

# Flood Frequency Analysis Using the Cox Regression Model

JAMES A. SMITH

*Interstate Commission on the Potomac River Basin, Rockville, Maryland*

ALAN F. KARR

*Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland*

Procedures for incorporating time-varying exogenous information into flood frequency analyses are developed using the Cox regression model for counting processes. In this statistical model the probability of occurrence of a flood peak in a short interval  $[t, t + dt)$  depends in an explicit manner on the values at  $t$  of  $k$  "covariate" processes  $Z_1, \dots, Z_k$ . Specifically, letting  $dN(t)$  be 1 if a flood peak occurs in  $[t, t + dt)$  and 0 otherwise,  $dN(t) = a(t) \exp \{ \sum_{j=1}^k b_j Z_j(t) \} + dM(t)$  where  $a$ , the "baseline intensity," is an unknown function,  $b$  is a vector of unknown "regression" parameters, and the error  $dM(t)$  is (conditionally) orthogonal to the past history. Two applications, assessment of relative importance of physical processes such as snow melt or soil moisture storage on flood frequency at a site and derivation of time-varying flood frequency estimates, are considered.

## INTRODUCTION

The traditional approach to flood frequency analysis utilizes information contained in the series of annual peak discharges. It is recognized, however, that additional information is often available. In this paper we develop procedures, based on the Cox regression model for counting processes, for incorporating time-varying exogenous information into flood frequency analyses. In the model, occurrence of a flood peak in a short time interval  $[t, t + dt)$  depends on the current values of  $k$  "covariate" processes  $Z_1, \dots, Z_k$ . Letting  $dN(t)$  equal 1 if a flood peak occurs in  $[t, t + dt)$  and 0 otherwise, we have

$$dN(t) = a(t) \exp \{ b_1 Z_1(t) + \dots + b_k Z_k(t) \} + dM(t) \quad (1)$$

where  $a$  is an unknown function that represents a (seasonal) "baseline" intensity of flood peak occurrences,  $b$  is a vector of unknown "regression" parameters and the error  $dM(t)$  is, in a suitable conditional sense, orthogonal to the past history of the process. More precisely,

$$\lambda(t) = a(t) \exp \{ b_1 Z_1(t) + \dots + b_k Z_k(t) \} \quad (2)$$

is the stochastic intensity of the counting process  $N$  and the error process  $\{M(t)\}$  is a martingale. The stochastic intensity  $\{\lambda(t)\}$  will be interpreted as a conditional flood frequency process, whose value at time  $t$  is the conditional flood frequency given information on flood peaks and covariates until  $t$ .

We illustrate two types of applications. The Cox regression model can be used to assess the relative importance of specific processes, such as snow melt, soil moisture storage or frozen ground, on flood frequency at a site. In section 3 we develop hypothesis testing procedures for this purpose. The Cox regression model can also be used to provide time-varying flood frequency estimates; in section 5 we present a formulation of the "flood-warning problem" [Yakowitz, 1985] based on the Cox regression model.

The modeling framework is illustrated in Figure 1, which shows hydrograph of river discharge (with dashed line representing discharge of magnitude  $x$ ) (Figure 1a), times of flood

peaks exceeding  $x$  (Figure 1b), and time series of two covariate processes  $Z_1$  and  $Z_2$  that are presumed to affect the frequency of occurrence of floods (Figure 1c). Throughout, the time interval  $[0, 1]$  will represent a single year. Dependence of the flood peak model on the discharge threshold  $x$  will be suppressed in the notation. We illustrate in section 4 that explicit representation of discharge threshold can be achieved using "thinning" methods for counting processes.

Our approach to flood frequency analysis is based on partial duration series, that is, the sequence of all flood peaks above a specified threshold [see Shane and Lynn, 1964; Todovic and Zelenhasic, 1970; Karr, 1976; North, 1980; Cervantes et al., 1983; Smith, 1984] rather than annual peaks. Historically, preference for annual peaks has been based on the argument that additional information provided by partial duration series only concerns the central part of the flood frequency distribution [Benson, 1962]. This argument may fail if floods result from several distinct processes and exogenous information can be used to classify flood peaks. Analyses of annual peaks using mixture distributions [Leytham, 1984; Waylen and Woo, 1982] illustrate that classification of events can have a major impact on flood frequency estimates at all quantiles. One perspective for viewing the flood frequency model we develop is that it provides a partial duration series analog to mixture distributions. It should also be pointed out that for "operational" problems such as the flood warning problem of section 5 the central part of the flood frequency distribution can be quite important.

