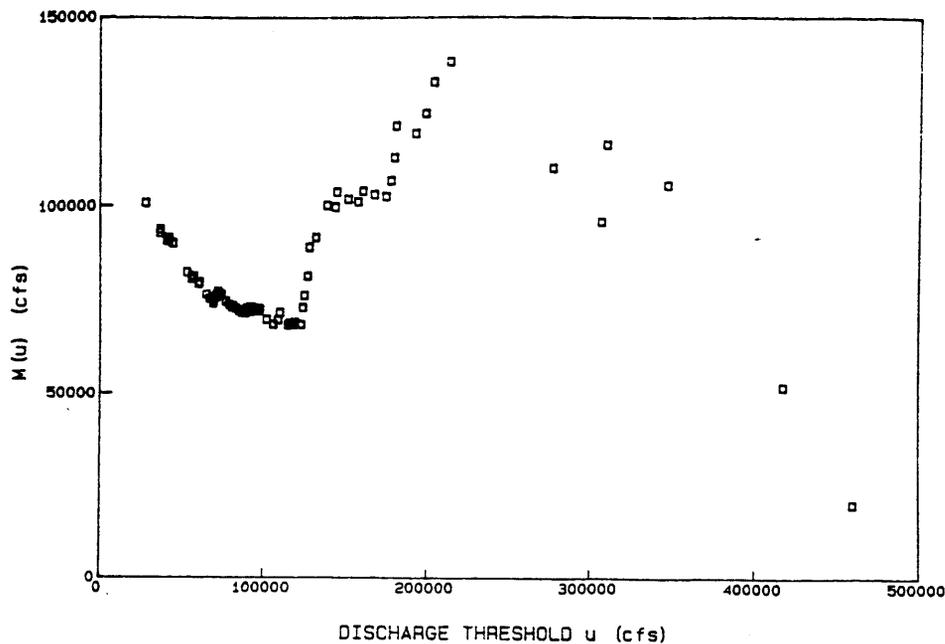


Fig. 1. Conditional mean exceedance function for annual flood peaks of the Potomac River. One cfs = 0.283 m³/s.

time-varying flood frequency estimation stems in part from reservoir regulation problems in which it is desired to allow conservation storage for flood protection to depend on seasonally varying flood risk (see, for example, *Smith and Karr* [1986]). For both estimation problems the underlying probability model for flood peaks is a peaks over threshold (partial duration series) model, which is presented in section 2. Both quantile estimation procedures are applied to instantaneous flood peak data for the Potomac River.

To motivate and support assumptions made in developing quantile estimators, results are presented in section 4 which characterize relationships between annual peak distributions and seasonally varying flood peak distributions. Section 4 serves to link problems of seasonal and annual peak quantile estimation. Results from this section are of independent interest in assessing the role of seasonal mixture distribution models (see, for example, *Waylen and Woo* [1982] and *Leytham* [1984]) in flood frequency analysis. In theorem 1 a general representation for the annual peak distribution of a seasonal mixture model is presented. Subsequent corollaries and examples illustrate pitfalls and insights that can be obtained from seasonal mixture models.

2. DEFINITIONS AND NOTATION

The times of occurrence of flood peaks, that is, exceedances of a discharge threshold u_0 , are modeled as a point process on the interval $[0, 1]$. Time 0 corresponds to the beginning of the year (which we take to be October 1) and time 1 corresponds to the end of the year (September 30). Denote by $N^i(1)$ the total number of flood peaks during year i and for $N^i(1)$ non-zero denote the occurrence times by $T_1^i, \dots, T_{N^i(1)}^i$ and the flood magnitudes by $Z_1^i, \dots, Z_{N^i(1)}^i$. The counting processes $\{N^i(t), t \in [0, 1], i = 1, 2, \dots\}$ are defined by

$$N^i(t) = 0 \quad N^i(1) = 0 \text{ or } t < T_1^i \quad (6a)$$

$$N^i(t) = n \quad T_n^i \leq t < T_{n+1}^i \quad (6b)$$

$$N^i(t) = N^i(1) \quad t \geq T_{N^i(1)}^i \quad (6c)$$

We assume that the "marked point processes" $\{(T^i, Z^i), i = 1, 2, \dots\}$ are IID (see *Karr* [1986] for additional definitions and results concerning marked point processes). Related point process models of flood peaks are presented by *Todorovic and Zelenhasic* [1970], *Gupta and Duckstein* [1976], and *Karr* [1976].

The sequence of annual peaks can be obtained from the marked point process as

$$Y_i = \max \{Z_j^i, j = 1, \dots, N^i(1)\} \quad N^i(1) > 0$$

$$Y_i = 0 \quad N^i(1) = 0 \quad (7)$$

It follows from the IID assumption on $\{(T^i, Z^i)\}$ that annual peaks are IID. We denote their common distribution function by

$$F(x) = P\{Y_i \leq x\} \quad x \geq 0 \quad (8)$$

We are particularly interested in estimating attributes of the upper tail of F . An important attribute of the upper tail is the upper bound of F

$$x_F = \sup \{x: F(x) < 1\} \quad (9)$$

with interest focusing on whether $x_F < +\infty$ or $x_F = +\infty$. If x_F is finite we are quite interested in estimating it. The quantile estimation procedure we present in section 3 yields an estimator of the upper bound in the case that F is bounded.

We conclude this section with the necessary background material from extreme value theory for development of quantile estimators. A detailed treatment of extreme value theory with numerous engineering examples can be found in the work by *Leadbetter et al.* [1983] (see also *Gumbel* [1958] and *de Haan* [1976]).

Let Y_1, Y_2, \dots be IID random variables with common distribution F . Denote the maxima of the first n by

$$M_n = \max (Y_1, \dots, Y_n) \quad (10)$$

(The annual peak notation Y_i and F is used intentionally to

suggest that the natural way to introduce extreme value theory into flood frequency analysis is through the tail of the annual peak distribution, not the maxima of peaks within a year.) According to the central result of extreme value theory, the extremal types theorem, if a nondegenerate limit distribution Λ exists such that

$$\lim_{n \rightarrow \infty} P\{a_n^{-1}(M_n - b_n) \leq x\} = \lim_{n \rightarrow \infty} F^n(a_n x + b_n) = \Lambda(x) \quad (11)$$

for appropriate scaling sequences $a_n > 0$ and b_n , then Λ must be one of the three extreme value distributions.

