On the application of a censored log-Normal distribution to partial duration series of storms

P.T. ADAMSON
Department of Environment Affairs, Directorate of Water Affairs, Private Bag X313, Pretoria 0001, South Africa

AND

W. ZUCCHINI
Department of Mathematical Statistics, University of Cape Town, Private Bag, Rondebosch 7700, South Africa

Abstract

A considerable diversity of opinion exists in the hydrological literature as to the relative merits of sampling a random sequence of extreme values either as a partial or annual maximum series. Theoretical aspects of the two approaches are considered and some practical criteria for model selection are given. The usual partial series model is to combine a Poisson arrival rate of exceedances with the assumption that these are exponentially distributed. The partial series are described as a censored process and in addition deal with the case of Negative Binomial arrivals. Attention is given to the distribution of storms in such a situation to be drawn from a censored log-Normal model and certain advantages are illustrated. The relative flexibility of the partial series approach is also illustrated. One hundred one day storm sequences from South Africa and Namibia are investigated and finally explicit algorithms are given with which to estimate the parameters and compute the percentiles of the censored log-Normal model.

Introduction

The search for a probability distribution that adequately describes the random behaviour of a given sample of extreme values to the joint satisfaction of the pure statistician on the one hand and the applied scientist on the other is an endeavour that has spanned a massive body of literature which itself continues to unify and generate a great variety of interesting topics. In the field of asymptotic extreme value theory the major work is still that of Gumbel (1958), although Galambos (1978) has brought together in a single text many of the more recent theoretical advances. The applied scientist in his assessment of the risk of extremes in hydrology and meteorology has not relied totally on asymptotic theory but has often turned to direct model fitting. Obvious amongst such models are, for example, the log-Normal and log-Pearson type III whose density functions satisfy the positive skew and heavy right tails characteristic of samples of annual flood peak or rainfall maxima. A number of contemporary studies have extended direct model fitting to compound distributions in cases where, for example, floods represent a mixture of generating processes (Houghton, 1978; Sing and Sinclair, 1972; Canfield *et al.*, 1980).

Standard hydrological practice has been to estimate the magnitude of the event with a mean recurrence interval expressed in years. This is then defined as "the T year event" and is usually computed from a sample of annual maxima. The alternative to sampling annual maxima is known as the partial duration series in which all events above a prescribed level are selected.

Usually the exponential model of partial duration series is assumed. However, if a model is proposed for the complete process, for example all floods or rainfall depths, then it can in a truncated or censored form be applied to the analysis of the extremes or those events greater than the selection level. In other words, an assumption about the complete distribution is used in the analysis of the upper tail.

In the present work the theoretical aspects of these approaches to the analysis of partial duration series are explored and some empirical comparisons made with alternative probabilistic models and sampling procedures.

Theoretical Background

On the definition of the "T year event"

The term "T year event" can be used to mean one of two different events. It can either be defined in the annual maximum sense or in what we shall call here the naive sense. For the annual maximum case let Q_a be the random variable which describes the biggest event, say a storm rainfall depth, within a particular year and let F_a be its distribution function, that is:

$$F_a(q) = \text{Probability}\{Q_a \leq q\} \quad (1)$$

The T year event, $q_a(T)$, in the annual maximum sense is defined as the solution to the equation.

$$F_a(q_a(T)) = 1 - 1/T, \qquad \text{i.e.} \quad (2)$$

$$q_a(T) = F_a^{-1}(1 - 1/T) \quad (3)$$

This event has, by definition, the following relative frequency interpretation: Suppose that storms are observed for n years and that $n_a(T)$ is the *number of years* during which the maximum storm exceeds $q_a(T)$, then $\lim_{n\to\infty} n/n_a(T) = T$. Note that there will be years during which more than one storm exceeds $q_a(T)$, but even for such years $n_a(T)$ is only incremented by one. In other words, on average in one year out of T there is *at least one* storm which exceeds $q_a(T)$.