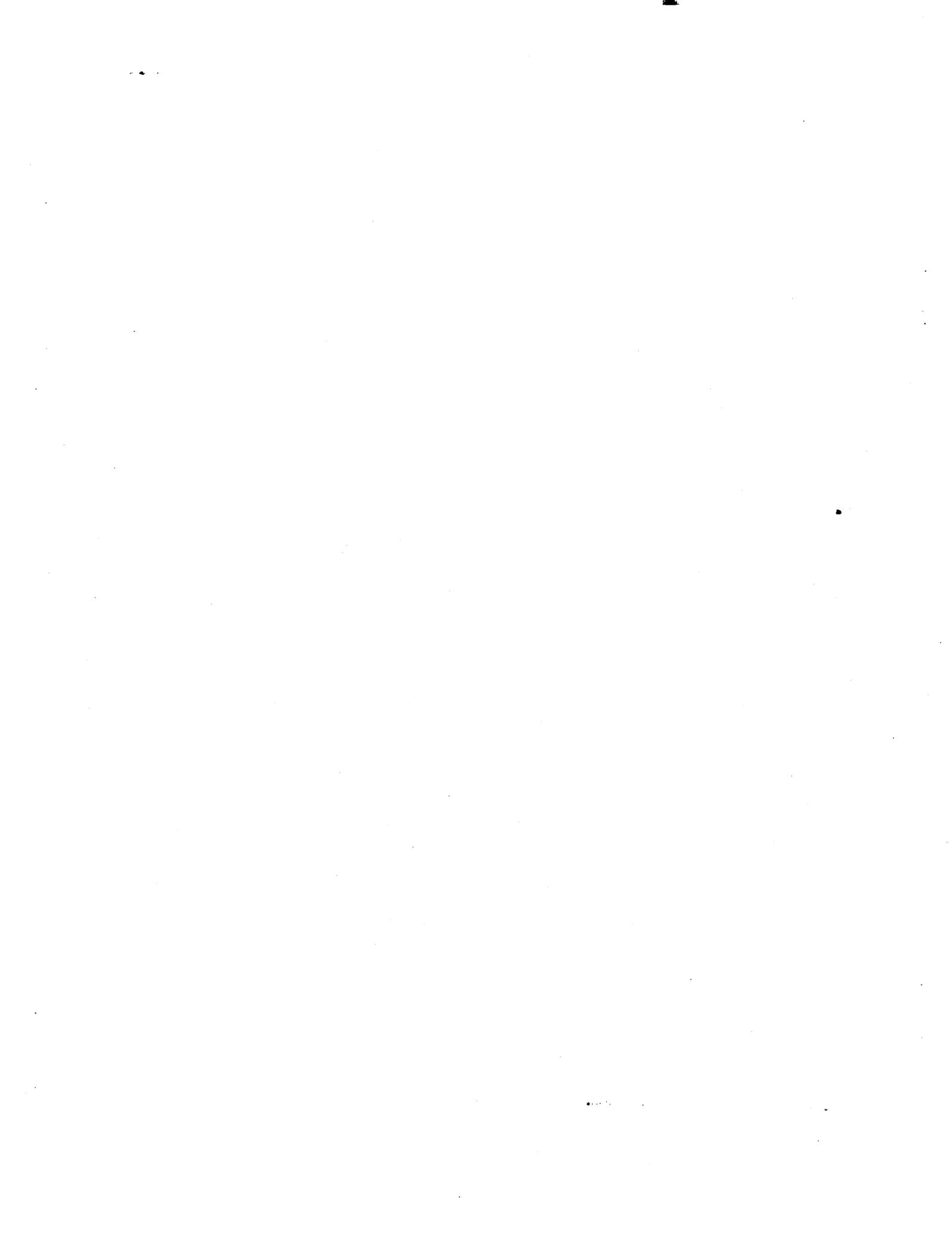
The Cox regression model was originally developed for analysis of survival times [Cox, 1972] and has been widely applied in medical and industrial lifetesting problems; a literature review containing numerous applications is given in Cox and Oakes [1984]. The formulation we use is based on the counting process development of Andersen and Gill [1982], for which Gill [1984] and Karr [1986] are the principal expository treatments.

## Definitions and Notation

The times of occurrence of flood peaks, that is, exceedances of a discharge level  $x$ , during a year will be modeled as a counting process (point process) on the interval  $[0, 1]$ . Thus time 0 corresponds to the beginning of the year (which we

Copyright 1986 by the American Geophysical Union.

Paper number 6W4420.  
0043-1397/86/006W-4420\$05.00



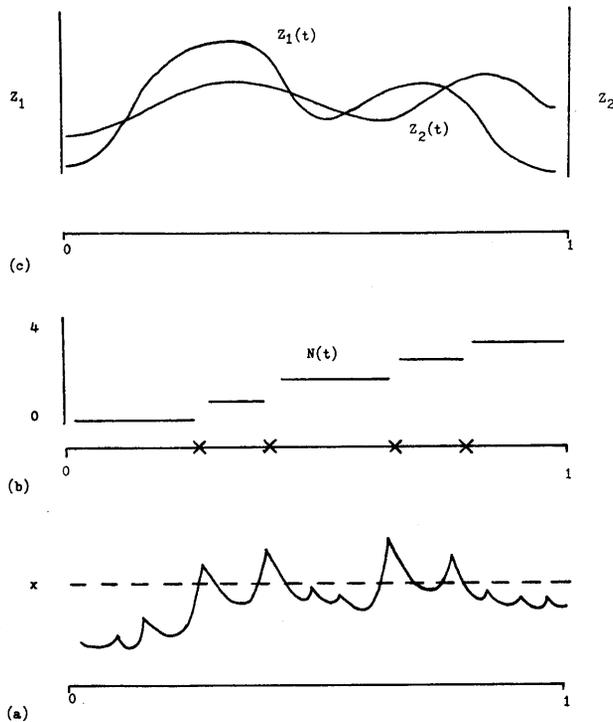


Fig. 1. Illustration of model components: (a) river discharge, (b) counting process for flood peaks of magnitude greater than  $x$ , and (c) covariate processes.

take to be October 1) and time 1 corresponds to the end of the year (September 30). Denote by  $N(1)$  the total number of flood peaks during the year and for  $N(1) > 0$  denote the occurrence times by  $T(1), \dots, T(N(1)) \in [0, 1]$ . The counting process  $\{N(t), t \in [0, 1]\}$  is defined by

$$\begin{aligned} N(t) &= 0 & N(1) &= 0 & \text{or } t &< T(1) \\ N(t) &= n & T(n) &\leq t < T(n+1) \\ N(t) &= N(1) & t &\geq T(N(1)) \end{aligned} \quad (3)$$