The three extreme value classes can be represented in terms of the generalized extreme value distribution as

$$\Lambda(x | \mu, \sigma, k) = \exp \{-[1 - k\sigma^{-1}(x - \mu)]^{1/k}\} \quad (12)$$

where $k(x - \mu) < \sigma$, $\sigma > 0$, and $\mu \in \mathbf{R}$. The three extreme value classes are distinguished as follows:

$$\text{extreme value type I: } k = 0 \quad (13a)$$

$$\text{extreme value type II: } k > 0 \quad (13b)$$

$$\text{extreme value type III: } k < 0 \quad (13c)$$

If (11) holds for a distribution F it is said to belong to the extreme value domain of attraction of Λ . Practical importance of the extremal types theorem derives from the fact that virtually all continuous "textbook" distributions have an extreme value domain of attraction.

The conditional exceedance distribution of F is defined, for $u < x_F$, by

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)} \quad y \leq x_F - u \quad (14)$$

For a fixed threshold u the conditional exceedance distribution $F_u(y)$ is the conditional probability that Y_i is less than or equal to $u + y$ given that it is larger than u , that is,

$$F_u(y) = P\{Y_i \leq u + y | Y_i > u\} \quad (15)$$

Closely associated with conditional exceedance distributions is the generalized Pareto distribution

$$G(y | \sigma, k) = 1 - (1 - k\sigma^{-1}y)^{1/k} \quad k \neq 0 \\ = 1 - \exp\{-\sigma^{-1}y\} \quad k = 0 \quad (16)$$

where $\sigma > 0$ and $k \in \mathbf{R}$. If $k > 0$, G has an upper bound given by

$$x_G = \sigma k^{-1} \quad (17)$$

The density of G is given by

$$g(y | \sigma, k) = \sigma^{-1}(1 - k\sigma^{-1}y)^{(k-1)} \quad k \neq 0 \\ = \sigma^{-1} \exp\{-\sigma^{-1}y\} \quad k = 0 \quad (18)$$

The distribution F has "generalized Pareto tail" [Pickands, 1975] with parameter k if

$$\lim_{u \rightarrow x_F} \inf_{0 < \sigma < \infty} \sup_{0 \leq y < \infty} \|F_u(y) - G(y | \sigma, k)\| = 0 \quad (19)$$

In words, F has a generalized Pareto tail if its conditional exceedance distribution can be approximated accurately by a generalized Pareto distribution. Importance of the generalized Pareto distribution stems from a theorem of Pickands [1975], which states that a distribution function F has generalized

Pareto tail with parameter k if and only if it has extreme value domain of attraction with parameter k . Because most continuous textbook distributions have an extreme value domain of attraction, it follows from Pickand's result that most continuous textbook distributions have generalized Pareto tails.

3. QUANTILE ESTIMATION FOR ANNUAL PEAK DISTRIBUTIONS

A number of quantile estimation procedures have been proposed which are based on Pickand's [1975] theorem. In this section we present and apply a quantile estimation procedure due largely to DuMouchel [1983] and Smith [1984], which is particularly well suited to flood peak data. The fundamental idea behind the method is that a discharge threshold u can be chosen such that the generalized Pareto approximation of Pickand's theorem holds as an equality for floods larger than u , that is,

$$P\{Y_i - u \leq x | Y_i > u\} = G(x) \quad (20)$$

where G is the generalized Pareto distribution (16).

The quantile function of F can be obtained from (20), and (14) for $x > F(u)$ as

$$Q(x) = u + G^{-1}\left(\frac{x - F(u)}{1 - F(u)}\right) \quad (21)$$

Further, using (21) and (16) for $k \neq 0$,

$$Q(x) = u + \sigma k^{-1} \left[1 - \left(\frac{1 - \alpha}{p} \right)^k \right] \quad (22)$$

where $p = 1 - F(u)$. To obtain an estimator of $Q(x)$ we must first estimate the parameters p , σ , and k .

To estimate p we note that the 0-1 random variables $\{1(Y_i > u), i = 1, 2, \dots\}$ is a sequence of Bernoulli trials with success probability $1 - F(u)$. We estimate p by

$$\hat{p} = n^{-1} \sum_{i=1}^n 1(Y_i > u) \quad (23)$$

Let \tilde{n} denote the random number of annual peaks larger than u and for $\tilde{n} > 0$ define

$$\tilde{Y}_1 = \min \{Y_i - u : Y_i - u > 0, i = 1, \dots, n\} \quad (24)$$

$$\tilde{Y}_j = \min \{Y_i - u : Y_i - u > \tilde{Y}_{j-1}, i = 1, \dots, n\} \quad (25)$$

for $1 < j \leq \tilde{n}$. The random variables $\tilde{Y}_1, \dots, \tilde{Y}_{\tilde{n}}$ are the ordered exceedances of u and by (20) comprise the order statistics of a random sample of size \tilde{n} from G . We can estimate σ and k from the log-likelihood function

$$L_n(\sigma, k) = \sum_{i=1}^{\tilde{n}} \log g(\tilde{Y}_i | \sigma, k) \quad (26)$$

where g is the density of the generalized Pareto distribution (18). We denote the estimators $\hat{\sigma}$ and \hat{k} . The estimator of $Q(x)$ becomes

$$\hat{Q}(x) = u + \hat{\sigma} \hat{k}^{-1} \left[1 - \left(\frac{1 - \alpha}{\hat{p}} \right)^{\hat{k}} \right] \quad (27)$$

For $\hat{k} > 0$ we obtain the following estimator for the upper bound of F ,

$$\hat{x}_F = u + \hat{\sigma} \hat{k}^{-1} \quad (28)$$

Smith [1985] shows that for $k < \frac{1}{2}$ standard asymptotic properties hold for maximum likelihood estimators of gener-

alized Pareto parameters. Denoting the "observed information matrix" by $V_n(k, \sigma)$, the "Fisher information matrix" by $I(k, \sigma)$, and the true parameters by k_0 and σ_0 we have

$$(\hat{k}, \hat{\sigma}) \xrightarrow{\mathcal{D}} (k_0, \sigma_0) \tag{29}$$

$$\bar{n}^{1/2}[(\hat{k}, \hat{\sigma}) - (k_0, \sigma_0)] \xrightarrow{\mathcal{D}} N(0, I(k_0, \sigma_0)^{-1}) \tag{30}$$

$$\bar{n}^{-1} V_n(\hat{k}, \hat{\sigma}) \xrightarrow{\mathcal{D}} I(k_0, \sigma_0) \tag{31}$$