For the definition of the T year event in the naive sense let Q_o be the random variable which describes a storm depth which

exceeds a given quantity, $q_o \geq 0$, and let F_o be its distribution function, that is:

$$F_o(q) = \text{Probability}\{Q_o \leq q\} \quad , \quad q > q_o \tag{4}$$

Suppose that there are an average of λ_o storms per year. Then the T year event, $q_o(T)$, in the naive sense is defined as the solution to the equation:

$$F_o(q_o(T)) = 1 - 1/(T\lambda_o) \quad , \text{ i.e.} \tag{5}$$

$$q_o(T) = F_o^{-1}(1 - 1/(T\lambda_o)) \tag{6}$$

The relative frequency interpretation of $q_o(T)$ is different to that of $q_a(T)$, being: Suppose that storms are observed for n years and that $n_o(T)$ is the *number of storms* which occur, then $\lim_{n\to\infty} n/n_o(T) = T$.

In this case *every* storm is counted. In other words: On average, once in every T years there will be a storm which exceeds $q_o(T)$.

Which of the two events $q_a(T)$ or $q_o(T)$ is to be preferred in practice should depend on the use to which it is to be put. For the type of distributions that are usually considered one has that

$$\lim_{T\to\infty} q_a(T)/q_o(T) = 1,$$

and so for sufficiently large T these two events are equivalent.

If one works with partial duration series it is simpler to derive formulae for $q_o(T)$ rather than for $q_a(T)$ and so $q_o(T)$ would be a more convenient definition of "T year event". However, it is standard hydrological practice to regard $q_a(T)$ as the "T year event". It is therefore necessary to derive the theoretical relationship between $q_o(T)$ and $q_a(T)$. The main result on which the relationship is based is given in (7a) and (7b) below. It is proved, e.g. in Todorovic and Zelenhasic (1970).

On some theoretical results relating to $q_a(T)$ and $q_o(T)$

Assuming that storms occur independently, define $p_o(m)$ as the probability function of the number of storms in any year which exceed q_o, then

$$Q_a \leq q_o \qquad \text{with probability } p_o(0) \tag{7a}$$

$$F_a(q) = \sum_{m=0}^{\infty} p_o(m)\, F_o(q)^m \qquad \text{for } q > q_o \tag{7b}$$

A number of results can be derived from (7) assuming that all random variables are independently distributed. Firstly, relationships between $F_a(q)$ and $F_o(q)$ can be established. If the process describing the arrival of storms is Poisson, i.e.

$$p_o(m) = \lambda_o^m \exp(-\lambda_o)/m! \qquad m = 0, 1, 2, \ldots$$

then (7b) becomes:

$$F_a(q) = \sum_{m=0}^{\infty} \lambda_o^m \exp(-\lambda_o)\, F_o(q)^m/m! \quad , \quad q > q_o$$

This sum converges and can be evaluated to:

$$F_a(q) = \exp\{-\lambda_o(1 - F_o(q))\} \quad , \quad q > q_o \tag{8}$$

The Poisson distribution has the property that the mean is equal to the variance. In some practical applications this condition is not met; in particular the variance can be much larger than the mean. For such cases the negative Binomial distribution can be used as an alternative to the Poisson, (see, e.g. Calenda *et al.* (1977)): i.e.

$$p_o(m) = \left(\frac{1}{1 + e_o\lambda_o}\right)^{1/e_o} \left(\frac{e_o\lambda_o}{1 + e_o\lambda_o}\right)^m \frac{1}{m!} \prod_{i=0}^{m-1} \left(\frac{1}{e_o} + i\right)$$

for which (7b) becomes:

$$F_a(q) = \{(1 - \alpha_o)/(1 - \alpha_o F_o(q))\}^{1/e_o} \quad , \quad q > q_o \tag{9}$$

where $e_o = \{\text{Var}(m) - E(m)\}/E(m)^2$

and

$$\alpha_o = e_o\lambda_o/(1 + e_o\lambda_o).$$

Having related the distribution functions of storm depths in the annual maximum and naive senses one can now also relate, for a given storm depth, the return period, T, in the annual maximum sense and the return period, T_o, in the naive sense.

