Rejoinder

Richard L. Smith

My thanks to all six discussants for their comments, which have focused attention on a number of issues raised by the paper. In my reply I have attempted to classify what the major issues are, dealing with the individual discussants' points under those general headings.

1. IS THERE REALLY A TREND?

Given the emphasis of the paper and the practical importance of the question, it is not surprising that several discussants have focused on this issue, and Raftery has provided an enlightening alternative analysis.

Let me first remind the reader of my own conclusions on this question. The evidence for a downwards trend is by no means clear-cut. Fitting model (4.1) with estimated β_j yielded nothing at all. A likelihood ratio test based on split data (Section 5) also failed to produce a significant result, though, as Raftery points out, it would have been worth trying some more parsimonious forms of the alternative hypothesis. Only the calculation of exceedance rates in Table 4 produced any solid evidence for a trend, and even there it is hard to be sure about their significance. This may indeed be rather a weak conclusion but I think it fairly reflects the evidence in the data.

Raftery has proposed an alternative analysis based on the point process of exceedances of a fixed level. By concentrating on cluster means and employing a time transformation to take account of both seasonal effects and the missing data, he creates a data set for which the null hypothesis of a homogeneous Poisson process would be reasonable, and he then tests this against the alternatives of: (1) a log-linear decay in the intensity, and (2) a change-point model. The evidence against homogeneity is stronger using the change-point alternative than a log-linear decay, but still "not worth more than a bare mention."

I think this is consistent with my own conclusions. Indeed, merely from the data in Table 5 it is possible to carry out a likelihood ratio test of whether the Poisson rate is the same over the two halves of the data. For exceedances of level 16, I obtain a deviance statistic (nominally χ_1^2) of 3.36, while, based on exceedances of level 20, the corresponding value is 1.74. Again, this is some evidence but hardly very strong.

As a technical aside, a full development of Raftery's analysis would presumably take account of the fact that the seasonal variation itself depends on estimated parameters. This would be tedious but straightforward to incorporate into the analysis of model (1), but model (2), with its nonregular features, may pose more problems.

The real import of these conclusions, however, can only be assessed in comparison with similar analyses carried out at other sites. My understanding is that ozone analyses at other sites in Texas have yielded far more clear-cut evidence of downward trend than this one. There is also an argument that the absence of a clear *upward* trend is evidence in itself that air quality regulations are having an effect. Thus it may be better to focus on *estimation* of a trend rather testing for its existence. In that case, the main message of the paper would be not to look for a simple additive trend but to measure it in terms of estimated exceedance rates of high levels, in which case the evidence for a trend may well depend on the level chosen. Fairley's comments reinforce this point; standard statistics such as mean or median may not give an accurate indication of what is happening at extreme levels.

2. FORMATION OF CLUSTERS

The other aspect which all the discussants mention in some way is the local behavior of the series near high exceedances. This can be subdivided into several individual points.

Identification of clusters. The paper used a very crude rule to identify clusters, and both Joe and Singpurwalla queried the appropriateness of this. To answer a point of Singpurwalla's, if the clustering rule failed to produce approximately independent clusters then equation (3.6) would indeed not be valid. In Smith (1984) and Davison and Smith (1989), an alternative method has been proposed based on assuming the full point process of exceedances to be a simple doubly stochastic process. However, the clustering procedure proposed is very similar to the one in this paper. One could also consider methods based on, say, the Neyman–Scott or Bartlett–Lewis models for a clustered point process. There is scope for more work here. In the present study I do place considerable weight on Table 3, but it is reasonable to ask what we would do if the results were not so good, as might indeed happen if there were longer clusters than there appear to be in the present data set. In such cases, a more careful treatment of the clustering problem might well be required.

Structure within a cluster. Having identified the clusters, my analysis used just the times and peak values, ignoring everything else that goes on within a cluster. One way to extend this is to consider both the size of a cluster (defined as the number of exceedances within a cluster) and its length (time from first exceedance to last). Based on hourly data, the means are shown in Table 1.

The extremal index mentioned by Weissman is essentially the reciprocal of mean cluster size. As can be seen, it *does* depend quite markedly on threshold and cluster interval, so we are clearly not far enough in the tail for the asymptotic theory mentioned by Weissman to hold. Nevertheless, these statistics should prove useful in assessing the broader impact of the ozone exceedances.