where

$$V_n(k, \sigma) = - \left[\begin{array}{cc} \sum_{i=1}^n \frac{\partial^2 \log(g(\bar{Y}_i | \sigma, k))}{\partial k^2} & \sum_{i=1}^n \frac{\partial \log(g(\bar{Y}_i | \sigma, k))}{\partial k} \frac{\partial \log(g(\bar{Y}_i | \sigma, k))}{\partial \sigma} \\ \sum_{i=1}^n \frac{\partial \log(g(\bar{Y}_i | \sigma, k))}{\partial \sigma} & \sum_{i=1}^n \frac{\partial^2 \log(g(\bar{Y}_i | \sigma, k))}{\partial \sigma^2} \end{array} \right] \tag{32}$$

$$I(k, \sigma)^{-1} = \begin{bmatrix} (1-k)^2 & \sigma(1-k) \\ \sigma(1-k) & 2\sigma^2(1-k) \end{bmatrix} \tag{33}$$

Computation of (32) from (18) is straightforward. Note that the index \bar{n} is random. The results (29)–(32) rely on limit theorems for random sums of random variables [see *Serfling*, 1980].

From (29) it follows that if sufficient data are available, maximum likelihood estimators of k and σ will be close to the true parameter values. Standard errors of parameter estimates can be assessed using (30)–(32). From (30) it is seen that the asymptotic distribution of $(\hat{k}, \hat{\sigma})$ is bivariate normal with covariance matrix equal to the inverse Fisher information matrix. From (31) it follows that the observed information matrix evaluated at the maximum likelihood estimators is a consistent estimator of the Fisher information matrix. In (32) we present the form of the observed information matrix for the generalized Pareto distribution. Knowing the asymptotic joint distribution of $(\hat{k}, \hat{\sigma})$ we can compute the asymptotic distribution of functions of $(\hat{k}, \hat{\sigma})$, such as $\hat{Q}(x)$, using theorem 3.3A in the work of *Serfling* [1980].

In practice, two approaches have been used for specifying the discharge threshold u above which the generalized Pareto approximation is assumed to hold. One can specify a priori a fixed percent of the largest observations; *DuMouchel* [1983] suggests 10%. Alternately, one can use graphical tools such as the conditional mean exceedance plot as a guide in specifying the discharge threshold. Recall that in the upper tails the conditional mean exceedance plot is approximately linear.

For Potomac annual peaks the two approaches yield similar thresholds. Note in Figure 1 that the conditional mean exceedance plot has a sharp change in slope in the vicinity of 195,000 cfs beyond which it is approximately linearly decreasing. There are 10 annual flood peaks larger than 195,000 cfs in the 92-year record so a threshold of 195,000 cfs yields approximately the largest 10% of the observations (Table 1 contains a listing of the annual peak values).

Estimates of the generalized Pareto parameters (from equation (26)) for the 10 exceedances of 195,000 cfs are $\hat{k} = 0.38$ and $\hat{\sigma} = 146,000$. Most notably the estimate of k is positive indicating that floods are bounded above. The estimate of the upper bound from (28) is 579,000 cfs. The estimated upper bound is only 20% larger than the flood of record (480,000 cfs).

The estimated value of k is quite large, resulting in an estimated upper bound close to the flood of record. If we are to conclude from this evidence that Potomac flood peaks are definitely bounded we are on shaky ground. The standard error of \hat{k} obtained from the observed information matrix (using equation (32)) is 0.49. Considering the standard error of \hat{k} we cannot rule out any form of upper tail behavior. The Fisher information matrix provides some insight into the "error of estimation" problem for upper quantiles. From (30)–(33) we have, conditional on \bar{n} ,

$$\hat{k} - k_0 \sim N(0, \bar{n}^{-1}(1 - k_0)^2) \tag{34}$$

Substituting \hat{k} into the right-hand side of (34) yields a standard error estimate that is smaller than that obtained using the observed information matrix (and, in fact, biased low; see *Prescott and Walden* [1983]). For $\bar{n} = 10$ and $\hat{k} = 0.38$ (34) yields a standard error estimate of 0.19. From the optimistic standard error estimate of (34) we obtain the pessimistic result that \bar{n} must be greater than 10 to conclude that an estimate of 0.38 is more than 2 standard errors from 0. If we are using the largest 10% of floods this implies that more than 100 years of data are needed to conclude with modest certainty that floods are bounded. The situation is, of course, much worse using the appropriate standard error estimates obtained from the observed information matrix. (*Hosking* [1984] and *Hosking et al.* [1985] present related discussions of tests for different forms of upper tail behavior of annual peak distributions; see also *Shen et al.* [1980].)

As will be seen below accepting estimates of k different from 0 (both positive and negative) leads to "extreme" quantile estimates. A middle ground approach to upper tail quantile estimation, which, in practice, will always be supportable in light of the error of estimates problem outlined above, is to specify $k = 0$. Recall that for $k = 0$ we obtain an exponential upper tail. Using the generalized Pareto procedure with k specified to be 0 and a threshold of 195,000 cfs we obtain $\hat{\sigma} = 110,000$ yielding as quantile estimator,

$$\begin{aligned} \hat{Q}(x) &= u - \hat{\sigma} \log \left(\frac{1-x}{\hat{p}} \right) \\ &= 195,000 - 110,000 \log \left(\frac{1-x}{0.11} \right) \end{aligned} \tag{35}$$

Do we obtain qualitatively different results if the entire sample is used to estimate upper tail quantiles? We give a qualified answer below. The 92-year record of annual peaks was used to estimate parameters of the generalized extreme value distribution using the maximum likelihood estimation procedure of *Prescott and Walden* [1980, 1983]. The parameters estimates are $\hat{\mu} = 90,800$, $\hat{\sigma} = 41,000$, and $\hat{k} = -0.42$; the quantile estimator is given by

$$\hat{Q}(x) = \hat{\mu} + \hat{\sigma} \hat{k}^{-1} [1 - (-\log x)^{\hat{k}}] \tag{36}$$

The large negative value of \hat{k} indicates that Potomac flood peaks have thick unbounded upper tails.

