Statistical Science
1989, Vol. 4, No. 4, 367–393

Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone

Richard L. Smith

Abstract. Several methods of analyzing extreme values are now known, most based on the extreme value limit distributions or related families. This paper reviews these techniques and proposes some extensions based on the point-process view of high-level exceedances. These ideas are illustrated with a detailed analysis of ozone data collected in Houston, Texas. There is particular interest in whether they is any trend in the data. The analysis reveals no trend in the overall levels of the series, but a marked downward trend in the extreme values.

Key words and phrases: Atmospheric pollution, clustered point processes, environmental time series, exceedances, extreme values, Generalized Extreme Value distribution, Generalized Pareto distribution, maximum likelihood, nonhomogeneous Poisson process, ozone, return levels.

1. INTRODUCTION

The traditional and best-known method of analyzing extreme values is based on the extreme value limiting distributions. These distributions, originally introduced by Fisher and Tippett (1928), arise as limits for the distribution of maxima in samples of independent, identically distributed random variables. In environmental applications, such as predicting extreme foods or sea levels, they are generally applied to the annual maxima of the series, though occasionally they are also applied to maxima over a different time period such as one month. The classical reference on these methods in Gumbel (1958), though there are a number of more recent proposals for fitting the distributions (e.g., Prescott and Walden 1980, 1983; Hosking, Wallis and Wood, 1985).

In recent years a number of alternative approaches have been studied. One method is to look at exceedances over high thresholds rather than maxima over fixed time periods. The idea of looking at extreme value problems from this point of view is very old, but the modern development seems to have started around 1970 with the "Peaks Over Threshold" or "POT" method (see, e.g., Todorovic and Zelenhasic, 1970), and further propounded in the English "Flood Studies Report" (NERC, 1975). This was paralleled by the mathematical development of procedures based on a certain number of extreme order statistics (Hill, 1975; Pickands, 1975; Weissman, 1978; to mention just three) and the Generalized Pareto distribution as a stable distribution for excesses over thresholds (Smith, 1984; Davison, 1984; Hosking and Wallis, 1987; Davison and Smith, 1989). Another approach, which partly combines the classical and threshold approaches, is to take a fixed number of order statistics from each year and to fit the appropriate extreme value distribution for the joint distribution of largest order statistics (Gomes, 1981; Smith, 1986; Tawn, 1988).

In this paper, ideas of extreme value theory are applied to the study of ozone in Houston, Texas. There has been much publicity recently of the ozone depletion problem in the upper atmosphere, but ground-level ozone is also a topic of considerable environmental concern, since excessive levels of ozone are taken as indicative of high air pollution generally. This is reflected in the formulation of U.S. air pollution standards in terms of the rate of exceedance by the ozone content of a specified threshold level (12 parts per 100 million). The current standards permit no more than three exceedances of this level in any 3-year period, but in a number of U.S. cities, including Houston, this standard is far from being met and the task of regulatory bodies such as the Texas Air Control Board is to introduce measures to reduce the frequency and level of high exceedances. The major contributory factors to the ozone problem are factory emissions

Richard L. Smith is Professor of Statistics, Department of Mathematics, University of Surrey, Guildford GU2 5XH, United Kingdom.

and vehicular exhaust, but the condition is exacerbated by particular meterological conditions, hot still weather being the worst from the point of view of allow high ozone levels to build up.

Extreme value theory has previously been used as a tool in studying air pollution problems (e.g., by Singpurwalla, 1972; Horowitz, 1980). This is partly because extreme values are of particular interest in assessing the impact of high air pollution, and partly because air-quality standards are formulated in terms of the highest level of permitted emissions. Roberts (1979) also reviewed the application of extreme value theory to air pollution problems and discussed further the role of statistical concepts in formulating air-quality standards. However, most existing methods are based on the traditional "annual maximum" approach to extremes, whereas the threshold approach to the statistical analysis has not so far been very well developed in this context.

The present paper takes the form of a case study. The data consist of hourly measurements of ozone over a 15-year period, and the problems of interest concern such issues as estimating the frequency with which specified high levels are exceeded, and especially whether there is any evidence of the frequency changing over the period of the study. The data feature many of the problems common in applying extreme value methods to environmental time series—short-range correlations, seasonal variation and the need to test for long-term trend—as well as the additional complication of missing values brought about by the measuring equipment being out of service for periods of anywhere between a few hours and a month. The combination of these features is such that none of the methods above is directly applicable; instead I propose a broader approach emphasizing the point-process viewpoint of high-level exceedances. This includes all the preceding methods as special cases and illustrates the wide applicability of extreme-value procedures.

The point-process approach to extreme value problems has been emphasized, though in quite different ways, in the books of Leadbetter, Lindgren and Rootzén (1983) and Resnick (1987), but both of these books are concerned primarily with the probabilistic rather than statistical aspects of extreme value theory. The present paper is intended to draw attention to the versatility of this approach and to discuss the practical problems present in any application of this nature.

2. THE DATA

The data consist of hourly readings of ozone in Houston, Texas from April 1973 to December 1986, nominally 119,905 values. Of these values, however, 24,472 (about 20%) are missing from periods when the equipment was out of service.

An earlier analysis of similar ozone data was performed by Davison and Hemphill (1987). That paper proposed models for the frequency of exceedances over a fixed threshold, justifying that approach on the grounds that air pollution standards are defined in that way. However, in cities such as Houston, where the legal threshold is exceeded very frequently, attention shifts to how best to control the problem. In assessing this, it is important to consider also the magnitudes of the exceedances. Particular interest is focused on whether there is any long-term trend in the results, since this may indicate the success or failure of the Air Control Board's efforts.

A histogram of the right-hand tail of the raw data is shown in Figure 1. Units are parts per 100 million, and the range of the histogram is from 12 to 34 (largest value). A raw histogram is, however, misleading as an indicator of how frequently high levels occur, partly because of the seasonality inherent in the data and more expecially because extreme values tend to occur in clusters.

The latter point suggests trying to identify clusters of high-level exceedances with the intention of concentrating on cluster maxima for the rest of the analysis. Some procedure of this form is standard in applications of the POT procedure mentioned in Section 1, but there is no universally accepted method for identifying the clusters. It is possible to take a model-based approach; for example, Smith (1984) used a doubly-stochastic model for the point process consisting of the times of high-level exceedances. More simple-minded approaches, however, seem to work just as well in practice. Such an approach is the one adopted here.