There is an additional problem in going from hourly to daily data. If I understand Chock's point (mentioned by Joe) correctly, it is to query the appropriateness of federal standards based on the total number of days on which the standard is exceeded, when the possibility of day-to-day dependence exists. My proposed solution to this problem is as follows. By taking a sample average over clusters, it is possible to estimate the probability that a cluster contains exactly k daily exceedances of whatever threshold is specified, for each of $k = 1, 2, \ldots$. Combined with the Poisson process of clusters, this then gives a compound Poisson distribution for the number of daily exceedances per year (compare the theoretical results in Hsing, Hüsler and Leadbetter, 1988). From this it is possible to calculate the mean number of daily exceedances per year, for any specified high threshold.

An even broader problem would be to describe the full joint distribution of high values within a cluster. The asymptotic theory has been given by Mori (1977) and Hsing (1987), but the general representation is very complicated, and depends further on the stability which was just found to be lacking in the case of extremal index. So far I have not pursued this, but it would be a useful practical exercise to describe other features of the structure of clusters.

Alternative approaches. The main alternative to the clustering approach used in the present paper is to model the full process generating the data and then to calculate extreme value distributions from that. In view of the very extensive literature now existing on extremes from stochastic processes, that ought to be an attainable programme. However, there is a real difficulty with deciding what stochastic models to fit. Along with Joe, I reject the idea of fitting standard time series models such as AR(1); even if the data are transformed to fit the marginal distribution, the exceedance behavior of linear Gaussian processes is too restricted to capture the clustering in the data. Joe's idea of using a first-order (nonlinear) Markov chain is much better, and can be extended to a kth order Markov chain. Jonathan Tawn, Sammy Yuen and I have been looking at these ideas as an extension of the multivariate extreme value theory of Smith, Tawn and Yuen (1989). It represents an interesting alternative approach presenting many new questions of both a theoretical and practical nature.

TABLE 1
Mean cluster size and mean cluster length

Threshold	Cluster interval	Mean size	Mean length
8	72	11.2	59.7
8	24	7.3	22.2
10	72	6.8	37.7
12	72	4.7	27.1

3. MEASUREMENT ERROR AND RECALIBRATION

Fairley mentions a number of practical problems with this kind of data set and implicitly asks whether the analysis can be extended to take account of measurement error. I am not aware of any theoretical treatment of this, though there are certainly other areas in which the problem arises, most notably in oceanography. In what follows, I present a few tentative thoughts on how the problem might be tackled.

Suppose we write $Y = X + \varepsilon$ where X is true ozone level, Y is measured ozone and ε is an independent random error. Following the ideas of the paper, we might assume that the exceedances of X over a high threshold follow the Generalized Pareto distribution, calculate the convolution of that distribution with that of ε, and approximate the resulting distribution of Y by another Generalized Pareto. If it were thus possible to relate the extreme value parameters of X and Y, that would provide a basis for inference about the extremes of X using data on those of Y.

To take this further, we need to know something about how to derive Generalized Pareto approximations. At present there is no established procedure for this, but in Smith (1989) I have made one proposal which is quite simple to understand and which turns out to have good properties. Recalling that the idea is to approximate $\{1 - F(u + y)\}/\{1 - F(u)\}$ by $(1 - ky/\sigma)^{1/k}$ (equation (3.5)), the proposal is simply to equate the first and second derivatives at $y = 0$, in other words to write

$$1/\sigma = f(u)/\{1 - F(u)\},$$

$$(1 - k)/\sigma^2 = f'(u)/\{1 - F(u)\}$$

where F is a distribution function and f the corresponding density. Note that the Generalized Pareto parameters σ and k must be combined with $1 - F(u)$, the probability of exceedance of the threshold u, to obtain a full description of the tail behavior.

Now suppose F is the distribution function of X and G that of Y and suppose $\varepsilon \sim N(0, \sigma^2)$ (the assumption of normal errors is not needed but is convenient for illustration). In the limit $\sigma \longrightarrow 0$ we have the expansion

$$G(u) = F(u) + \frac{\sigma^2}{2} F''(u) + \frac{\sigma^4}{8} F^{iv}(u) + \ldots$$

and we can differentiate this to obtain corresponding expansions for $g(u) = G'(u)$ and $g'(u)$. The assumption that σ is small seems reasonable since if the measurement error were not very much smaller than the variability of X there would be no hope of doing anything worthwhile.