Table 2 shows estimates of the 100-, 1000- and 10,000-year



TABLE 1. Instantaneous Annual Flood Peaks for the Potomac River at Point of Rocks

Year	Flood Peak, cfs
1895	66,800
1896	56,000
1897	204,000
1898	127,000
1899	128,000
1900	57,000
1901	161,000
1902	219,000
1903	110,000
1904	44,500
1905	71,400
1906	81,300
1907	119,000
1908	152,000
1909	83,000
1910	168,000
1911	106,000
1912	95,400
1913	139,000
1914	73,900
1915	139,000
1916	124,000
1917	123,000
1918	127,000
1919	80,500
1920	109,000
1921	88,800
1922	78,800
1923	40,700
1924	277,000
1925	89,000
1926	60,500
1927	89,900
1928	145,000
1929	180,000
1930	110,000
1931	36,800
1932	158,000
1933	123,000
1934	36,700
1935	128,000
1936	480,000
1937	310,000
1938	175,000
1939	124,000
1940	93,600
1941	69,000
1942	125,000
1943	418,000
1944	70,300
1945	139,000
1946	53,100
1947	42,100
1948	97,000
1949	132,000
1950	64,700
1951	128,000
1952	127,000
1953	118,000
1954	109,000
1955	214,000
1956	60,000
1957	69,200
1958	72,000
1959	55,700
1960	124,000
1961	102,000
1962	116,000
1963	125,000
1964	87,000

TABLE 1. (continued)

Year	Flood Peak, cfs
1965	97,800
1966	71,300
1967	144,000
1968	76,800
1969	27,800
1970	92,100
1971	86,400
1972	347,000
1973	106,000
1974	132,000
1975	181,000
1976	109,000
1977	193,000
1978	139,000
1979	178,000
1980	69,500
1981	41,900
1982	92,000
1983	115,000
1984	199,000
1985	84,700
1986	307,000

The drainage area of the Potomac River at Point of Rocks is 9651 m². One square mile equals 2.590 km²; 1 cfs = 0.283 m³/s.

floods from (1) generalized Pareto procedure (equation (27)), (2) the generalized Pareto procedure with $k = 0$ (equation (35)), and (3) the generalized extreme value distribution (equation (36)). The range of the three estimates for each return period is striking. The estimate of the 10,000-year flood from the generalized extreme value distribution is an order of magnitude larger than the estimate from the generalized Pareto procedure. Note that the generalized Pareto procedure with $k = 0$ provides "middle of the road" quantile estimates.

To conclude this section we examine sensitivity of generalized Pareto quantile estimates to the thinning threshold u . Table 3 contains parameter and quantile estimates (with standard errors in parentheses) for threshold values ranging from 120,000 to 190,000 cfs. Note that the estimate of k switches from positive to negative in the vicinity of 170,000 cfs. The most striking features are the results for 120,000 cfs (for which there are 40 exceedances). The estimate of k (-0.53) is smaller than the estimate obtained from the entire sample for the generalized extreme value distribution. With reference to Figure 1, there appear to be three distinct segments to the conditional mean exceedance function. Below 120,000 cfs (involving the smallest 52 floods) the cme function is decreasing. There is a sharp change in slope around 120,000 cfs and the cme function is increasing from 120,000 to 195,000 cfs. For the largest 10 floods, the cme function is approximately linearly decreasing.

It is clear that censored maximum likelihood applied to the largest 50–90% of observations will give radically different quantile estimates than the generalized Pareto procedure applied to the largest 10% of observations. A fundamental problem for estimating recurrence intervals of large floods is determining how much of the annual peak sample is relevant to the upper tail.

4. SEASONAL MIXTURE DISTRIBUTION MODELS

In this section we examine relationships between annual peak and seasonal flood peak distributions. This section serves

TABLE 2. Quantile Estimates for Generalized Pareto Procedure GP 1, Generalized Pareto Procedure with $k = 0$, GP 2, and Generalized Extreme Value Distribution

Recurrence Interval, years	GP 1	GP 2	GEV
100	425,000	451,000	652,000
1,000	515,000	704,000	1,728,000
10,000	553,000	958,000	4,556,000

GEV, generalized extreme value.

to link the problems of annual peak quantile estimation of section 3 and seasonal peak quantile estimation of section 5. Results of this section are pertinent to both problems.

Recall from section 2 that $\{T_j^i, Z_j^i\}$ is the marked point process of flood occurrence times and magnitudes. We will denote the seasonal distribution of flood magnitudes by

$$H(x|t) = P\{Z_j^i \leq x | T_j^i = t\} \quad x \geq 0 \quad (37)$$

that is, $H(x|t)$ is the conditional probability that a flood magnitude is less than or equal to x given that it occurs at time t during the year ($t \in [0, 1]$). Similarly, we define the seasonal conditional exceedance distribution by

$$H_u(x|t) = P\{Z_j^i - u \leq x | Z_j^i > u, T_j^i = t\} \quad (38)$$

Closely associated with the seasonal conditional exceedance distribution is the point process of flood peaks larger than u , which is defined by

$$N_u^i(t) = \sum_{j=1}^{N^i(t)} 1\{Z_j^i > u\} \quad t \in [0, 1] \quad (39)$$

The following theorem and corollaries characterize relationships between the annual peak distribution F and seasonal peak distributions $H(x|t)$. Proofs are given in the appendix.

4.1. Theorem 1

If $\{N^i\}$ is a Poisson process with intensity function $\lambda(t)$ then

$$F(x) = \exp \left\{ - \int_0^1 \lambda(s) [1 - H(x|s)] ds \right\} \quad x \geq 0 \quad (40)$$

For development of quantile estimators we would like to accommodate the possibility that the point process $\{N^i\}$ may not be Poisson. *Cervantes et al.* [1983] and *Smith and Karr* [1986] have shown that peaks over threshold records may exhibit clustering if small to moderate floods are included. Validity of the Poisson assumption for large floods, however, is supported by considerable empirical evidence [*Todorovic*, 1978], as well as theoretical arguments based on the Poisson limit theorem [*Cinlar*, 1972]. Because we are only interested in the upper tail of flood peak distributions there is a simple way of modifying (40) to account for the possibility that floods above u_0 are not Poisson: we raise the (arbitrary) base level u_0 to a sufficiently high threshold u and consider only the tail distribution F_u . The following corollary provides the necessary modifications.