The approach taken to identify clusters consists of specifying a threshold and a cluster interval. Two

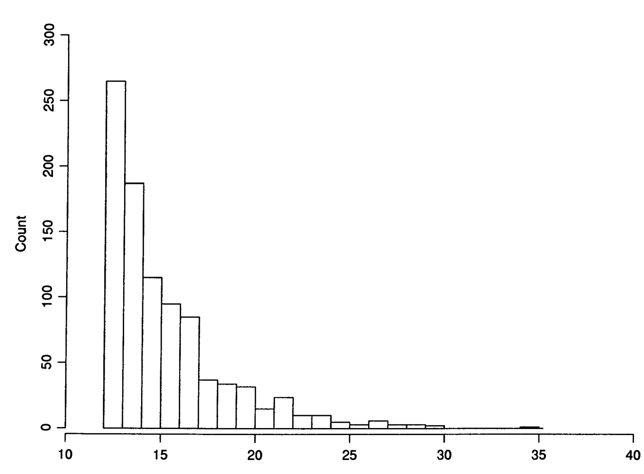


FIG. 1. *Histogram of ozone values over* 12.

exceedances of the threshold which are closer together than the cluster interval are deemed part of the same cluster, but when the time interval between successive exceedances is longer than the cluster interval it is considered that the old cluster has finished and a new one begun. In this way, clusters are defined and cluster maxima established. The threshold and cluster interval are both to some extent arbitrary, but the sort of considerations which go into their choice will become apparent from the subsequent discussion. In any case, it is recommended that different values be used for comparison.

For most of the analysis in this paper, a threshold of 8 (parts per 100 million) was adopted. This was chosen so as to be high enough to exclude the great bulk of the data which are of no relevance for extremes, but still to include enough data to allow detailed statistical analysis. Initially a cluster interval of 24 hours was chosen, this representing the rough intuitive judgment that exceedances further apart than that could be treated as effectively independent. When clusters were formed using the procedure just outlined, however, it was found that there were still a number of instances where only a few days separated clusters, so the procedure was repeated with a cluster interval of 72 hours. This produced more satisfactory results, and a cluster interval of 72 hours was adopted for most of the following analysis.

Figure 2 shows a histogram of cluster maxima formed by this procedure. The shape is very different from Figure 1, with a much longer tail reflecting the fact that very large values are much more likely to be cluster maxima than only moderately large values.

A number of other plots were drawn to try to identify which features of the data were important. Figure 3 shows a boxplot of the mean number of peaks per month, calculated for each month of the year. As

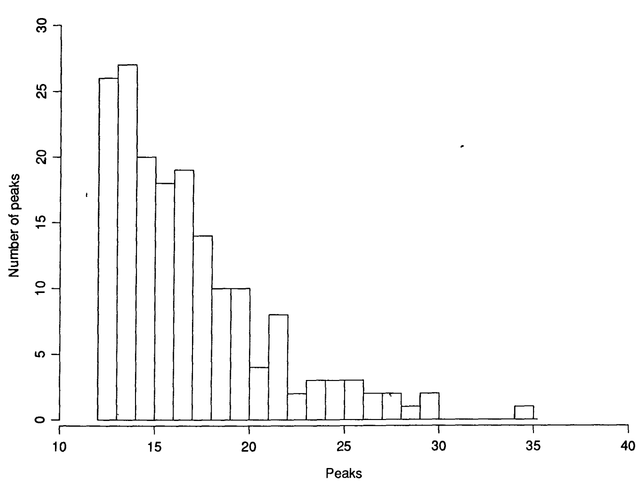


FIG. 2. *Histogram of cluster peaks over* 12.

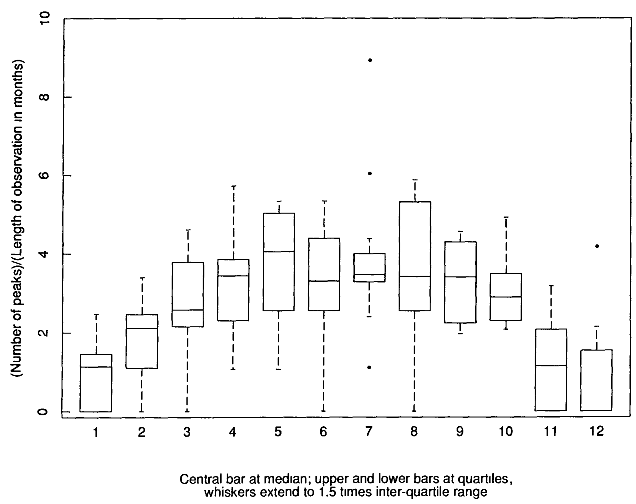


FIG. 3. *Boxplot for mean numbers of peaks/month.*

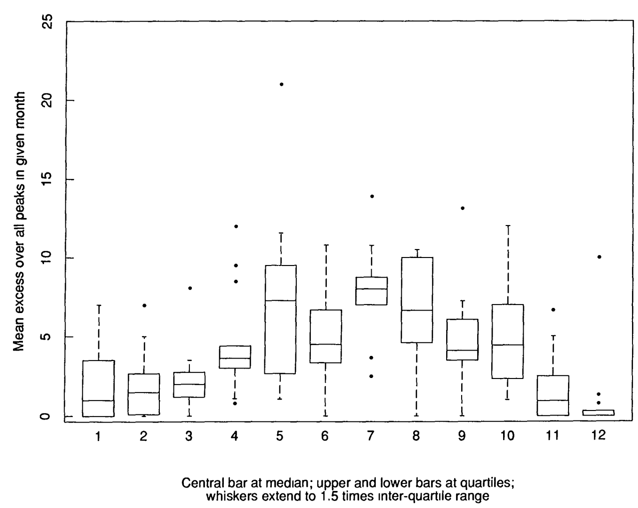


FIG. 4. *Boxplot for mean excess for each month.*

expected, there is a strong seasonal variation, with high values throughout the summer months through to October. However, there is also wide variability in the estimates, as indicated by the interquartile ranges and whiskers. Figure 4 shows a similar plot based on the mean excess over the threshold, i.e., the difference between mean cluster maxima and the threshold. This also shows a strong but irregular seasonal pattern.

Finally, in an attempt to make an initial judgment about long-term trend, the numbers of exceedances and mean excesses were plotted for each month over the period of the data, in Figures 5 and 6, respectively. Superimposed on these plots are moving averages to estimate the deseasonalized trend. These were calculated by applying a moving average filter to the 13 months centered on that point (Kendall, 1973, page 38). There is one prominent outlier in each plot but no other visual indication of any long-term trend.