From (3) and (6):

$$q_a(T) = q_o(T_o) \Rightarrow F_a^{-1}(1 - 1/T) = F_o^{-1}(1 - 1/(T_o\lambda_o))$$

for which it follows that

$$T = \left[1 - F_a\{F_o^{-1}(1 - 1/(T_o\lambda_o))\}\right]^{-1} \tag{10}$$

For Poisson arrivals then, and recalling (8) we have

$$\begin{aligned} F_a\{F_o^{-1}(1 - 1/(T_o\lambda_o))\} &= \exp\{-\lambda_o(1 - F_o[F_o^{-1}(1 - 1/(T_o\lambda_o))])\} \\ &= \exp\{-1/T_o\} \end{aligned}$$

Equation (10) therefore becomes:

$$T = [1 - \exp\{-1/T_o\}]^{-1} \tag{11a}$$

Consequently T_o in terms of T, for Poisson arrivals becomes:

$$T_o = \{\ln(T) - \ln(T - 1)\}^{-1} \tag{11b}$$

Turning now to a consideration of negative Binomial arrivals, $F_a(q)$ is given in (9), hence:

$$\begin{aligned} &F_a\{F_o^{-1}(1 - 1/(T_o\lambda_o))\} \\ &= \{(1 - \alpha_o)/(1 - \alpha_o F_o[F_o^{-1}(1 - 1/(T_o\lambda_o))])\}^{1/e_o} \\ &= \{(1 - \alpha_o)/(1 - \alpha_o(1 - 1/(T_o\lambda_o)))\}^{1/e_o} \\ &= (1 + e_o/T_o)^{-1/e_o} \end{aligned}$$

where α_o has been replaced by $(\lambda_o e_o/(1 + \lambda_o e_o))$.

Equation (10) then becomes:

$$T = \{1 - (1 - e_o/T_o)^{-1/e_o}\}^{-1} \tag{12a}$$

Consequently T_o in terms of T for negative Binomial arrivals is:

$$T_o = \{e_o/[(1 - 1/T)^{-e_o} - 1]\} \tag{12b}$$

On the distribution of exceedances

So far no assumptions have been made about the distribution of the exceedances of q_o. The above results hold for all F_o. We now consider two cases: the exponential and censored log-Normal distribution and derive from these the T year event in the annual maximum sense when the process of arrivals is Poisson. This follows the usual hydrological practice of expressing $q_o(T)$ in terms of $q_a(T)$.

Exponentially distributed exceedances

The exponential probability function is defined by:

$$F_o(q) = 1 - \exp\{-(q - q_o)/\beta\} \quad , \quad q > q_o \tag{13}$$

We note that the scale parameter β does not have a subscript since for any truncation level $q_o' > q_o$ the exceedances are also exponentially distributed with the same parameter β.

Substituting (13) in (8) results in:

$$F_a(q) = \exp\{-\lambda_o \exp(-(q - q_o)/\beta)\} \ , \ q > q_o \ , \text{ or} \tag{14a}$$

$$F_a(q) = \exp\{-\exp(-(q - q_o - \beta \ln\lambda_o)/\beta)\} \ , \ q > q_o \tag{14b}$$

which we see to be the Gumbel or Extreme Value Type I distribution with parameters $(q_o + \beta\ln\lambda_o)$ and β. Using (2) an event can now be expressed in the annual maximum sense as

$$q_a(T) = q_o + \beta(\ln\lambda_o + y(T)) \tag{15}$$

where $y(T) = -\ln[-\ln(1 - 1/T)]$.

Censored log-Normal exceedances.

In this case

$$F_o(q) = \frac{\Phi((\ln(q) - \mu)/\sigma) - \Phi(\xi_o)}{1 - \Phi(\xi_o)} \quad , \ q > q_o \tag{16}$$

where Φ is the distribution function of the standard normal random variable.

For the same reasons as given above the parameters μ and σ are not subscripted but ξ_o is because of its dependence on q_o.