The history of  $N$  at time  $t$  is, heuristically, the information about the process that has accumulated until  $t$ . It includes the occurrence times of all flood peaks before  $t$ , as well as any exogenous information that has been obtained. We assume that relevant exogenous information is contained in  $k$  (left-continuous) random processes  $\{Z_1(t), \dots, Z_k(t), t \in [0, 1]\}$ , which, in the terminology of the Cox regression model, are called covariates. (In survival analysis the covariates represent additional observable characteristics of a patient, such as age or weight.) Formally, we define the history  $\{H_t, t \in [0, 1]\}$  of the flood occurrence process by

$$H_t = \sigma\{N(u), Z_j(u); j = 1, \dots, k, u < t\} \quad (4)$$

that is,  $H_t$  is the  $\sigma$ -algebra generated by  $\{N(u), u < t\}$  and  $\{Z_j(u); j = 1, \dots, k, u < t\}$  [see Karr, 1986]. The history  $H_t$  can be viewed as shorthand notation for use in conditional expectations; thus for a random variable  $X$ ,

$$E[X|H_t] = E[X|N(u), Z_j(u); j = 1, \dots, k, u < t] \quad (5)$$

Properties of the error process  $\{M(t)\}$  in (1) play a central role in inference procedures for the flood frequency model. By analogy with classical regression theory we would like error terms for successive time increments to be independent and

identically normally distributed. From the jump process representation of  $\{M(t)\}$  (equation (1)), it is clear that errors cannot be normally distributed. While we must abandon hope for IID normal residuals, we do not, however, have to relax our goals too much (as will be shown below), due to the fact that  $\{M(t)\}$  is a martingale. An  $H$ -martingale is a right-continuous, (adapted) process  $M$  satisfying the martingale equality

$$E[M(t)|H_s] = M(s) \quad (6)$$

for all  $s < t \in [0, 1]$ . The term "adapted" means that for each  $t$ , the value of the process at time  $t$  is a function of the random variables comprising the history of  $N$  and  $Z$  at time  $t$ . An equivalent form of (6) is

$$E[M(t) - M(s)|H_s] = 0$$

which shows that each increment in the error process is conditionally orthogonal to the past history of the process; this can be viewed as a weakened version of the classical situation.

The stochastic intensity of a point process  $N$  is a left-continuous, adapted process  $\{\lambda(t)\}$  such that

$$M(t) = N(t) - \int_0^t \lambda(u) du \quad (7)$$

is an  $H$ -martingale. It is important to note that the stochastic intensity depends on the history  $H$ ; as illustrated below,  $\lambda(t)$  can be interpreted as the conditional rate of occurrence of events of  $N$ , given the information  $H_t$ .

Our flood frequency model is specified to have the stochastic intensity

$$\lambda(t) = a(t) \exp \langle b, Z(t) \rangle \quad (8)$$

where we use the inner product notation

$$\langle b, Z(t) \rangle = b_1 Z_1(t) + \dots + b_k Z_k(t) \quad (9)$$

The model reduces to a nonstationary Poisson process with intensity function  $a(t)$  if all of the  $b$ 's equal zero, and to a stationary Poisson process if, additionally,  $a(t)$  is constant.

Interpretation of  $\lambda$  as a flood frequency follows from the fact that if  $\lambda$  is the stochastic intensity of  $N$  then

$$\lambda(t) = \lim_{s \downarrow 0} (1/s) P\{N(t+s) - N(t) \geq 1 | H_t\} \quad (10)$$

[see Karr, 1986]; that is,  $\lambda(t) dt$  is the conditional probability of a flood occurring during  $[t, t + dt]$  and  $\lambda(t)^{-1}$  is the "recurrence interval" (for a flood of magnitude  $x$ ) at time  $t$ .

Statistical inference for the flood frequency model will be based on  $n$  years of flood peak data, which will be denoted  $\{N^i(t), t \in [0, 1], i = 1, \dots, n\}$  and  $n$  years of corresponding data for the covariate processes  $\{Z_j^i(t); t \in [0, 1], j = 1, \dots, k; i = 1, \dots, n\}$ . We assume that the annual processes  $(N^i, Z^i)$  are independent and identically distributed. For fixed,  $i$ ,  $N^i$  and  $Z^i$  are not independent; indeed a key property of the model is that  $N^i$  depends in a specific manner on  $Z^i$ . Validity of the independent and identically distributed (IID) assumption for the example presented in section 4, in which snow pack and soil moisture storage are the two covariate processes, depends heavily on choosing the time interval  $[0, 1]$  to represent a "water year" (October–September) rather than a calendar year. The following three points are of particular importance: (1) late summer is the period of lowest flood frequency (see Figure 2); (2) late summer is typically the period of maximum soil moisture depletion; and (3) snow occurs only during the months October–April.

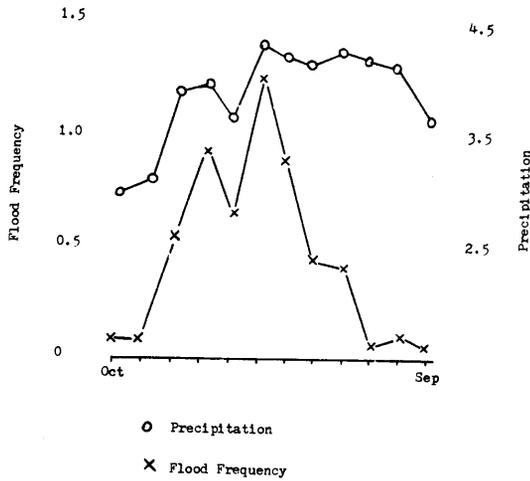


Fig. 2. Precipitation (in inches per month) and flood frequency (in events per month) for the North Branch Potomac River basin (one inch = 2.54 cm).