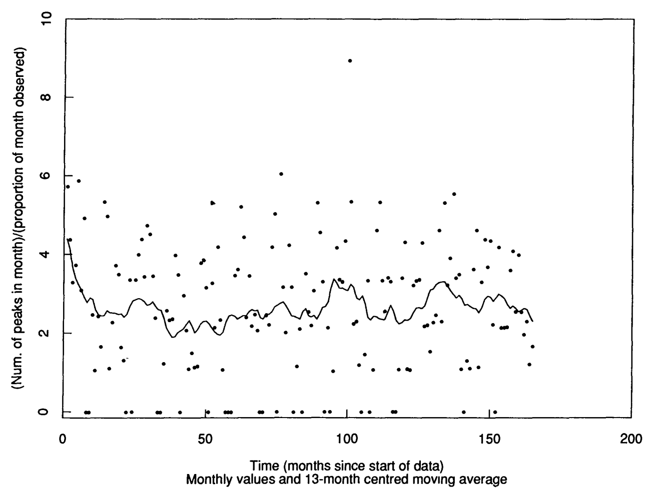


FIG. 5. *Crossing rates by month over threshold* 8.

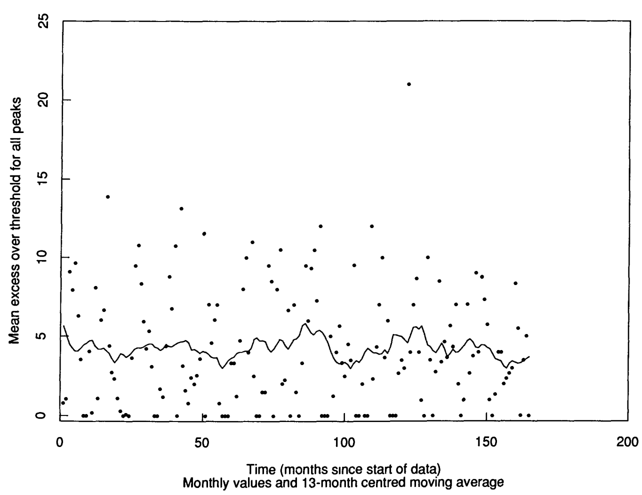


FIG. 6. *Mean excesses over threshold* 8.

3. EXTREME VALUE MODELS

This section reviews classical extreme value theory, leading up to the point process viewpoint which is used as the basis of the statistical method in Section 4. The initial discussion is restricted to independent, identically distributed (iid) random variables, though later that restriction will be removed.

Suppose $X_1, X_2, \cdots$ is an iid sequence with common distribution function F, and $M_n = \max(X_1, \cdots, X_n)$. Classical extreme value theory looks for normalizing sequences $a_n > 0$, b_n such that $(M_n - b_n)/a_n$ converges in distribution, so that

$$\Pr\{(M_n - b_n)/a_n \le x\} = F^n(a_n x + b_n) \to H(x), \tag{3.1}$$

where H is a nondegenerate distribution function. The convergence in (3.1) occurs if and only if

$$n\{1 - F(a_n x + b_n)\} \to -\log H(x). \tag{3.2}$$

Following the theory originally developed by Fisher and Tippett (1928) and Gnedenko (1943), it is known that H must be one of three types of limiting distributions. These three types may be combined into the single Generalized Extreme Value distribution

$$H(x; \mu, \sigma, k) = \exp[-\{1 - k(x - \mu)/\sigma\}^{1/k}] \tag{3.3}$$

valid on the range $\{x\colon 1 - k(x - \mu)/\sigma > 0\}$. Here $\sigma > 0$ and μ, k may be any real numbers, the case $k = 0$ being interpreted as the limit $k \to 0$,

$$H(x; \mu, \sigma, 0) = \exp[-\exp\{-(x - \mu)/\sigma\}], \tag{3.4}$$

widely called the Gumbel distribution.

The parameter k is called the shape parameter and may be used to model a wide range of tail behavior. The case $k < 0$ is that of a polynomially decreasing tail function and therefore corresponds to a long-tailed parent distribution. The case $k = 0$ is that of an exponentially decreasing tail, while $k > 0$ is the case of a finite upper endpoint and therefore short-tailed.

The family (3.3) may be fitted to data by numerical maximum likelihood, and an algorithm for this was published by Hosking (1985). Of the alternative methods of fitting, the most serious competitor to maximum likelihood is the "probability weighted moments" approach advocated by Hosking, Wallis and Wood (1985). However, the latter is at present restricted to single samples from a common distribution and is not therefore suitable for the more complicated types of modeling illustrated here. The regularity conditions of maximum likelihood are satisfied when $k < 0.5$, and alternatives are available when this condition is violated (Smith, 1985), but it is generally considered that the condition $k < 0.5$ is valid for most environmental applications.

The threshold approach is based on the distribution of exceedances over a high threshold u (say). Given that an observation exceeds u, the probability that it exceeds by at least y is $\{1 - F(u + y)\}/\{(1 - F(u)\}$. Under the same conditions as lead to (3.1), this may be approximated for large u by the family

$$G(y; \sigma, k) = 1 - (1 - ky/\sigma)^{1/k} \tag{3.5}$$

valid on $0 < y < \infty$ $(k \le 0)$ or on $0 < y < \sigma/k$ $(k > 0)$. This is the *Generalized Pareto Distribution* of Pickands (1975). The case $k = 0$ is the exponential distribution, which was used in all early applications of the POT method. Its applications to hydrological and air-pollution data have been discussed by Smith (1984), Davison (1984) and Hosking and Wallis (1987). Another approach is based on the joint distribution of several large order statistics from a sample, instead of just the maximum. This may be derived as an extension of (3.2) and (3.3), or using point-process arguments as given below. A procedure using this joint distribution was proposed by Weissman (1978) in the case of a single sample of iid data, and subsequently extended by a number of other authors. Smith (1986) advocated an approach to hydrological data analysis in which a fixed number r of largest values in each year was taken, the joint distribution of r largest values being fitted by maximum likelihood. This approach seems to improve on the classical ($r = 1$) approach, but it is important not to take r too large, as then the fit of the model is not nearly so good (Smith suggested $r = 5$ as a reasonable compromise for the data analyzed there). This method was extended by Tawn (1988), who also recommended declustering the data before picking out the order statistics.