Substituting (16) in (8) results in:

$$F_a(q) = \exp\left\{-\lambda_o \frac{\Phi((\ln(q) - \mu)/\sigma) - \Phi(\xi_o)}{1 - \Phi(\xi_o)}\right\} \ , q > q_o \tag{17}$$

The T year event in the annual maximum sense can now be found by substituting (17) in equation (2). From this one obtains:

$$\Phi((\ln q_a(T) - \mu)/\sigma) = (1 - \Phi(\xi_o)) \ln(1 - 1/T)/\lambda_o + 1$$

$$= p, \text{ say.}$$

Hence $$q_a(T) = \exp\{Z(p)\sigma + \mu\} \tag{18}$$

where $Z(p)$ is the standard normal deviate corresponding to $Z = \Phi^{-1}$ and efficiently approximated by the functional approximation given in Abramovitz and Stegun (1972, p 933) which is given in the Appendix.

The Poisson assumption for partial duration series

The above theoretical results for partial duration series with exceedances distributed either exponentially or as censored log-Normal have assumed that the arrival process for such events is Poisson. Under the Poisson assumption the number of events $q > q_o$ is considered a random variable with mean and variance λ_o. It is not a necessary condition that the mean rate of arrival be constant within the year so long as whole years or multiples of whole years are considered. Following Cunnane (1979) it is seen that if the number of occurrences in successive years, designated $M_1, M_2, \ldots, M_N$ have a mean $\bar{M}$ which is the estimate of λ_o, then as λ_o becomes larger ($\lambda_o > 5$) the M_i become normally distributed as $N(\lambda_o, \lambda_o^{1/2})$. Consequently:

$$(M_i - \lambda_o)/\lambda_o^{1/2} \sim N(0,1) \quad , \text{ and} \tag{19}$$

$$\sum_{i=1}^{N} (M_i - \lambda_o)^2/\lambda_o \sim \chi^2_N \tag{20}$$

The so-called Fisher dispersion test statistic, d, is obtained by replacing λ_o by its estimate $\bar{M}$. Its distribution is still χ^2 but with one less degree of freedom:

$$d = \sum_{i=1}^{N} (M_i - \bar{M})^2/\bar{M} \sim \chi^2_{N-1} \tag{21}$$

This statistic provides a means of statistically assessing whether the number of exceedance events within a fixed time interval is Poisson distributed.

Deviations from the Poisson distribution do occur as has been shown in a number of hydrological studies (see, e.g. Cunnane, 1979). In particular the sample variance can be much larger than the sample mean, i.e. the observed distribution is more dispersed than is expected under the Poisson distribution. As already mentioned such cases may be modelled using the negative Binomial distribution. Formulae analogous to those given in equations (13) to (18) can be derived for the negative Binomial case without any essential difficulties.

On the application of partial duration series to n-day storm depth

Preliminary

For the estimation of extreme storm risk the partial duration series is easier to apply than it is to flood peak analysis. No objective definition of a flood exists although one could perhaps use "bankfull" discharge. For rainfall, however, an event can be defined as that for which precipitation depth exceeds a given value and in the following study of n-day point rainfall extremes we define a "rain day" as one during which 0,25 mm or more is recorded.

With the process defined it is now necessary to choose a truncation value q_o. In order to preserve an equal range of data for partial duration and annual maximum series for comparative studies, q_o is chosen as the minimum annual maximum event. Consequently all n-day storm rainfalls above this value are initially chosen for analysis. However, to attempt to ensure independence between selected events all $q > q_o$ are removed when the gap between them is less than 4 days.