PARAMETER ESTIMATION AND HYPOTHESIS TESTING

The utility of specifying the flood frequency model in terms of its stochastic intensity is due in large part to the physical interpretation of the stochastic intensity and to availability of inference procedures. In particular, the likelihood function for a point process can be expressed in terms of the stochastic intensity. We have the following expression for the log-likelihood function given independent observations over  $n$  years [Karr, 1986]

$$L(\theta) = \sum_{i=1}^n \int_0^1 [1 - \lambda^i(u)] du + \sum_{i=1}^n \int_0^1 \log(\lambda^i(u)) dN^i(u) \\ = \sum_{i=1}^n \int_0^1 [1 - \lambda^i(u)] du + \sum_{i=1}^n \sum_{j=1}^{N^i(1)} \log(\lambda^i(T^i(j))) \quad (11)$$

where  $\theta$  is the vector of unknown "parameters" of  $N$ , some of which, as in our model, may be functions rather than finite-dimensional parameters. Likelihood-based inference for point process models of rainfall are described in the work by Smith and Karr [1985].

Direct analysis of the Cox regression model using maximum likelihood estimation is impossible. Because the baseline intensity function  $a$  is entirely unknown the full log-likelihood function (omitting a constant term)

$$L(a, b) = - \sum_{i=1}^n \int_0^1 a(u) \exp\{\langle b, Z^i(u) \rangle\} du \\ + \sum_{i=1}^n \int_0^1 \log(a(u)) dN^i(u) \\ + \sum_{i=1}^n \int_0^1 \langle b, Z^i(u) \rangle dN^i(u) \quad (12)$$

is not bounded above, and hence (joint) maximum likelihood estimators of  $a$  and  $b$  do not exist. Partly because of this, partly because the regression parameter  $b$  is often of paramount interest, and on the rationale that differences among covariates are manifested only through  $b$ , Cox proposed that  $b$  be estimated by maximizing the "partial (log-) likelihood" function

$$C(b) = \sum_{i=1}^n \int_0^1 \langle b, Z^i(u) \rangle dN^i(u) \\ - \int_0^1 \log \left[ \sum_{i=1}^n \exp\{\langle b, Z^i(u) \rangle\} \right] d\bar{N}(u) \quad (13)$$

which does not depend on  $a$ . In (13),  $\bar{N}$  is the superposition (sum) of  $N^1, \dots, N^n$ , that is,

$$\bar{N}(t) = \sum_{i=1}^n N^i(t) \quad (14)$$

Thus  $C(b)$  in effect compares individual covariate effects to the overall effect of the covariates on the superposition  $\bar{N}$ . The Cox estimator of  $b$  is any solution  $\hat{b}$  to the likelihood equation

$$\nabla C(b) = 0 \quad (15)$$

where " $\nabla$ " denotes gradient. For additional discussion and interpretation see Cox [1975], Johansen [1983], and Karr [1986].

Andersen and Gill [1982] show that partial likelihood estimators have asymptotic properties similar to those of ordinary maximum likelihood estimators. Of particular importance are the following results, which are valid under suitable technical restrictions [see Andersen and Gill, 1982]. Let  $b_0$  denote the "true value" of the regression parameter  $b$ .

Consistency

$$\hat{b} \xrightarrow{P} b_0 \quad (16)$$

Asymptotic normality

$$n^{1/2}(\hat{b} - b_0) \xrightarrow{D} N(0, \Lambda^{-1}) \quad (17)$$

the matrix  $\Lambda$  can be calculated explicitly (we omit the result) and, more importantly, can be estimated from the observations.

Consistency of estimators of asymptotic covariance matrix

$$(1/n)I(\hat{b}) \xrightarrow{P} \Lambda \quad (18)$$

where

$$I(b) = \int_0^1 \left[ \frac{\sum_{i=1}^n Z^i(s)^{\otimes 2} \exp(\langle b, Z^i(s) \rangle)}{\sum_{i=1}^n \exp(\langle b, Z^i(s) \rangle)} - \left( \frac{\sum_{i=1}^n Z^i(s) \exp(\langle b, Z^i(s) \rangle)}{\sum_{i=1}^n \exp(\langle b, Z^i(s) \rangle)} \right)^{\otimes 2} \right] d\bar{N}(s) \quad (19)$$

For a vector  $z = (z(1), \dots, z(k))$  the notation  $z^{\otimes 2}$  denotes the  $k \times k$  matrix whose  $(i, j)$  entry is  $z(i)z(j)$ .

For estimation of the integrated baseline intensity function

$$A(t) = \int_0^t a(u) du \quad (20)$$

one uses martingale estimators [see Karr, 1984; 1986]

$$\hat{A}(t) = \int_0^t \left[ \sum_{i=1}^n \exp(\langle \hat{b}, Z^i(u) \rangle) \right]^{-1} d\bar{N}(u) \quad (21)$$

where  $\hat{b}$  is the estimator given by (15).

An important feature of the partial likelihood method is that inference for the regression coefficients  $b$  does not require parametric assumptions concerning the baseline intensity  $a(u)$ . Explicit estimators of  $a(u)$  (using either equation (21) or (12) for standard maximum likelihood estimation) may be obtained for certain specified parametric forms of the baseline intensity. For example, one may assume that the baseline intensity is constant over the entire year or that the baseline intensity assumes constant values for seasons. A serious problem with standard likelihood-based inference procedures (especially for the hypothesis testing problems described below) is that it is difficult to eliminate the effect of (improper) parametric assumptions for the baseline intensity on inferences concerning the regression parameters. Consequently the non-parametric approach represented by (21) is often preferred.