The point-process approach to these problems was originally introduced by Pickands (1971). Suppose (3.2) holds for some normalizing sequences a_n and b_n. Let $X_1 \cdots, X_n$ denote a random sample from F and let $Y_{n,i} = (X_i - b_n)/a_n$, $i = 1, \cdots, n$. Let P_n denote the point process on $\mathscr{R}^2$ with points at $(i/(n+1), Y_{n,i})$, $i = 1, \cdots, n$. The ordinates of P_n will tend to cluster near the lower endpoint of the (rescaled) distribution, but away from the boundary the process will look like a nonhomogeneous Poisson process. Weak convergence of P_n to a process P is easily established under a topology which essentially excludes sets bordering the lower boundary. The intensity measure of the limiting process is derived from (3.2) and (3.3), as

$$\begin{aligned}\Lambda\{(t_1, t_2) \times (x, \infty)\} \\ = (t_2 - t_1)[1 - k(x-\mu)/\sigma]^{1/k},\end{aligned} \tag{3.6}$$

whenever $0 \le t_1 \le t_2 \le 1$ and $1 - k(x - \mu)/\sigma > 0$.

All the previously mentioned results in extreme value theory may be derived from this representation. For instance, the probability that $(M_n - b_n)/a_n$ is less than x is just the probability that P_n has no points in $(0, 1) \times (x, \infty)$. Under the Poisson process limit with intensity (3.6), this is precisely (3.3). The distribution of the rth largest order statistic, or the joint distribution of the r largest (for fixed r as n grows), may be derived almost as easily. The Generalized Pareto distribution may be derived from this approach as well: the limiting conditional probability that $Y_{n,i} > u + y$ given $Y_{n,i} > u$ is

$$\frac{\Lambda\{(0,1)\times(u+y,\infty)\}}{\Lambda\{(0,1)\times(u,\infty)\}} = \left[1 - \frac{ky}{\sigma - ku + k\mu}\right]^{1/k} \tag{3.7}$$

which is just the Generalized Pareto Distribution on replacing σ in (3.5) by $\sigma - ku + k\mu$.

The statistical approach advocated in this paper is based on viewing the high-level exceedances as points of a Poisson process. The intensity function depends on the parameters μ, σ and k (which may themselves depend on additional parameters such as the slope of a trend) and may be fitted to the data by maximum likelihood. Of course, in the statistical context we do not know a_n and b_n, but this is effectively taken care of in the estimation of μ and σ. All the previous approaches may be regarded as special cases of this approach.

So far, the discussion has been confined to iid sequences. Environmental data tends to depart from this in two respects: first, in being heavily seasonal, and second, in exhibiting short-range dependence leading to clustering of high-level exceedances. As we saw in Section 2, both of these features are very strongly present in the data set under discussion. The remainder of this section is devoted to a brief discussion of how the ideas presented here may be modified to take account of these features.

The theory of extreme values in dependent stochastic processes has been very extensively developed and is summarized in the book of Leadbetter, Lindgren and Rootzén (1983) and the review article of Leadbetter and Rootzén (1988). For stationary processes, the classical extreme value laws (3.3) remain valid under a mild mixing condition (Leadbetter's condition D) which covers a very wide range of processes. There is then a general relation of the form

$$\Pr\{M_n \le x\} \approx [F(x)]^{n\theta} \tag{3.8}$$

where $0 \le \theta \le 1$ is a parameter called the *extremal index* for the process. This is a measure of the amount of clustering in the process, $1/\theta$ being the limiting mean cluster size. When $\theta > 0$, knowledge of θ together with F is all that is needed to determine the limiting distribution of sample maxima.

Given a specific model, it is often possibly to verify (3.8) and to calculate θ. In cases such as ours, in which no specific model for the dependence is being considered, it seems reasonable to proceed directly by identifying clusters of exceedances and studying the distribution of cluster maxima. The general theory supports our approach by showing that the idea of a limiting Poisson process of cluster maxima is valid under very broad assumptions.

So much for stationary processes. What about nonstationary sequences? In this case, the general theory is much less helpful. Extreme value theory of independent, non-identically distributed summands was developed by Mejzler in the 1940's (Galambos, 1987 reviewed this topic well), but the class of limiting distributions is much too wide to be of use in identifying parametric statistical models. Modern developments are due to Weissman (1975) for the point process approach, and Hüsler (1986) for nonstationary stochastic processes. In practical statistical terms, two approaches have been considered for seasonal series. One is to decompose the whole series into a sum of an effectively deterministic seasonal component and a random noise, the latter being assumed stationary.

This idea was developed under the name of "joint probability method" for separating out tidal and surge effects in sea level studies (Pugh and Vassie, 1980); in a different form it was also used by Smith (1984). A recent extension of the joint probability method, incorporating modern developments in extreme value theory, is due to Tawn and Vassie (1989). This approach seems reasonable when there is a strong physical mechanism underlying the seasonal variation, as is obviously the case with tides. In other cases, however, including the example under discussion here, there is no reason to suppose that a simple decomposition will work. The second approach which has been tried, and is of broader utility at the cost of more parameters, is to allow all the parameters of the process to be seasonally dependent. This approach is the one adopted here.

4. STATISTICAL PROCEDURE

In this section a detailed procedure is described, which is motivated by the general discussion of Section 3.

It is assumed that the data have already been declustered using a procedure similar to that in Section 2, so that the data under study are maxima of clusters of exceedances over a high threshold. The year is divided up into M-day periods where, for instance, $M = 31$ for (approximately) monthly sections, and a separate model fitted to each period. Thus, no specific form of seasonal variation (such as a sinusoid) is being assumed. Let N_{ij} denote the number of cluster maxima in period j of year i, and let Y_{ijm} (for m between 1 and N_{ij}) denote the individual excesses. Let p_{ij} denote the length of observation (in years) in period j of year i. The p_{ij}'s are not all equal, because of the missing data.

For period j of year i, let the extreme value parameters be μ_{ij}, σ_{ij}, k_{ij}. In the following discussion we shall only consider models of form

$$\mu_{ij} = \alpha_j + i\beta_j, \quad \sigma_{ij} = \sigma_j, \quad k_{ij} = k_j, \tag{4.1}$$

so that the parameters depend on i only through a possible linear drift β_j, and in many instances this will also be ignored. It is of course possible to test other hypotheses, such as whether the k_j's are all equal, within the general context of model (4.1).