4.2. Corollary 1

If for some $u \geq u_0$, $\{N_u^i\}$ is a Poisson process with intensity function $\lambda_u(t)$ then

$$F_u(x) = \exp \left\{ - \int_0^1 \lambda_u(s) [1 - H_u(x|s)] ds \right\} \quad x \geq 0 \quad (41)$$

4.3. Example 1

If we take flood magnitudes to have a seasonally varying exponential distribution

$$H(x|t) = 1 - \exp(-\beta(t)x) \quad (42)$$

it follows from theorem 1 that

$$F(x) = \exp \left\{ - \int_0^1 \lambda(s) \exp(-\beta(s)x) ds \right\} \quad (43)$$

If $\beta(t) = \beta$ for all t , so that the only seasonal variation is in the intensity function $\lambda(t)$, we have

$$F(x) = \exp \left\{ - \exp \left[-\beta x + \log \left(\int_0^1 \lambda(s) ds \right) \right] \right\} \quad x \geq 0 \quad (44)$$

Thus F has a (truncated) Gumbel distribution. Note that

$$\begin{aligned} F(0) &= \exp \left\{ - \int_0^1 \lambda(s) ds \right\} \\ &= P\{N^i(1) = 0\} \end{aligned} \quad (45)$$

"Truncation" at 0 thus accounts for the probability that no events occur during the year.

A compelling reason for using the generalized Pareto procedure for annual peak quantile estimation is that it is very difficult to specify the correct parametric form of the annual peak distribution. Theorem 1 and corollary 1 illustrate that seasonality plays an important role in determining the complexity of parametrizing annual peak distributions.

We now consider relationships between the upper tail of the annual peak distribution F and the upper tails of $H(x|t)$, $t \in [0, 1]$. We assume that for each t , $H(x|t)$ has an extreme value domain of attraction; we denote the generalized Pareto tail parameter by $k(t)$. The following result characterizes dependence of the annual peak tail on seasonal tails that have distinct "seasons." (Note that we use the term "uppertail" in this discussion in the sense of (19).)

4.4. Corollary 2

If there exist disjoint intervals A_1 and A_2 such that $A_1 \cup A_2 = [0, 1]$ and

$$\begin{aligned} H(x|t) &= H_1(x) \quad t \in A_1 \\ &= H_2(x) \quad t \in A_2 \end{aligned} \quad (46)$$

$$\begin{aligned} k(t) &= k_1 \quad t \in A_1 \\ &= k_2 \quad t \in A_2 \end{aligned} \quad (47)$$

TABLE 3. Generalized Pareto Estimates for Varying Thresholds

Threshold (No. of Floods)	\hat{k}	$\hat{\sigma}$	$\hat{Q}(0.99)$	$\hat{Q}(0.999)$	$\hat{Q}(0.9999)$
190,000	0.22	122,000	422,000	548,000	623,000
(11)	(0.49)	(70,000)	(54,000)	(182,000)	(355,000)
182,000	0.20	120,000	422,000	555,000	640,000
(12)	(0.45)	(64,000)	(57,000)	(189,000)	(370,000)
176,000	0.02	95,600	424,000	625,000	817,000
(13)	(0.49)	(54,600)	(76,000)	(340,000)	(823,000)
150,000	-0.08	81,200	427,000	687,000	997,000
(19)	(0.40)	(33,000)	(89,000)	(365,000)	(943,000)
120,000	-0.53	33,500	521,000	1,620,000	5,320,000
(40)	(0.28)	(10,300)	(213,000)	(1,520,000)	(8,476,000)

Standard errors computed from the observed information matrix are given in parentheses.

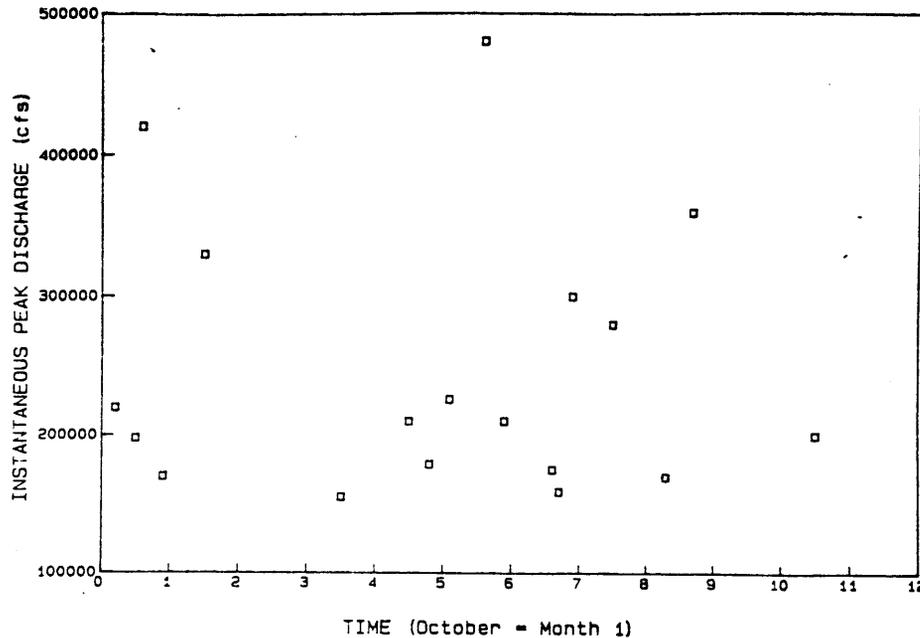


Fig. 2. Seasonal distribution of Potomac flood peaks larger than 150,000 cfs. Numbers shown are year of flood. One cfs = 0.283 m³/s.

then F has generalized Pareto tail with parameter k specified as follows: (1) if $k_1 > 0$ and $k_2 > 0$ and $x_{H_1} > x_{H_2}$ then $k = k_1$; and (2) if $k_1 \leq k_2$ and $k_1 \leq 0$ then $k = k_1$.

The main point of corollary 2 is that if flood peak distributions vary seasonally then the upper tail of the annual peak distribution depends only on the season with the "thickest upper tail." Corollary 2 can be generalized to an arbitrary number of seasons. If, for example, there are m seasons all with finite upper bounds, then the tail parameter of the annual peak distribution is determined by the season with the largest upper bound. If any of the seasons is unbounded, the annual peak tail is determined by the season with the smallest tail parameter.

Figure 2 shows time of occurrence and magnitude of all Potomac floods larger than 150,000 cfs. A notable feature is that the three largest floods are separated (seasonally) by at least 3 months and are similar in magnitude. A traditional "model" of floods for the northeastern United States (see, for example, Benson [1962]) holds that winter/spring is a season of numerous floods of small magnitude (often dominated by "snowmelt floods"), while summer/fall is a season of infrequent large floods (often dominated by "hurricane floods"). The largest Potomac flood occurs, however, during "snowmelt season" (March 1936). The second largest flood (October 1942) occurs during "hurricane season" but is not a hurricane flood. The third largest flood (June 1972) is a hurricane flood but occurs months before the peak of hurricane season.