The difficulty with this is that to obtain even the first-order approximations to $g(u)$ and $g'(u)$ we have to go as far as the fourth derivative of F, so there is no direct way of establishing a one-one correspondence between $F(u)$, $f(u)$ and $f'(u)$ on the one hand, and $G(u)$, $g(u)$ and $g'(u)$ on the other. If we knew the full distribution of X (say, normal or lognormal), then it would be possible to use these ideas to relate the extreme value limits of X and Y; but the whole spirit of this paper has been to avoid making parametric assumptions on the whole distribution. I therefore see the need for a more powerful approach to modeling if this problem were to be dealt with.

Fairley and Joe also mention possible biases caused by recalibration. If the approach of the last four paragraphs were adopted, it might be possible to deal with this aspect by varying the distribution of ε to take account of known changes in the calibration procedure. In the absence of that, I could only suggest a deterministic adjustment. I was not previously aware of this problem and do not know how it would affect the Houston analysis.

Finally there is the disturbing question of whether the supposed drift in extreme values could in fact be due to errors in the data collection procedure. Both Fairley and Raftery mention this possibility. In particular, Guttorp, via Raftery, makes a specific suggestion implying that the variance of individual measurements might have decreased over the period of the study.

In Table 2 I have computed standard deviations of the original data, separately for each hour of the day from 10 a.m. up to 6 p.m. (all the high exceedances occurred during these hours), averaged over the years 1974–1980 and 1981–1986, and separately for periods 3 and 4 (dividing the year up into six 61-day periods; I quote only periods 3 and 4 because these are the most interesting for high exceedances). As can be seen, the results support Guttorp's hypothesis.

TABLE 2
Standard deviation for each hour

Hour	Period 3		Period 4	
	1974–1980	1981–1986	1974–1980	1981–1986
10	2.371	2.126	2.113	1.811
11	3.060	2.644	2.750	2.350
12	3.635	2.933	3.382	2.814
13	3.839	3.120	3.920	3.032
14	3.884	3.222	4.226	3.291
15	4.107	3.039	4.352	3.513
16	3.943	2.816	3.734	3.438
17	3.657	2.685	3.562	3.259
18	3.291	2.439	3.079	2.811

What effect might this have on the extreme values? There is no direct way to check, but one might expect the effect to increase the serial correlations, and this might have an effect on probabilities of consecutive high exceedances. The following analysis uses level 16, chosen because, of the levels considered in Table 5 of the paper and in Raftery's discussion, this was the one which gave strongest evidence for a drift. We could try computing the extremal index for each year, as suggested by Weissman, but for this specific problem it seems to me that only the lag-1 correlation is relevant. Therefore I computed: (a) total number of exceedances (not now cluster maxima), and (b) total number of pairs of consecutive exceedances, on the same hourly basis as in Table 2. In each case, the results are expressed as mean number per year, as seen in Table 3.

Perhaps the most meaningful way to summarize Table 3 is in terms of the ratio of total number of pairs of exceedances to total number of exceedances. In Period 3 this ratio is .532 for 1974–1980, and .559 for 1981–1986. The corresponding figures for period 4 are .480 and .548. In each case there is only a very slight increase.

In summary, Table 2 does show that there is a decrease in variance, supporting Guttorp's hypothesis. However, the attempt in Table 3 to see whether this affected the correlation of extreme values was negative. Whether this is indeed the explanation of the trend requires a more detailed examination of the method of data collection, but it is clear that this additional feature of the data needs to be taken into account.