Inference procedures based on partial likelihood also possess significant computational advantages over maximum likelihood methods. Numerical optimization techniques must be used for either maximum likelihood or partial likelihood parameter estimation. Note, however, that the partial likelihood function can be expressed in terms of the covariates evaluated only at the times of flood peaks, whereas the maximum likelihood method, even under parametric assumptions regarding  $a$ , requires numerical integration of the covariate functions over the interval  $[0, 1]$  (the first term of equation (12)), and in particular, knowledge of values of the covariates at every time  $t$ .

Formal significance tests for the Cox regression model can be constructed from partial likelihood ratios

$$\Lambda_n = -2[C(b_0) - C(\hat{b})] \quad (22)$$

where  $\hat{b}$  is the partial likelihood estimator and  $b_0$  is the (hypothesized) true parameter of the model. Rejection levels can be calculated using the following proposition.

Proposition: under the hypothesis that  $b = (b_1, \dots, b_k) = b_0$ , and assuming that the estimators  $\hat{b}$  are asymptotically normal,

$$\Lambda_n = -2[C(b_0) - C(\hat{b})] \quad (23)$$

converges in distribution to a  $\chi^2$  random variable with  $k$  degrees of freedom. The proof is sketched in the appendix.

#### ILLUSTRATION OF INFERENCE PROCEDURES

In this section parameter estimation and hypothesis testing procedures for the flood frequency model are applied to a 240 square mile catchment located on the Appalachian Plateau in western Maryland. The covariate processes we consider are soil moisture storage and snow pack.

The North Branch of the Potomac River receives an average of 50 inches of precipitation annually, of which approximately 10% is in the form of snow. Figure 2 shows mean monthly precipitation together with the monthly frequency of floods greater than 2000 cubic feet per second (cfs). The sharp winter-spring peak in flood frequency, together with the absence of comparable seasonal contrasts in precipitation, suggests that soil moisture storage and snow melt are important processes in flood frequency for the North Branch. It is natural to ask whether the role of soil moisture (snow pack) results simply from the annual cycle of evapotranspiration demand (temperature), without change from year to year, or whether flood frequency varies from year to year in response to random fluctuations of soil moisture and snow pack. We illustrate below how the Cox regression model can be used to examine these issues.

Direct measurements of soil moisture or snow pack are not widely available, but surrogates for these variables can be obtained from readily available precipitation and temperature data. The soil moisture data we use for flood frequency analyses are obtained from the Sacramento soil moisture accounting model [Burnash *et al.*, 1973], while our snow pack data are obtained using the National Weather Service (NWS) Snow Accumulation and Ablation model [Anderson, 1973].

The Sacramento model is a rainfall runoff model that routes rainfall and snow melt through a series of conceptual storage zones, with the ultimate destination being either channel inflow or evapotranspiration. Six-hourly rain plus melt data for the North Branch were used to produce a 26 year record of daily soil moisture storage. The covariate value on day  $t$  for soil moisture storage was taken to be the value of lower zone free primary storage (see Burnash *et al.* [1973] for definitions) at the beginning of day  $t - 1$ .

The NWS snow model accounts for a range of processes including areal extent of snow cover, energy balance of the snow pack, and total water equivalent of the snow pack. Six-hourly records of mean areal precipitation and temperature for the North Branch were used to create a 26 year record of daily total water equivalent. The covariate value on day  $t$  was taken to be the average total water equivalent over the preceding ten days. Over the period of record the months for which nonzero values for the snow variable occur are October-April.

Table 1 shows estimates of the regression coefficients of a model for which soil moisture storage is covariate 1 and snow pack is covariate 2. Estimates are presented for flood thresholds of 2000, 3000, and 4000 cfs. Standard deviation and correlation of the estimators, obtained from (17) and (18), are also presented in Table 1. For each of the regression coefficients, the standard deviation is small compared with the value of the estimator, confirming that the covariates do play a significant role in the flood process. Note also that correlation between the estimators is small.

The baseline flood frequency  $a(t)$  was parameterized to assume a constant value  $a(1)$  in October and November,  $a(2)$  in December and January, etc. Estimates of the baseline intensity are presented in Table 2. Note that the largest values occur in summer, the period of lowest flood frequency. The form of the estimated baseline intensity (especially, the summer peak in baseline intensity) reflects both seasonal contrasts in the covariates and seasonal contrasts in flood producing storms. High rainfall intensity storms are most common in the summer season. The largest flood during the period of record occurred in September of 1955 following hurricane Dianne. The results suggest that during the summer season the effects of soil moisture storage on the flood process are subordinate to the variability in rainfall intensity of summer season storms, resulting in "large" values of baseline intensity relative to the covariate term.

TABLE 1. Estimates of Regression Coefficients

	Number of Events	$\hat{b}_1$	Var ( $\hat{b}_1$ )	$\hat{b}_2$	Var ( $\hat{b}_2$ )	Cor ( $\hat{b}_1, \hat{b}_2$ )
Threshold						
2000	133	0.21	0.02	0.019	0.004	-0.07
3000	69	0.25	0.03	0.021	0.005	-0.08
4000	35	0.25	0.04	0.026	0.006	-0.07

Threshold given in cubic feet per second.

TABLE 2. Estimates of Baseline Intensity Parameters for Increasing Discharge Threshold

	Parameters					
	$\hat{a}(1)$	$\hat{a}(2)$	$\hat{a}(3)$	$\hat{a}(4)$	$\hat{a}(5)$	$\hat{a}(6)$
2000	33.1	5.3	2.5	2.7	4.3	20.2
3000	7.4	0.5	0.2	0.3	0.4	2.1
4000	2.6	0.2	0.1	0.1	0.1	2.1

Threshold given in cubic feet per second; units in  $10^{-4}$  events per year.