For the data within period j of year i, the exceedance times and excesses are taken to form a nonhomogeneous Poisson process with intensity measure given by (3.6) with μ_{ij}, σ_j, k_j substituted for μ, σ, k. With the threshold taken as 0, the likelihood function is

$$L = \prod_{i,j} \left[\exp\{-p_{ij}(1 + k_j \mu_{ij}/\sigma_j)^{1/k_j}\} \cdot \prod_{m=1}^{N_{ij}} \{(1 - k_j(y_{ijm} - \mu_{ij})/\sigma_j)^{1/k_j - 1}/\sigma_j\} \right]. \tag{4.2}$$

Special cases such as $k_j \equiv 0$ (the Gumbel model) or $\beta_j \equiv 0$ (no drift) may be derived from (4.2).

Estimation proceeds by minimization of $-\log L$ using a Newton-Raphson or quasi-Newton iteration, a routine for the latter being taken from Nash (1979). Note that L factorizes on j, so it is possible to minimize separately for each of the M-day periods. Even so, the numerical minimization is nontrivial, and it was found that the quasi-Newton method (in which second-order derivatives are not calculated explicitly but an approximation produced by the algorithm) performed better than a Newton-Raphson method based on exact first- and second-order derivatives. It is not clear why this should be the case, but there is certainly a difficulty in that there are large regions of the parameter space where the log likelihood is not concave, and it is possible that in this region the performance of the quasi-Newton algorithm is superior. Starting values were obtained by using the method of moments to fit a Gumbel distribution to the M-day maxima. The maximized likelihood was then used to produce parameter estimates, with standard errors from the inverse of the observed information matrix, and likelihood ratios for discrimination among different models.

As a final part to this procedure, it is possible to use the fitted model to estimate the mean rate of exceedance in year i over a specified high level y. This is given by

$$\tau(y) = (M/365) \sum_j \{1 - k_j(y - \mu_{ij})/\sigma_j\}^{1/k_j} \tag{4.3}$$

where the summand denotes the mean rate of exceedance in period j. This, in turn, is multiplied by $M/365$, the length in years of each period. An estimate of $\tau(y)$ and its standard error may be obtained directly from the fitted model.

5. RESULTS

The model of Section 4 will now be applied to the ozone data. There are still many decisions to be made concerning the initial specification of threshold and cluster interval, the length of the periods into which the year is broken, and the relations among the various extreme value parameters.

A first analysis was based on threshold 8, cluster interval 72 hours and period $M = 31$ days, so that the year is in effect being broken up into months. Table 1 shows the results of fitting model (4.1) with $\beta_j \equiv 0$, i.e., no trend. No fit was obtained for periods 1 and 12, but this is not surprising as the number of exceedances in these periods was very small. For the remaining periods, there is a strong seasonal pattern in the location parameter α, but the other two, σ and k, vary haphazardly. Also shown is $\alpha + \sigma(1 - 0.1^k)/k$, which represents the value that would be attained once every

TABLE 1
Parameter estimates for 31-day periods with no trend

Period	Number of observations	α	σ	k	$\alpha + \sigma(1 - 0.1^k)/k$
1	10	—	—	—	—
2	12	15.0 (0.8)	1.0 (0.7)	.66 (.46)	16.2
3	24	16.1 (0.7)	0.9 (0.4)	.54 (.19)	17.3
4	28	23.5 (2.4)	2.7 (1.5)	.29 (.24)	28.0
5	31	26.3 (1.8)	2.4 (0.8)	.39 (.14)	29.9
6	33	25.8 (2.0)	2.7 (0.9)	.31 (.14)	30.2
7	39	27.3 (1.4)	2.0 (0.6)	.44 (.11)	30.2
8	35	25.3 (0.9)	1.1 (0.7)	.64 (.19)	26.7
9	34	26.8 (3.5)	4.6 (1.8)	.07 (.14)	36.5
10	31	25.3 (2.3)	3.0 (1.3)	.28 (.19)	30.4
11	11	16.7 (2.5)	3.1 (1.4)	.14 (.29)	22.8
12	6	—	—	—	—

Standard errors in parentheses.

10 years on average, if the fitted model were valid over the whole year. This and the value of α follow the pattern expected from Figures 3 and 4, i.e., of a strong seasonal pattern with a peak in the summer months extending through to October. All the values of the shape parameter k are positive and three are over 0.5, the value at which the regularity of maximum likelihood fails. The wide variation among the estimates may reflect statistical variability more than anything else, but the estimates do correspond to a short upper tail to the distribution. It may be that this in itself reflects efforts being made by the authorities to limit the occurrence of very high levels.

As mentioned earlier, there is considerable interest in whether there is trend in the data, so the models were re-fitted including the β_j parameters. Only in two cases—periods 2 and 6—was a significant β_j found, in both cases negative.

The results so far are hard to interpret. One reason may be that the model is just overparametrized. It is possible to reduce the number of parameters by breaking up the year into larger periods, so a new analysis based on $M = 61$ days was tried. There is some doubt over whether it is valid to combine periods 3 and 4, in view of the results of Table 1, but otherwise it seems reasonable. Results are in Table 2.

To investigate sensitivity of the results to M and to initial threshold and cluster interval, the same models were fitted (without trend) to data based on eight different initial choices, with results in Table 3. This comparison is based on estimates of return values for 3, 10 and 50 years. The n-year return value is defined as that value of y for which $\tau(y)$ is $1/n$. This can be estimated by evaluating (4.3) for many values of y and inverting. Those periods for which no estimates were obtained were ignored in these calculations, noting that, because they were in a part of the year when ozone levels are lowest, it is unlikely that they would have a significant influence on the calculations. As can be seen, even for 50-year estimates there is very little variability due to the initial choice of threshold and cluster interval.

Table 3 shows good agreement amongst the point estimates of return values, and this suggests that we should also try to construct confidence intervals. In principle, this can also be done from (4.3): we can obtain a standard error and hence a confidence interval for each $\tau(y)$ and can then invert the lower and upper confidence bounds to obtain a confidence interval for return value. However, this did not yield such good results: for the model in Table 1, approximate 95% confidence intervals of (25.8, 30.3), (27.1, 32.3) and (28.0, 40.8), respectively, for the 3-, 10- and 50-year levels were obtained. These results are not so satisfactory, which is one reason why in subsequent analysis attention is focused on exceedance rates rather than return levels.