This somewhat arbitrary gap is based on a study of weather cycle lengths by Gabriel and Neumann (1957) who found that the modal length of dry/wet periods for Tel Aviv cannot be less than 3 days and represents the average time of passage of the

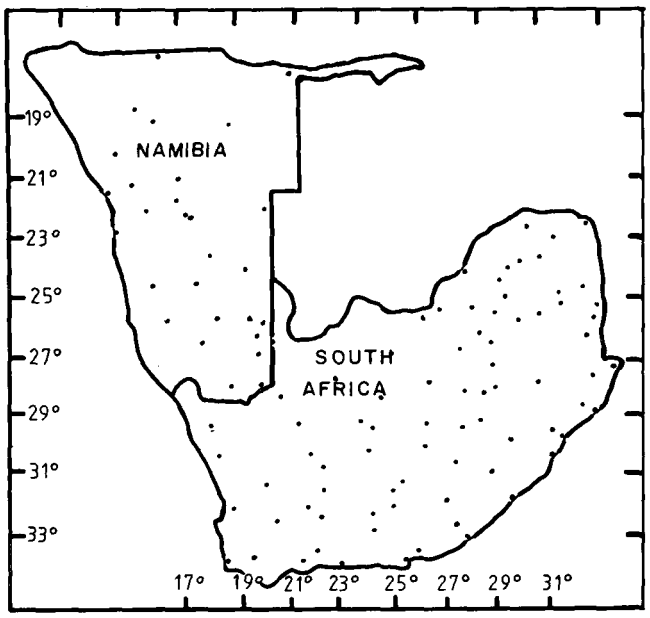

Figure 1
Location of 100 daily rainfall stations used in comparative studies.

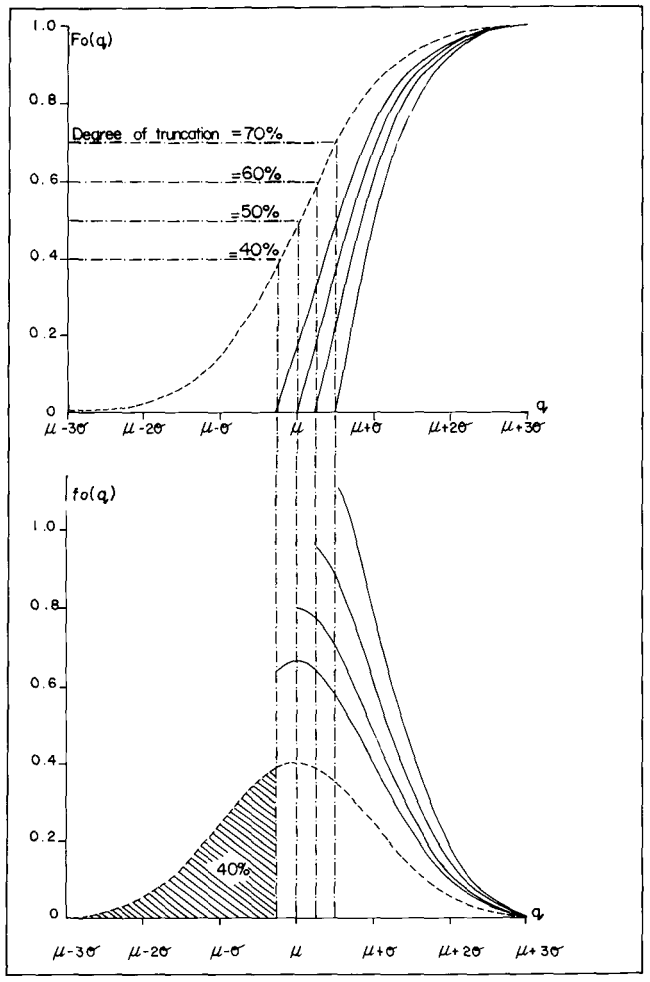

Figure 2
Distribution function and density function for various degrees of truncation of a normal probability model

weather generating process. To ensure that the assumption of independence is not grossly violated one can carry out a test of the hypothesis that the serial correlation for the series is zero. A suitable test is given in Box and Jenkins (1976).

One hundred rainfall stations were chosen for the assessment of the partial duration series approach to storm risk estimation. As far as possible the various climatic regions of South Africa and Namibia are represented and the location of the gauges is shown in Figure 1.