An interesting feature of the estimation results is that estimates of the regression coefficients increase slightly with increasing discharge threshold, so that even though the overall number of events is sharply decreasing, the flood frequency component attributed to snow pack and soil moisture remains virtually constant or increases slightly. The following discussion sheds light on this result.

Let  $N$  be a point process with history  $\{H_t\}$  and let  $\{Y(i), i = 1, 2, \dots\}$  be IID random variables independent of  $\{H_t\}$ , with

$$P\{Y(i) = 1\} = p = 1 - P\{Y(i) = 0\} \quad (24)$$

For each  $t$  let

$$\tilde{N}(t) = \sum_{i=1}^{N(t)} Y(i) \quad (25)$$

thus  $\tilde{N}$  is obtained from  $N$  by randomly deleting events from the original process. In particular, each event of  $N$  is deleted with probability  $1 - p$ , independently of all other events. The process  $\tilde{N}$  is termed a  $p$ -thinning of  $N$ .

In our setting, the occurrence process for flood peaks larger than 3000 cfs can be expressed as

$$\tilde{N}(t) = \sum_{i=1}^{N(t)} Y(i) \quad (26)$$

where  $Y(i) = 1(X(i) > 3000)$ , with  $X(i)$  the magnitude of the  $i$ th event, and where  $N$  represents flood peaks larger than 2000 cfs.

If (1) the magnitudes  $X(i)$  of flood peaks are IID, (2) the magnitudes of flood peaks are independent of the occurrence process  $N$  and the covariates  $Z_1$  and  $Z_2$ , and (3)  $N$  has stochastic intensity

$$\lambda(t) = a(t) \exp \{ \langle b, Z(t) \rangle \} \quad (27)$$

then [see Kallenberg, 1983; Karr, 1986]  $\tilde{N}$  is a  $p$ -thinning of  $N$  and its stochastic intensity is

$$\lambda(t) = pa(t) \exp \{ \langle b, Z(t) \rangle \} \quad (28)$$

where

$$p = P\{X(i) > 3000\} \quad (29)$$

Thus if conditions 1 and 2 above hold, regression coefficients do not vary with increasing threshold values. Conversely, if regression coefficients are not constant with thinning, explanation can be based on violation of conditions 1 or 2 (or both). Regression coefficients should increase with thinning if larger flood peaks are more closely related to covariate processes, and decrease if thinning produces events with decreasing dependency on the covariates. Results of Table 1 suggest that over the range of thinning thresholds examined influence of the covariates on flood frequency remains virtually constant.

Table 3 contains results of partial likelihood ratio tests of significance for soil moisture storage and snow pack. A partial likelihood ratio test for model significance is given by  $\Lambda(0, 0)$ , which under the null hypothesis has a  $\chi^2$  distribution with 2 degrees of freedom. A test of the significance of soil moisture given that snow pack is included in the model is given by  $\lambda(0, \hat{b}_2)$ , which under the null hypothesis has a  $\chi^2$  distribution with one degree of freedom. Similarly,  $\Lambda(\hat{b}_1, 0)$  provides a test of the significance of snow pack given that soil moisture is included in the model.

All of the test statistics are highly significant. Thus in this example, the partial likelihood ratio tests merely substantiate conclusions based on the covariance results. Decreasing likelihood ratios with thinning reflect the decline in total number of events from 139 at 2000 cfs to 35 at 4000 cfs; a corresponding increase can be noted in standard deviation of regression coefficient estimators in Table 1. Note, for example, that while estimates of  $b_1$  increase from 0.21 to 0.25 with thinning, the number of standard deviations from 0 decreases from 10 to 6. The latter is the germane point for likelihood ratio tests. It should also be noted that the likelihood ratio tests provide a qualitative measure of the relative importance of snow pack and soil moisture storage; this information is much more difficult to derive from covariance results.

#### THE FLOOD WARNING PROBLEM

In this section we formulate a version of the flood warning problem of Yakowitz [1985] based on the Cox regression model. Our development is motivated by a reservoir operation problem on the North Branch of the Potomac River; the formulation presented below, however, is a highly simplified version of the actual operating problem.

The essentials of the flood-warning problem are as follows. Large releases from a reservoir must be made from the lowest port of a multiport release system. Water quality at the flood port level is poor and large releases from this port will decimate a downstream fishery. The operations problem is to allocate flood control storage based on time varying estimates of flood risk so as to minimize the expected loss from fishery damage and excess release. This is accomplished by maintaining the flood pool at level L1 when flood risk is low and lowering the flood pool to level L2 when flood risk is high. When the flood pool is at level L1 floods above magnitude  $x$  require releases which destroy the fishery. "All" floods can be managed without damaging the fishery when flood pool is at level L2. Flood control operations only concern the nondraw-down period November-April, which we now take to be the time interval  $[0, 1]$ .

To replace a destroyed fishery, which is necessary when a flood occurs, costs A dollars, while the cost per time unit of operating the reservoir at the reduced level L2 is B dollars. Decisions to raise or lower the flood pool (for simplicity we assume that changes can be effected instantaneously) are based on estimates of the probability that a flood of magnitude greater than  $x$  will occur in the immediate future. These

TABLE 3. Partial Likelihood Ratio Test Results

	$\Lambda(0, 0)$	$\Lambda(0, \hat{b}_2)$	$\Lambda(\hat{b}_1, 0)$
2000	169.	156.	15.
3000	130.	104.	14.
4000	78.	52.	13.