So far, the investigation of trend has been inconclusive. Even if there is a trend, however, there is no reason to suppose that it has to take the form of a linear additive trend, as required by (4.1). An alternative approach is to split the data into two parts and fit the model without trend separately to each. The question is also raised by Table 2 of whether we could assume a common k for the whole year. This could be investigated by minimizing the negative log likelihood under the assumption of fixed k, for several trial values of k. Minimum values 101.0, 96.0, 92.8, 92.9, 97.5, 106.2 for $k = 0, 0.1, \cdots, 0.5$ were obtained, suggesting strong evidence against $k = 0$ but little to choose in the range 0.2 to 0.3. Trying $k = 0.25$ produced the slightly smaller value 92.3. Therefore, both the "fixed k" and "separate k" models were fitted both to the whole data and separately to the years 1973–1980, 1981–1986 ($M = 61$) with the following results:

Separate k, full data:
Neg log likelihood 85.4 (18 parameters)

TABLE 2
Parameter estimates for 61-day periods with no trend

Period	Number of observations	α	σ	k	$\alpha + \sigma(1 - 0.1^k)/k$
1	22	15.0 (0.6)	0.9 (0.4)	.76 (.26)	16.0
2	50	21.1 (1.9)	3.2 (1.1)	.12 (.14)	37.5
3	64	25.9 (1.3)	2.4 (0.5)	.36 (.09)	29.8
4	74	26.9 (1.0)	1.9 (0.4)	.46 (.08)	29.6
5	69	26.0 (2.1)	4.0 (1.0)	.13 (.10)	34.0
6	17	20.5 (3.1)	5.3 (2.5)	−.05 (.28)	34.0

Standard errors in parentheses.

TABLE 3
Comparison of return values under different choices for threshold, clustering interval and period

Threshold	Cluster interval (hours)	Period (days)	3-year	10-year	50-year
8	72	31	27.4	29.7	33.4
8	24	31	27.4	29.9	33.7
10	72	31	27.2	30.0	33.9
12	72	31	27.2	29.8	33.8
8	72	61	27.8	29.5	35.9
8	24	61	27.6	30.5	36.4
10	72	61	27.5	29.9	33.9
12	72	61	27.6	30.3	34.6

All under three-parameter model for each period, no trend.

Fixed $k = .25$, full data:
Neg log likelihood 92.3 (12 parameters)
Separate k, split data:
Neg log likelihood 77.1 (36 parameters)
Fixed $k = .25$, split data:
Neg log likelihood 87.4 (24 parameters)

Based on standard likelihood ratio considerations, there seems no reason to split the data but some evidence against a common k.

It is also possible, however, to investigate estimated exceedance rates under these models. These are shown in Table 4 for three moderately high thresholds and a number of different models. The results here are qualitatively quite different from those given for the parameter estimates: the distinction between "fixed k" and "separate k," for instance, does not now seem so important, but it is remarkable that all the models allowing for a trend (whether linear or by splitting data) make substantially lower estimates for 1988 than for 1973. It appears that the evidence for or against a particular model is stronger when attention is focused on a particular quantity of interest than when considered on the basis of the overall fit to the model.

Table 4 also raises the question of whether, in the "fixed k" case, different values should be taken for the two halves of the data. This was tried but with no significant result.

To summarize our conclusions so far:

1. The initial analysis, based on 31-day periods, showed clearly that $k > 0$ but gave no conclusive evidence about trend. The wide variability of the estimates suggests that the model is overparametrized.
2. Investigations of a range of models showed no strong sensitivity to the choice of threshold, cluster interval or period (i.e., 31 or 61 days) within the values considered, but an attempt to provide confidence intervals for return values was not successful.
3. Further analysis based on 61-day periods and allowing for trends showed no significant differences based on the overall fit, but there were significant differences in computed crossing rates of high levels. This conclusion is obtained for both the "fixed k" and "separate k" models.

Thus our overall conclusion is that there is a significant downward trend in the crossing rates of high levels. A natural question is to what extent this can be observed in a purely empirical analysis of the data. In Table 5, crossing rates by cluster maxima over several high levels are given separately for the two halves into which we have split the data, and it can be seen that they do confirm the picture that at the

TABLE 4
Calculations of exceedance rates for "separate k" and "fixed k" models based on full data and on split data

Level	26		28		39	
Separate k, $M = 31$, no drift	0.56	(.17)	0.23	(.11)	0.08	(.07)
Separate k, $M = 31$, linear drift[a]	1.27	(.52)	0.65	(.34)	0.24	(.18)
Separate k, $M = 31$, linear drift[b]	0.39	(.30)	0.15	(.17)	0.06	(.08)
Separate k, $M = 61$, no drift	0.67	(.18)	0.30	(.12)	0.12	(.08)
Separate k, $M = 61$, linear drift[a]	1.05	(.40)	0.51	(.24)	0.20	(.13)
Separate k, $M = 61$, linear drift[b]	0.31	(.20)	0.15	(.10)	0.08	(.07)
Fix $k = 0.25$, $M = 61$, no drift	0.73	(.19)	0.39	(.14)	0.19	(.09)
Separate k, $M = 61$, 1973–1980	0.96	(.31)	0.42	(.20)	0.13	(.12)
Fix $k = 0.25$, $M = 61$, 1973–1980	1.07	(.34)	0.65	(.26)	0.13	(.10)
Separate k, $M = 61$, 1981–1986	0.26	(.17)	0.12	(.12)	0.06	(.09)
Fix $k = 0.25$, $M = 61$, 1981–1986	0.33	(.18)	0.13	(.10)	0.04	(.04)

Standard errors in parentheses.
[a] Estimates made for year 1973.
[b] Estimates made for year 1988.

TABLE 5
Empirical exceedance rates for split data

Level	8	12	16	20	26
Frequency 1973–1980[a]	175	96	56	21	5
Rate/year	22.8	12.5	7.3	2.7	0.7
Frequency 1981–1986	161	79	29	10	3
Rate/year	26.8	13.2	4.8	1.7	0.5

[a] From April 28, 1973; 7.68 years.

highest levels (though not elsewhere) there is a downward trend. However, there are very few exceedances at the levels considered in Table 4, so some form of model-based analysis is essential if quantitative estimates are to be given at such high levels. It might also be pointed out that the levels being studied are much higher than the official standard of 12 parts per hundred million.