Because the largest floods are spread throughout the year and are of comparable magnitude, there is little justification for concluding from the data that the seasonal peak distributions $H(x|t)$, $t \in [0, 1]$, differ in the upper tail. In light of corollary 2 we can either assume that (1) the importance of seasonality diminishes for extreme flood magnitudes or (2) we have too little data to assess seasonal features of extreme flood magnitudes. Assumption 1 is consistent with the available data for the Potomac, while assumption 2 is immune to judgement from data. If we accept assumption 2 we must also admit

that estimating return intervals for floods larger than the flood of record is virtually impossible.

As a final comment on Figure 2 we note that all Potomac floods larger than 150,000 cfs are annual peaks. The largest peak over threshold flood that is not an annual peak has a magnitude of 132,000 cfs. There are 26 annual peaks larger than 132,000 cfs. It follows that annual peak data contain virtually all of the information about the upper tail of annual peaks and seasonal flood peak distributions.

5. ESTIMATION OF SEASONAL QUANTILES

The main topic of this section is development of an estimation procedure for seasonal flood quantiles. Interest in seasonally varying flood frequency estimates stems in part from reservoir regulation problems in which it is desired to allow conservation storage for flood protection to vary seasonally in response to seasonally varying flood risk. Smith and Karr [1986] develop seasonally varying flood frequency estimators which incorporate covariate information such as snow pack and soil moisture storage. Their approach is designed for the central portion of flood frequency distributions and is not directly applicable to upper tails. Interest in this section, as in the entire paper, is strictly with the upper tails.

Our first task is to define a seasonally varying quantile function $Q_t(\alpha)$, $t \in [0, 1]$. Intuitively, Q_t should be the quantile function of a distribution F_t , that is the product of two terms (1) the conditional distribution of flood magnitudes at time t , $H(x|t)$, and (2) the probability of a flood at time t . A direct distributional approach to defining Q_t runs into trouble with the second term. As will be seen below, the natural approach for dealing with seasonal quantiles is the point process approach.

Recall that $\{N^i\}$ is the point process of floods larger than the base level u_0 . Its intensity function is defined by

$$\lambda(t) = \lim_{s \rightarrow 0} \left(\frac{1}{s} \right) P\{N^i(t+s) - N^i(t) \geq 1\} \quad (48)$$



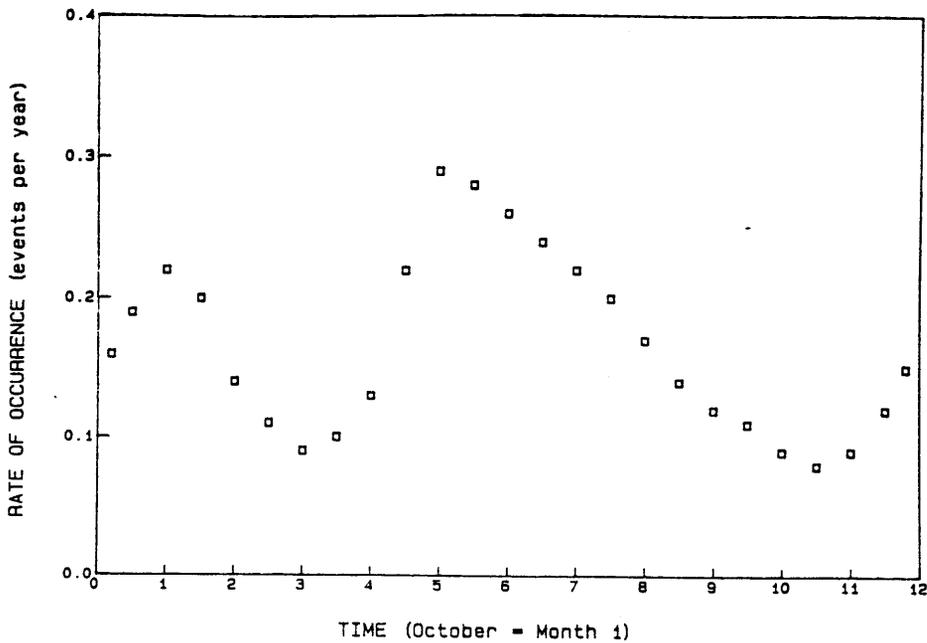


Fig. 3. Estimated seasonal rate of occurrence for Potomac floods larger than 150,000 cfs. (1 cfs = 0.283 m³/s).

The point process $\{N_u^i\}$ is the point process of all floods larger than u ($u > u_0$); its intensity function is denoted $\lambda_u(t)$. The quantile function Q_t can be defined as follows:

$$Q_t(\alpha) = \inf \{u: \lambda_u(t) \leq 1 - \alpha\} \quad \alpha \in [0, 1] \quad \alpha > 1 - \lambda(t) \quad (49)$$

Note that $\lambda_u(t)^{-1}$ is the "recurrence interval" at time t of a flood of magnitude u . We illustrate the definition for the peaks over threshold model of Todorovic and Zelenhasic [1970].

5.1. Example 2

In the model of Todorovic and Zelenhasic [1970], N^i is assumed to be a nonstationary Poisson process with intensity function $\lambda(t)$. Flood magnitudes are assumed to be IID with

$$H(x|t) = 1 - \exp(-\beta x) \quad (50)$$

Thus flood magnitudes are exponentially distributed and do not depend on time of year. In this case it is easy to show that N_u^i is a Poisson process with intensity function

$$\lambda_u(t) = \exp(-\beta u)\lambda(t) \quad (51)$$

(see, for example, Karr [1986]). It follows that

$$Q_t(\alpha) = -\beta^{-1} \log \left\{ \frac{1 - \alpha}{\lambda(t)} \right\} \quad \alpha > 1 - \lambda(t) \quad (52)$$

Our seasonal quantile estimation procedure is based on the following assumptions, which are supported for Potomac flood peaks by arguments in section 4.