Threshold given in cubic feet per second.

estimates are based in turn on previous observations of the flood peaks and covariate processes. In the context of our Cox regression model for flood peaks, provided that the baseline intensity function  $a$  and regression parameter  $b$  are known (more on this point below), this probability can be approximated as

$$P\{N(t + dt) - N(t) \geq 1 | H_t\} \cong a(t) \exp \{ \langle b, Z(t) \rangle \} dt \quad (30)$$

The simplest kind of operating policy has the form: maintain reservoir level L1 when the stochastic intensity  $\lambda(t) = a(t) \exp \{ \langle b, Z(t) \rangle \}$  falls below a threshold value  $y$  (this is equivalent to the predicted flood probability given by (30) being below a different threshold), and reduce the level to L2 when the stochastic intensity exceeds  $y$ . The random loss incurred by following this policy over the course of one November-April period is

$$L(y) = A \int_0^1 1(\lambda(u) \leq y) dN(u) + B \int_0^1 1(\lambda(u) > y) du \quad (31)$$

In (31) the first term is  $A$  times the number of flood peaks that occur when  $\lambda(u) \leq y$  and the flood pool is hence at L1, and represents the cost of destroyed fisheries; the second is  $B$  times the length of time during which the level is maintained at L2.

Suppose now that we wish to choose  $y$  to minimize the expected cost

$$\begin{aligned} C(y) &= E[L(y)] \\ &= AE \left[ \int_0^1 1(\lambda(u) \leq y) dN(u) \right] + BE \left[ \int_0^1 1(\lambda(u) > y) du \right] \end{aligned} \quad (32)$$

To simplify the analysis we stipulate that for each  $u$  the random variable  $\lambda(u) = a(u) \exp \{ \langle b, Z(u) \rangle \}$  admits a continuous density function  $f_u(x)$ :

$$P\{\lambda(u) \leq z\} = \int_0^z f_u(v) dv \quad (33)$$

This amounts to a corresponding assumption on the random vector  $Z(u)$  and does not seem unreasonable physically. Because  $\lambda$  is the stochastic intensity of  $N$  (and because the process  $1(\lambda(u) \leq y)$  is predictable; this point is technical but absolutely essential, and is discussed in Karr [1986]),

$$\begin{aligned} C(y) &= AE \left[ \int_0^1 1(\lambda(u) \leq y) \lambda(u) du \right] \\ &\quad + BE \left[ \int_0^1 1(\lambda(u) > y) du \right] \\ &= AE \left[ \int_0^1 1(\lambda(u) \leq y) \lambda(u) du \right] \\ &\quad + BE \left[ \int_0^1 \{1 - 1(\lambda(u) \leq y)\} du \right] \\ &= \int_0^1 E[(A\lambda(u) - B)1(\lambda(u) \leq y)] du + B \\ &= \int_0^1 \left( \int_0^y (Az - B)f_u(z) dz \right) du + B \end{aligned} \quad (34)$$

To minimize  $C$  it suffices to minimize the first term, which can be done by calculus. We have

$$C'(y) = \int_0^1 (Ay - B)f_u(y) du \quad (35)$$

which is zero for  $y = B/A$ , and it is easily checked that  $C''(B/A) > 0$ , so that  $y^* = B/A$  is the unique value of the threshold that minimizes expected total costs.

Note that the value of  $y^*$  does not depend on either the baseline intensity function  $a$  or the regression parameter  $b$ , which is particularly useful in practice since these are not known. Thus  $y^*$  can be determined even when  $a$  and  $b$  are unknown. Nevertheless, issues of implementing the policy remain, because the stochastic intensity is not observable; only the covariate process  $Z$  is observable, and since  $a$  and  $b$  are unknown, one cannot calculate from  $Z(u)$  the value  $\lambda(u) = a(u) \exp \{ \langle b, Z(u) \rangle \}$  needed for comparison to the threshold. In the formulation of Karr [1986] this becomes a problem of combined statistical inference and state estimation. State estimation is the optimal prediction of unobserved random variables; nearly always the predictors, as in (30), are conditional expectations of the unobservable random variables given the observations. But also, as (30) makes explicit, computation of conditional expectations usually entails knowledge of whatever "parameters" comprise the statistical model. In practice and in theory this difficulty is addressed by replacing unknown parameters by estimators of them, yielding in our case "pseudo-" state estimators

$$\hat{P}\{N(t + dt) - N(t) \geq 1 | H_t\} = \hat{a}(t) \exp \{ \langle \hat{b}, Z(t) \rangle \} dt \quad (36)$$

(Recall that  $Z(t)$  is observable.) In (36),  $\hat{a}$  and  $\hat{b}$  are estimators of  $a$  and  $b$ , respectively, derived from previous years' data using methods described in section 3. By virtue of consistency, given large data sets these estimators are close to the "true" values  $a$  and  $b$ , and hence  $\hat{P}\{N(t + dt) - N(t) \geq 1 | H_t\}$  does not differ significantly from  $P\{N(t + dt) - N(t) \geq 1 | H_t\}$ .

In terms of the flood warning policy, one would define the "pseudo-" stochastic intensity

$$\hat{\lambda}(t) = \hat{a}(t) \exp \{ \langle \hat{b}, Z(t) \rangle \} \quad (37)$$

and follow the policy "operate at L1 if  $\hat{\lambda}(t) \leq y^* = B/A$  and at L2 if  $\hat{\lambda}(t) > B/A$ ." For practical purposes this policy has the same optimality properties as the policy based on the stochastic intensity  $\lambda(t)$ .

The problem described above is representative of a range of water resources problems for which the primary goals are to characterize the frequency of occurrence of water quality problems related to rare hydrologic events, and to design control and abatement strategies which utilize time varying information pertaining to hydrologic response.