The last question I shall address here concerns testing the fit of the extreme-value distribution. Motivated by the relations among the various extreme value properties in Section 3, two studies are proposed: (a) how closely the excesses over the threshold fit a Generalized Pareto distribution and (b) how closely the period maxima follow a Generalized Extreme Value distribution. In Figure 7, a plot of observed versus expected values is shown for excesses over a threshold, for each of the six periods of the year (superimposed). This is done: (a) fitting the exponential distribution (i.e., $k = 0$) and (b) fitting the Generalized Pareto distribution with separate k. In each case the values were standardized, via equation (3.7), to have scale parameter 1 before plotting. Of course, it is expected from the previous analysis that the Generalized Pareto fits better than the exponential, but Figure 7 provides visual confirmation that the

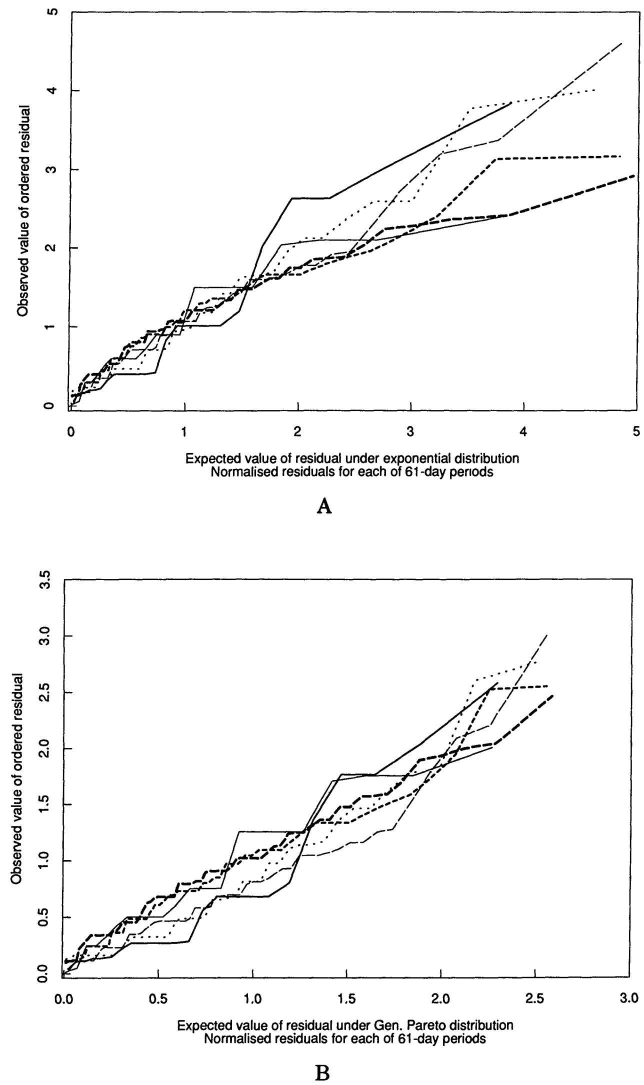
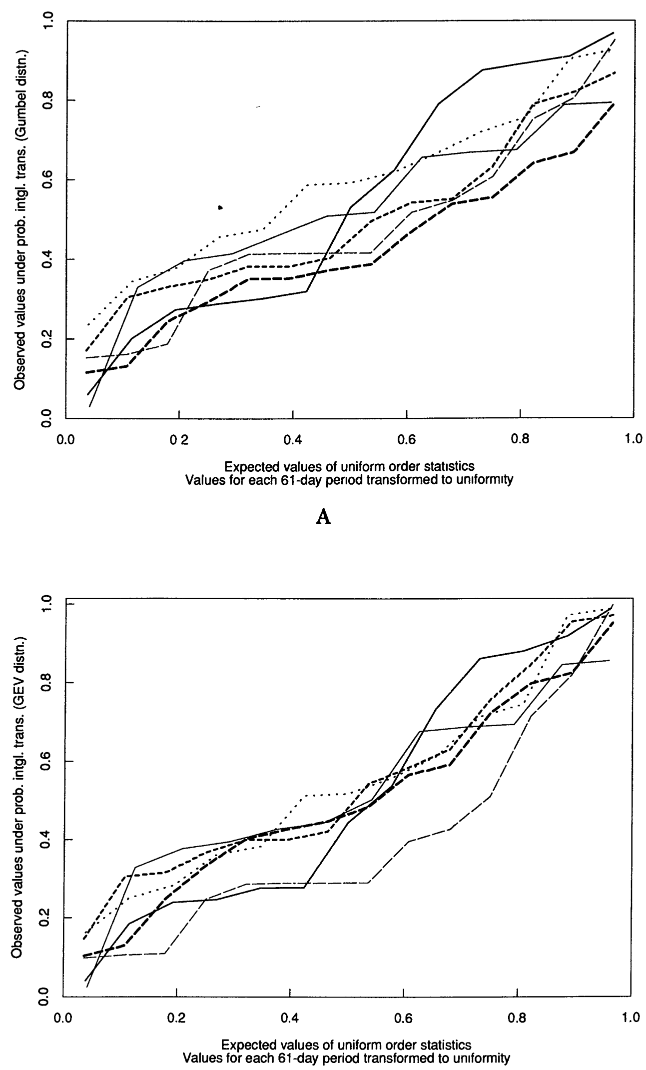


FIG. 7. (A) *Residual plot under exponential model, and* (B) *residual plot under GPD model.*



B

FIG. 8. (A) *Probability plot of* 61-*day maxima* (*Gumbel model*), *and* (B) *probability plot of* 61-*day maxima* (*GEV model*).

exponential is indeed a poor fit in the upper tail, whereas the Generalized Pareto distribution seems quite adequate. Figure 8 shows a differently constructed plot for the period maxima, in which the probability integral transform is used to transform to uniformity before plotting the empirical distribution function. Again two plots were constructed, one based on $k = 0$ and the other on k varying. Again the second plot is much better in the upper tail. Neither plot is a very good fit in the lower tail, but this is of less concern in view of the emphasis on upper extremes in our analysis. Smith (1986) showed how to construct similar plots for the marginal distribution of the jth largest order statistic for fixed $j \geq 2$. This was also tried here but with results similar to Figure 8.

Overall, these plots provide confirmation that the models fitted are sensible.

6. CONCLUSIONS AND SUMMARY

Classical extreme value theory has suggested a number of techniques for statistical modeling. These have been widely applied but are deficient at handling the more complicated features of practical data. The modeling approach proposed here uses the ideas of point-process theory to suggest a very general strategy.