1. A threshold \bar{u} can be chosen such that (1) $N_{\bar{u}}^i$ is a Poisson process with intensity function $\lambda_{\bar{u}}$ and (2) the distribution of flood peaks larger than \bar{u} does not depend on time of year; that is,

$$H_{\bar{u}}(x|t) = H_{\bar{u}}(x) \quad (53)$$

2. A threshold $u > \bar{u}$ can be chosen such that

$$H_u(x) = G(x|k, \sigma) \quad (54)$$

that is, a generalized Pareto approximation holds for flood peaks larger than u .

5.2. Theorem 2

Under the above assumptions,

$$Q_t(\alpha) = u + \sigma k^{-1} \left[1 - \left(\frac{1 - \alpha}{p \lambda_u(t)} \right)^k \right] \quad \alpha > 1 - p \lambda_u(t) \quad (55)$$

where

$$p = P\{Z_j^i > u | Z_j^i > \bar{u}\} \quad (56)$$

5.3. Proof

The result follows from the fact that N_{u+x}^i is a Poisson process with intensity function

$$\lambda_{u+x}(t) = p \lambda_u(t) [1 - G(x)] \quad (57)$$

To implement (55) we must estimate the parameters p , k , σ , and $\lambda_u(t)$, $t \in [0, 1]$, which we do as follows. By analogy with (24) p is estimated by

$$\hat{p} = \frac{\sum_{i=1}^n \sum_{j=1}^{N^i(1)} 1(Z_j^i > u)}{\sum_{i=1}^n \sum_{j=1}^{N^i(1)} 1(Z_j^i > \bar{u})} \quad (58)$$

The generalized Pareto parameters k and σ are estimated from the exceedances of u precisely as in (26). We denote the estimators, as before \hat{k} and $\hat{\sigma}$. The intensity function $\lambda_u(t)$ is estimated using a maximum likelihood procedure [see Karr, 1986].

For threshold values $\bar{u} = 150,000$ and $u = 195,000$, we obtain $\hat{k} = 0.38$, $\hat{\sigma} = 146,000$, $\hat{p} = 0.55$. Figure 3 shows the estimated intensity function. The intensity ranges from a maximum of 0.30 (units are events per year) in March to a minimum of 0.06 in August. Using (55) we can now estimate the time-varying 100-year flood quantile. The 100-year flood



quantile ranges from a maximum of 450,000 cfs in March to a minimum of 330,000 cfs in August.

6. SUMMARY AND CONCLUSIONS

In section 3 the generalized Pareto procedure for annual peak quantile estimation is described and applied to Potomac River flood peak data. Major conclusions are as follows.

1. The estimate of the tail parameter k of the annual peak distribution is positive (0.38), implying that Potomac flood peaks are bounded. This result suggests that the common assumption that flood peak distributions are unbounded above should be examined more closely.

2. The standard error of the estimate of k is large enough that we cannot definitively rule out any form of upper tail behavior. Due to the error of estimates problem for k it may be prudent to assume that the upper tail is exponential, that is, $k = 0$. On the other hand, small positive (or negative) values of k will lead to very different quantile estimates for floods of large return interval.

3. Standard errors of quantile estimates of very large recurrence interval floods (in the range of 1,000–10,000 year floods) are of the same order of magnitude as the estimates. Estimates in this range are thus of no utility. An important component of any quantile estimation procedure that purports to estimate very large recurrence interval floods is a method of assessing error of the estimates.

4. For the generalized Pareto procedure, severe censoring yields very different quantile estimates from moderate censoring. Using the largest 10% of annual peaks we conclude that flood peaks are bounded with an upper bound only 20% larger than the flood of record. Using the largest 40% of annual peaks we conclude that flood peaks are unbounded with very thick tails. For the Potomac data set the issue is not removing a few unrepresentative small floods, but rather deciding what the "upper tail" of Potomac flood peaks really is.

The main results of section 4 are theorem 1 which provides a general representation for the annual peak distribution of a seasonal mixture model and corollary 2 which characterizes dependence of the annual peak distribution on seasonal tails. Two major issues for flood frequency analysis are raised in this section. The principal motivation for estimating the upper tail of an annual peak distribution F from the largest order statistics is uncertainty in specification of the parametric form of F . We argue in section 4 that uncertainty in specification of F is well justified if seasonality is an important feature of the flood process. The second issue we consider is seasonality of extreme floods. It follows from corollary 2 that if seasonality is a prominent feature of extreme floods, serious difficulties for both annual peak and seasonal flood frequency analysis will result.

In section 5 we introduce the seasonal quantile function $Q_t(x)$, $t \in [0, 1]$ and develop generalized Pareto quantile estimators $\hat{Q}_t(x)$. Importance of seasonal quantile estimation stems in part from reservoir regulation problems in which it is desired to tie flood control operation to seasonally varying flood risk. The seasonal quantile estimation procedure is applied to Potomac flood peak data yielding a time-varying estimate of the 100-year flood that ranges from a maximum of 450,000 cfs in March to a minimum of 330,000 cfs in August.

Let $(\lambda(t))$ be a nonnegative function on $[0, 1]$. The point process N^i is a Poisson process on $[0, 1]$ with intensity function λ if (1) $\{N^i\}$ has independent increments, that is, for all k and $0 \leq s_1 < t_1 \leq \dots \leq s_k < t_k \leq 1$, the random variables $N^i(t_k) - N^i(s_k), \dots, N^i(t_1) - N^i(s_1)$ are independent and (2) for each $t \in [0, 1]$, $N^i(t)$ has a Poisson distribution with mean $\int_0^t \lambda(s) ds$.

Lemma: The zero probability function of a Poisson process $\{N^i\}$ on $[0, 1]$ with intensity function λ is given by

$$P\{N^i(t) = 0\} = \exp \left\{ - \int_0^t \lambda(s) ds \right\} \quad t \in [0, 1] \quad (A1)$$

The proof of theorem 4.1 follows from (A1) as follows. The annual peak distribution is given by

$$F(x) = P\{Y_i \leq x\} \\ = P\{N_x^i(1) = 0\} \quad (A2)$$

Recall that N_x^i is the counting process of floods larger than x and can be represented as

$$N_x^i(t) = \sum_{j=1}^{N^i(t)} 1(Z_j^i > x) \quad t \in [0, 1] \quad (A3)$$

By assumption N^i is a Poisson process with intensity function λ . It follows from (A3) that N_x^i is a Poisson process with intensity function

$$\lambda_x(t) = \lambda(t)(1 - H(x|t)) \quad t \in [0, 1] \quad (A4)$$

The theorem follows by applying the lemma to $N_x^i(1)$.