4. METHODS OF INFERENCE

In this paper, I have concentrated on the choice of model rather than the method of inference and have mostly been content with numerical maximum likelihood for the latter. Weissman points out the stability of the procedure with respect to the estimation of

TABLE 3
Exceedances and pairs of exceedances

	Exceedances per year in period 3		Pairs of exceedance per year in period 3		Exceedances per year in period 4		Pairs of exceedances per year in period 4	
Hour	1974–1980	1981–1986	1974–1980	1981–1986	1974–1980	1981–1986	1974–1980	1981–1986
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
11	0.143	0.167	0.000	0.000	0.571	0.167	0.000	0.000
12	0.714	0.667	0.143	0.000	1.000	0.667	0.429	0.167
13	1.000	0.833	0.571	0.667	1.143	1.167	0.857	0.667
14	1.714	0.667	0.571	0.500	1.714	0.833	0.571	0.667
15	1.429	0.667	1.143	0.333	1.571	0.833	1.143	0.667
16	0.714	0.500	0.571	0.333	1.143	0.667	0.714	0.333
17	0.571	0.333	0.429	0.167	0.714	0.833	0.571	0.333
18	0.429	0.333	0.143	0.000	0.571	0.000	0.429	0.000

N-year return level. This is a reassuring feature of the present data, but it is not always the case in this kind of analysis. Davison and Smith (1989) have an example, based on the River Nidd in the North of England, where the precise opposite behavior is observed! There is plenty of evidence that the asymptotic properties of maximum likelihood can be a long way from reality when applied to extreme value distributions (see, for example, Hosking, Wallis and Wood, 1985; Smith and Naylor, 1987), so we should be wary and on the look-out for alternatives.

Singpurwalla, Raftery, and Pickands propose Bayesian alternatives. Singpurwalla proposes this in place of the maximum likelihood analysis of (4.2). In principle I am sympathetic to this (see Smith and Naylor, 1987, where an argument was made for the practical efficacy of Bayesian procedures in problems where the log likelihood is far from quadratic in shape), though in the case of the present data I doubt whether it would make much difference. Also I wonder how realistic it is to try to obtain prior information about the β_j's. Raftery's Bayesianism is of a different form, concentrating on Bayes factors as a means of model selection. Raftery's own work has done much to draw attention to the usefulness of this concept, especially in nonregular problems such as estimation of a change point, and I would be interested to see how far it can be taken in the extreme values context. Pickands' proposal to correct the log likelihood for the number of parameters is in similar spirit, though I am assuming that the reader will mentally adjust for the number of parameters in interpreting the quoted log likelihoods.

I thank Joe for suggesting an alternative initialization procedure in the context of maximum likelihood. Given the convergence difficulties that already exist, this might well be useful.

5. SPATIAL ASPECTS

Fairley argues for the need to consider the spatial distribution of ozone data. Apart from requiring data at several sites, this also involves modeling spatial variability in the analysis. Another field where spatial aspects feature heavily is rainfall data; Reed and Dales (1988) proposed methods of reconciling the point and areal distributions of rainfall extremes and, in work not yet written up, S. Neil and I have proposed an alternative approach based on a model for spatial extremes. Fairley's comments, however, touch on a variety of aspects relating to the spatial distribution of ozone concentrations, and suggest a whole range of both practical and theoretical questions. Perhaps the main point we should be making as statisticians is to emphasize the possible biases that can arise if these effects are ignored, so that scientists and politicians understand the need for extensive sampling.

6. LONG-RANGE DEPENDENCE

Joe refers to work of Hirtzel and Quon which suggests that correlation persists at large time lags (15–20 days) in ozone data, and Raftery mentions his own paper with Haslett which has examined long-range dependence as one of a number of issues in the modeling of wind data, asking in particular what effect this might have on an extreme value analysis.

In the specific context of the ozone data, I have not examined this issue and agree that it is yet another which would repay further study. On the broader question of the effect of long-range dependence on extremes, the only class of models for which the question has been answered explicitly are stationary Gaussian processes, and there the surprising answer is that long-range dependence does not matter unless it is very strong indeed. Specifically, if $\rho_n \log n \to 0$ where $\{\rho_n\}$ is the correlation function, then the asymptotic properties are the same as in independent sequences (Berman, 1964). The cases where $\rho_n \log n \to C$ with $0 < C \leq \infty$ are another matter (Mittal and Ylvisaker, 1975), but the standard long-range dependence models such as fractional Brownian motion and fractional difference processes have polynomially decreasing correlation functions and therefore satisfy Berman's condition. For general stochastic processes the main condition is Leadbetter's Condition D (Leadbetter, Lindgren and Rootzén, 1983), which is essentially a mixing condition far weaker than conditions such as strong mixing or ϕ-mixing used in central limit theory. On the basis of these theoretical results, I think there is some justification for ignoring long-range dependence, but the precise effect of correlations such as those observed by Hirtzel and Quon has not been determined and would be worth exploring.