The analysis of the ozone data included the comparison of numerous different models within the general structure proposed in Section 4 and led to the conclusion of a downward trend in crossing rates at the very highest levels. This would seem to be some indication that the work of the regulatory bodies is having some effect in at least limiting the frequency of very high emissions, and indeed one could argue that without such regulation there would have been a clear increasing trend. As far as the *impact* of the suggested downward trend is concerned, one interpretation of Table 4 is that the crossing rate of level 26 found at the end (in the range 0.3–0.4) corresponds to a level between 28 and 30 at the beginning; it seems justified, therefore, to say that the most extreme emissions have been reduced by about 3 parts per 100 million over the period of the study. They are still well above the federal standard, but it does indicate some measure of improvement. It should be pointed out, however, that the present study is only for one site, and it would be necessary to repeat the analysis with data collected at other sites to get a firm indication of this.

The other noticeable feature of our analysis is that, in contrast to many extreme value problems in which the shape parameter k is estimated to be near 0, in this case k had significant positive values. It should be pointed out, however, that in many cases where k turns out significantly negative, the analysis is much harder since the tail is very long, and it becomes very hard to give anything like precise estimates of quantiles. This difficulty is discussed at some length by Davison and Smith (1989).

A final comment is that the analysis has only considered a single long time series, with no attempt to relate ozone emissions to other covariates (e.g., temperature) or to relate values at different sites. The former gives rise to regression-type problems, the latter to problems of multivariate extremes. Davison (1984) and Davison and Smith (1989) have discussed the use of regression in this kind of analysis, but there have still not been many practical applications. Concerning multivariate extremes, this is the subject of much current research, and extensions of the methodology proposed here are anticipated.

ACKNOWLEDGMENTS

I am grateful to Professor Tom Shively, of the University of Texas at Austin, for sending me the data and for discussions about the ozone problem. I would like to thank Professor J. V. Zidek and Dr. Harry Joe for inviting me to the University of British Columbia, where part of the research was supported by the Natural Sciences and Engineering Research Council of Canada. The revision was prepared at the Department of Statistics, University of Chicago, and made use of computer facilities supported in part by the National Science Foundation under grants DMS-86-01732 and DMS 87-03942 awarded to the Department, and by the University of Chicago Block Fund.

REFERENCES

Davison, A. C. (1984). Modelling excesses over high thresholds, with an application. In *Statistical Extremes and Applications* (J. Tiago de Oliveira, ed.) 461–482. Reidel, Dordrecht.

Davison, A. C. and Hemphill, M. W. (1987). On the statistical analysis of ambient ozone data when measurements are missing. *Atmospheric Environment* **21** 629–639.

Davison, A. C. and Smith, R. L. (1989). Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B.* To appear.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proc. Cambridge Philos. Soc.* **24** 180–190.

Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*, 2nd ed. Krieger, Melbourne, Fla.

Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. of Math.* (2) **44** 423–453.

Gomes, M. I. (1981). An i-dimensional limiting distribution function of largest values and its relevance to the statistical theory of extremes. In *Statistical Distributions in Scientific Work* (C. Taillie, G. P. Patil and B. A. Baldessari, eds.) **6** 389–410. Reidel, Dordrecht.

Gumbel, E. J. (1958). *Statistics of Extremes.* Columbia Univ. Press, New York.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174.

Horowitz, J. (1980). Extreme values from a nonstationary stochastic process: An application to air quality analysis (with discussion). *Technometrics* **22** 469–482.

Hosking, J. R. M. (1985). Algorithm AS215: Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Appl. Statist.* **34** 301–310.

Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics* **29** 339–349.

Hosking, J. R. M., Wallis, J. R. and Wood, E. F. (1985). Estimation of the generalised extreme-value distribution by the method of probability-weighted moments. *Technometrics* **27** 251–261.

Hüsler, J. (1986). Extreme values of non-stationary random sequences. *J. Appl. Probab.* **23** 937–950.

Kendall, M. G. (1973). *Time Series.* Griffin, London.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Series.* Springer, New York.

Leadbetter, M. R. and Rootzén, H. (1988). Extremal theory for stochastic processes. *Ann. Probab.* **16** 431–478.

Nash, J. C. (1979). *Compact Numerical Algorithms for Computers.* Adam Hilger, Bristol.

NERC (1975). *The Flood Studies Report.* The Natural Environment Research Council, London.

Pickands, J. (1971). The two-dimensional Poisson process and extremal processes. *J. Appl. Probab.* **8** 745–756.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3** 119–131.

Prescott, P. and Walden, A. T. (1980). Maximum likelihood estimation of the parameters of the generalized extreme value distribution. *Biometrika* **67** 723–724.

Prescott, P. and Walden, A. T. (1983). Maximum likelihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples. *J. Statist. Comput. Simulation* **16** 241–250.

Pugh, D. T. and Vassie, J. M. (1980). Applications of the joint probability method for extreme sea level computations. *Proc. Inst. Civ. Engrs. Part 2* **69** 959–975.

Resnick, S. (1987). *Extreme Values, Point Processes and Regular Variation.* Springer, New York.

Roberts, E. M. (1979). Review of statistics of extreme values with applications to air quality data. Part I: Review. Part II: Applications. *J. Air Pollution Control Assoc.* **29** 632–637, 733–740.

Singpurwalla, N. D. (1972). Extreme values from a lognormal law with application to air pollution problems. *Technometrics* **14** 703–711.

Smith, R. L. (1984). Threshold methods for sample extremes. In *Statistical Extremes and Applications* (J. Tiago de Oliveira, ed.) 621–638. Reidel, Dordrecht.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika* **72** 67–92.

Smith, R. L. (1986). Extreme value theory based on the r largest annual events. *J. Hydrology* **86** 27–43.

Tawn, J. A. (1988). An extreme value theory model for dependent observations. *J. Hydrology* **101** 227–250.

Tawn, J. A. and Vassie, J. M. (1989). Extreme sea levels: The joint probability method revised and revisited. *Proc. Inst. Civ. Engrs. Part 2.* To appear.

Todorovic, P. and Zelenhasic, E. (1970). A stochastic model for flood analysis. *Water Resources Research* **6** 1641–1648.

Weissman, I. (1975). Extremal processes generated by independent non-identically distributed random variables. *Ann. Probab.* **3** 172–177.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.* **73** 812–815.