To prove corollary 2 we note first that for sequences $a_n > 0$ and b_n ,

$$\lim_{n \rightarrow \infty} F^n(a_n x + b_n) = \Psi(x) \quad (A5)$$

for some nondegenerate limit distribution Ψ if and only if

$$\lim_{n \rightarrow \infty} n \log F(a_n x + b_n) = \log \Psi(x) \quad (A6)$$

We also have that $-\log \Psi(x) \approx (1 - \Psi(x))$ as $\Psi(x) \rightarrow 1$. It follows from theorem 1 that

$$\begin{aligned} & \lim_{n \rightarrow \infty} n \log F(a_n x + b_n) \\ &= \lim_{n \rightarrow \infty} -n \int_0^1 \lambda(u) [1 - H(a_n x + b_n | u)] du \\ &= - \lim_{n \rightarrow \infty} \left[\int_{A_1} \lambda(u) n [1 - H_1(a_n x + b_n)] du \right. \\ & \quad \left. + \int_{A_2} \lambda(u) n [1 - H_2(a_n x + b_n)] du \right] \\ &= - \int_{A_1} \lambda(u) du \left[\lim_{n \rightarrow \infty} n [1 - H_1(a_n x + b_n)] \right] \\ & \quad - \int_{A_2} \lambda(u) du \left[\lim_{n \rightarrow \infty} n [1 - H_2(a_n x + b_n)] \right] \\ &\approx - \int_{A_1} \lambda(u) du \left[\lim_{n \rightarrow \infty} n \log H_1(a_n x + b_n) \right] \\ & \quad - \int_{A_2} \lambda(u) du \left[\lim_{n \rightarrow \infty} n \log H_2(a_n x + b_n) \right] \quad (A7) \end{aligned}$$

APPENDIX

In this appendix we present proofs of theorem 1 and corollary 2. Before commencing with the proof of theorem 1 we need a definition and a lemma.

if both $H_1(a_n x + b_n) \rightarrow 1$ and $H_2(a_n x + b_n) \rightarrow 1$ (otherwise $F^n(a_n x + b_n)$ has a degenerate limit distribution). If both H_1 and H_2 are bounded with upper bounds $x_{H_1} > x_{H_2}$ then there

exists \bar{n} such that $H_2(a_n x + b_n) = 1$ for $n > \bar{n}$, from which the first assertion of corollary 2 follows. The second assertion follows from corollary 1.6.3 of Leadbetter et al. [1983].

Acknowledgments. This research was carried out at the Center for Mathematics and Computer Science in Amsterdam with the support of a Fulbright postdoctoral research grant.

REFERENCES

- Benson, M., Evolution of methods for evaluating the occurrence of floods, *U.S. Geol. Surv. Water Supply Pap.*, 1580-A, 30 pp., 1962.
- Bryson, M. C., Heavy-tailed distributions: Properties and tests, *Technometrics*, 16(1), 61-67, 1974.
- Cervantes, J. E., M. L. Kavvas, and J. W. Delleur, A cluster model for flood analysis, *Water Resour. Res.*, 19(1), 209-224, 1983.
- Cinlar, E., Superposition of point processes, in *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, edited by P. A. W. Lewis, pp. 549-606, Wiley-Interscience, New York, 1972.
- De Haan, L., Sample extremes: An elementary introduction, *Stat. Neerlandica*, 30(1), 161-172, 1976.
- DuMouchel, W. H., Estimating the stable index α in order to measure tail thickness: A critique, *Ann. Stat.*, 11(4), 1019-1031, 1983.
- Gumbel, E. J., *Statistics of Extremes*, Columbia University Press, New York, 1958.
- Gupta, V., L. Duckstein, and R. W. Peebles, On the joint distribution of the largest flood and its time of occurrence, *Water Resour. Res.*, 12(2), 295-304, 1976.
- Hosking, J. R. M., Testing whether the shape parameter is zero in the Generalized Extreme-Value distribution, *Biometrika*, 71, 367-374, 1984.
- Hosking, J. R. M., J. R. Wallis, and E. F. Wood, Estimation of the Generalized Extreme-Value distribution by the method of probability-weighted moments, *Technometrics*, 27(3), 231-262, 1985.
- Karr, A. F., Two extreme value processes arising in hydrology, *J. Appl. Probab.*, 13, 190-194, 1976.
- Karr, A. F., *Point Processes and Their Statistical Inference*, Marcel Dekker, New York, 1986.
- Leadbetter, M. R., G. Lindgren, and H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York, 1983.
- Leytham, K. M., Maximum likelihood estimates for the parameters of mixture distribution, *Water Resour. Res.*, 20(7), 896-902, 1984.
- Pickands, J., Statistical inference using extreme order statistics, *Ann. Stat.*, 3, 119-130, 1975.
- Prescott, P., and A. Walden, Maximum likelihood estimation of the parameters of the generalized extreme value distribution, *Biometrika*, 67, 723-724, 1980.
- Prescott, P., and A. Walden, Maximum likelihood estimation of the parameters of the three-parameter generalized extreme value distribution, *J. Stat. Comput. Simul.*, 16, 241-250, 1983.
- Serfling, R. J., *Approximation Theorems of Mathematical Statistics*, John Wiley, New York, 1980.
- Shen, H. W., M. C. Bryson, and I. D. Ochoa, Effect of tail behavior assumption on flood predictions, *Water Resour. Res.*, 16(2), 361-364, 1980.
- Smith, R. L., Threshold models for sample extremes, in *Statistical Extremes and Applications*, edited by J. Tiago de Oliveira, pp. 621-638, D. Reidel, Hingham, Mass., 1984.
- Smith, R. L., Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, 72, 67-90, 1985.
- Smith, J. A., and A. F. Karr, Flood frequency analysis using the Cox regression model, *Water Resour. Res.*, 22(6), 890-896, 1986.
- Todorovic, P., Stochastic models of floods, *Water Resour. Res.*, 14, 345-356, 1978.
- Todorovic, P., and E. Zelenhasic, A stochastic model for flood analysis, *Water Resour. Res.*, 6(6), 1641-1648, 1970.
- Waylen, P., and M. Woo, Prediction of annual floods generated by mixed processes, *Water Resour. Res.*, 18(4), 1283-1286, 1982.
- J. A. Smith, Interstate Commission on the Potomac River Basin, 6110 Executive Boulevard, Suite 300, Rockville, MD 20852.

(Received June 20, 1986;
revised April 9, 1987;
accepted April 10, 1987.)