#### SUMMARY AND CONCLUSIONS

The Cox regression model provides a flexible tool for incorporating time-varying exogenous information pertaining to processes such as soil moisture storage, snow pack, and frozen ground into flood frequency analyses. The flood frequency model we develop based on the Cox regression model simultaneously generalizes partial duration series models [Shane and Lynn, 1966; Todorovic, 1972] and mixture distribution models [Leytham, 1984; Waylen and Woo, 1982].

Two applications of the model are illustrated. The model can be used in a hypothesis testing framework to assess the

importance of specific processes on flood frequency at a site. Partial likelihood inference procedures for this purpose are presented in section 3. Specifically, the estimation procedures of *Andersen and Gill* [1982] are presented and a partial likelihood ratio test is developed for assessing the significance of "covariates." Inference procedures are applied to a 240 square mile catchment on the Appalachian Plateau. In this example the covariate processes are snow pack and soil moisture storage.

The Cox regression model can also be used to provide time-varying flood frequency estimates for water resource management problems. In section 5 a formulation of the "flood warning problem" of *Yakowitz* [1985] is presented based on the Cox regression model. In this example, operation of flood control storage of a reservoir is tied to time-varying estimates of flood risk.

#### APPENDIX

The proof of (23) is sketched below. A Taylor series expansion of the log partial likelihood function  $C(b)$  about  $\hat{b}$  yields

$$C(b) = C(\hat{b}) + (b - \hat{b})\nabla C(\hat{b}) - (1/2)(b - \hat{b})I(b^*)(b - \hat{b})^T \quad (A1)$$

where  $b^*$  is on the line segment between  $b$  and  $\hat{b}$ .

Evaluating (A1) at the true parameter  $b_0$  yields

$$C(b_0) = C(\hat{b}) - (1/2)(b_0 - \hat{b})I(b^*)(b_0 - \hat{b})^T \quad (A2)$$

noting that  $\nabla C(\hat{b}) = 0$  by definition of the partial likelihood estimator. Rearranging terms yields

$$\begin{aligned} & -2(C(b_0) - C(\hat{b})) \\ & = n^{1/2}(b_0 - \hat{b})[(1/n)I(b^*)]n^{1/2}(b_0 - \hat{b})^T \end{aligned} \quad (A3)$$

The result follows from consistency of the partial likelihood estimator (16), consistency of the estimator of the asymptotic covariance matrix (18), and asymptotic normality of partial likelihood estimators (17).

#### REFERENCES

- Anderson, E. A., National weather service river forecast system-snow accumulation and ablation model, *NOAA Tech. Memo. NWS HYDRO-17*, 217 pp., Silver Spring, Md., 1973.
- Andersen, P. K., and R. D. Gill, Cox's regression model for counting processes: A large sample study, *Ann. Stat.*, 10(4), 1100-1120, 1982.
- Benson, M. A., Evolution of methods for evaluating the occurrence of floods, *U.S. Geol. Surv. Water Supply Pap.*, 1580-E, 56 pp., 1965.
- Burnash, R. J. C., R. L. Ferral, and R. A. McGuire, A generalized streamflow simulation system, technical report, Joint Fed. State River Forecast Cent., U.S. Nat. Weather Serv. and Calif. Dep. of Water Resour., Sacramento, Calif., 1973.
- Cervantes, J. E., M. L. Kavvas, and J. W. Delleur, A cluster model for flood analysis, *Water Resour. Res.*, 19(1), 209-224, 1983.
- Cox, D. R., Regression models and life tables (with discussion), *J. R. Stat. Soc. Ser. B*, 34, 187-220, 1972.
- Cox, D. R., Partial likelihood, *Biometrika*, 62, 269-276, 1975.
- Cox, D. R., and D. Oakes, *Analysis of Survival Times*, Methuen, London, 1984.
- Gill, R. D., Understanding Cox's regression model: A martingale approach, *J. Am. Stat. Assoc.*, 79, 441-447, 1984.
- Johansen, S., An extension of Cox's regression model, *Int. Stat. Rev.*, 51, 165-174, 1983.
- Kallenberg, O., *Random Measures*, Academic, Orlando, Fla., 1983.
- Karr, A. F., Two extreme value processes arising in hydrology, *J. Appl. Probab.*, 13, 190-194, 1976.
- Karr, A. F., The martingale method: Introductory sketch and access to the literature, *Oper. Res. Lett.*, 3(2), 59-63, 1984.
- Karr, A. F., *Point Processes and their Statistical Inference*, Dekker, New York, 1986.
- Leytham, K. M., Maximum likelihood estimates for the parameters of mixture distributions, *Water Resour. Res.*, 20(7), 896-902, 1984.
- North, M., Time dependent stochastic models of floods, *J. Hydraul. Div. Am. Soc. Civ. Eng.*, 106, 649-655, 1980.
- Shane, R. M., and W. R. Lynn, Mathematical model for flood risk evaluation, *J. Hydraul. Div. Am. Soc. Civ. Eng.*, 90(HY6), 1-20, 1964.
- Smith, J. A., and A. F. Karr, Statistical inference for point process models of rainfall, *Water Resour. Res.*, 21(1), 73-79, 1985.
- Smith, R. L., Threshold methods for sample extremes, in *Statistical Extremes and Applications*, edited by J. Tiago de Oliveira, D. Reidel, Hingham, Mass., 1984.
- 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.
- Yakowitz, S., Markov flow models and the flood warning problem, *Water Resour. Res.*, 21(1), 81-88, 1985.
- A. F. Karr, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218.
- J. A. Smith, Interstate Commission on the Potomac River Basin, 6110 Executive Boulevard, Suite 300, Rockville, MD 20852.

(Received January 22, 1986;  
accepted January 28, 1986.)