7. OTHER QUESTIONS

Singpurwalla asks whether there is a connection with the point-process model of Ferguson and Klass. I am not aware of any; the question asked by Ferguson and Klass is a much broader one, namely how to characterize all processes with independent increments lacking a Gaussian component. This indeed leads to a Poisson jump process, but the jumps do not have a parametric distribution. This would make the statistical application of the model appreciably harder.

Singpurwalla also raises the questions of decision making and control. I think he is actually making two points here: first, the possibility of short-term forecasting and control, and secondly, the need to consider more general measures of extreme value behavior.

The first of these points is not entirely germane to the problem under study here, though in other contexts (e.g., when to sound a flood warning) it could be highly relevant. The main work in the existing literature is Lindgren's on the "Slepian" model and its extensions for high-level exceedances of continuous-time processes (e.g., Lindgren, 1985). However, an analysis based on the stochastic properties of the exceedance clusters would represent an entirely different approach and deserves to be explored. Singpurwalla is kind enough to mention my paper with Miller which attempted one version of this, via a state-space model for extreme value distributions, but the class of models we proposed is very restrictive. It is not easy to generalize the class of models without losing the exact Bayesian predictive analysis which Miller and I were able to provide.

The second point is very important and is also implied by Fairley in his comments about the health effects of ozone. Is it sensible to formulate legislation purely in terms of the number of crossings of a fixed high level? I think not. The approach of this paper would lend itself very well to the calculation of other functionals of the point process of high-level exceedances, and I would certainly like to see this aspect explored.

Finally, I agree entirely with the desirability, mentioned by several discussants, of including covariates in the analysis and would see this as a strong area of future development.

8. SUMMARY

The discussion has drawn attention to many aspects not mentioned, or only glossed over, in the paper. I would regard the point-process approach as merely the starting point; once this is accepted as a basic approach, there are many possibilities to extend it to incorporate other features of practical importance. In particular, the discussion has drawn attention to the need for more detailed modeling of short-range time dependence (clustering) and has suggested problems related to spatial aspects.

Fairley's final sentence raises an important point about how statisticians should approach such problems. It is undeniable that, however powerful the statistical technology used to tackle these questions, there remain very real uncertainties about the final conclusions. If statisticians use this as an excuse to avoid the problems, they are leaving the field open to others less qualified to address the real difficulties. My view of the duty of statisticians is that they should answer the questions as best they can, while making clear the inherent uncertainties about their answers. In certain circumstances, this may indeed result in a declaration that no meaningful answer can be given to a certain question.

ADDITIONAL REFERENCES

BERMAN, S. M. (1964). Limit theorems for the maximum term in stationary sequences. *Ann. Math. Statist.* **35** 502–516.

HSING, T. (1987). On the characterization of certain point processes. *Stochastic Process. Appl.* **26** 297–316.

HSING, T., HÜSLER, J. and LEADBETTER, M. R. (1988). On the exceedance point process for a stationary sequence. *Probab. Theory Related Fields* **78** 97–112.

LINDGREN, G. (1985). Optimal prediction of level crossings in Gaussian processes and sequences. *Ann. Probab.* **13** 804–824.

MITTAL, Y. and YLVISAKER, D. (1975). Limit distributions for the maxima of stationary Gaussian processes. *Stochastic Process. Appl.* **3** 1–18.

MORI, T. (1977). Limit distributions of two-dimensional point processes generated by strong mixing sequences. *Yokohama Math. J.* **25** 155–168.

REED, D. W. and DALES, M. Y. (1988). Regional rainfall risk: A study of spatial dependence. IAHR Conference, Birmingham, England.

SMITH, R. L. (1989). Approximations in extreme value theory. Technical Report 205, Center for Stochastic Processes, Univ. North Carolina.

SMITH, R. L. and NAYLOR, J. C. (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Appl. Statist.* **36** 358–369.

SMITH, R. L., TAWN, J. A. and YUEN, H.-K. (1989). Statistics of multivariate extremes. *Internat. Statist. Rev.* To appear.