Maximum likelihood estimation of parameters

Exponentially distributed exceedances

If $q_1, q_2, \ldots, q_n$ are independent random variables each having the distribution function (13) and with q_o assumed known then the maximum likelihood estimator of β is: (Johnson and Kotz 1970).

$$\hat{\beta} = n^{-1} \sum_{i=1}^{n} (q_i - q_o) = \bar{q} - q_o \qquad (22)$$

Censored log-Normal exceedances

Since for the study of rainfall exceedances it is simple to establish the number of events less than q_o, (n_2), and the number greater than q_o, ($n_1 = n - n_2$), then the maximum likelihood estimators for μ and σ given by Cohen (1950) are appropriate:

$$[1 - \hat{\xi}_o (Y(\hat{\xi}_o) - \hat{\xi}_o)]/(Y(\hat{\xi}_o) - \hat{\xi}_o)^2 - v_2/v_1^2 = 0 \qquad (23)$$

$$\sigma = v_1/(Y(\hat{\xi}_o) - \hat{\xi}_o) \quad \text{and} \quad \hat{\mu} = q_o - \hat{\sigma}\hat{\xi}_o \qquad (24)$$

where $Y(\xi_o) = \{n_2 \phi(\xi_o)\}/\{n_1 \Phi(\xi_o)\}$ (25)

and $$v_K = \sum_{i=1}^{n_1} (q_i - q_o)^K / n_1 \qquad (26)$$

where $\phi(.)$ is the density of a standard random normal deviate. It should be noted that this system of equations has two solutions only one of which is correct, but the algorithm given in the Appendix converges to the correct one for all practical values of n_1 and n_2. Since for rainfall data n_1 can be found the parameter estimation procedure is for a censored model. In the truncated case no information is available for the process below q_o and the alternative maximum likelihood estimators for μ and σ are to be found in Cohen (1950) or Gupta (1952). Figure 2 shows the effect of the degree of truncation (or censoring) on the density and distribution function of the Normal model.

On the validity of the Poisson assumption

For the 100 sites shown in Figure 1 the validity of the Poisson

TABLE 1
PERCENTAGE NUMBER OF TIMES THAT THE POISSON ASSUMPTION FOR ANNUAL DISTRIBUTION OF THE EXCEEDANCES OF q_o WAS REJECTED FOR n-day RAINFALL DEPTHS (100 stations)

Level of significance	Duration (days)			
	1	2	3	7
> 10	55	29	17	2
> 05	50	23	11	0
> 01	34	18	7	0

assumption for the arrival of the exceedances of q_0 for 1, 2, 3, and 7 day rainfall totals was tested using the Fisher dispersion statistic (21). The value of χ^2 was estimated using the well-known Wilson-Hilferty approximation. The results are shown in Table 1, with q_0 the minimum annual maximum n-day rainfall.

It is apparent that the assumption becomes dubious as the duration of storms decreases. For one day storms we don't have a Poisson process. This may be partially explained by the following facts:

Firstly, selecting q_0 as the minimum annual maximum results in every year having at least one storm, i.e. the sample distribution is truncated at zero. Secondly $\bar{M}$ was often less than 5 which invalidates the approximations on which the critical values of the test are based. Thirdly there is a distortion due to missing observations. Fourthly the numbers of storms at the different stations are not independently distributed.

However, it is felt that the poorness of the Poisson model of storm arrivals for durations of one day is not entirely explained by these considerations and further causes are possibly evident. This question is currently being investigated and preliminary results indicate that even when the above distortions are removed the process of storm arrivals in still not Poisson. This implies that at least one of the conditions which are sufficient for the process to be Poisson is not met.

Firstly the function which determines the rate of arrivals is not periodic, that is the weather changes. It appears that this function is itself random rather than deterministic. Secondly the assumption of independence is violated because the dependence between storms can last for more than the four day period inserted between the selected storms. A third possibility is that such dependence is introduced by some large scale and slow moving weather generating process, sea temperature for example. Finally the four day gap may itself be the problem because even under the Poisson assumptions the gap between storms can be smaller than 4 days. However, very few events were removed ($^1/_2$%) although there was a tendency for more removals for those stations where the deviation from the Poisson distribution was most marked.

The fact that the 7-day storm totals fitted the Poisson process of arrivals improbably well gave some cause for concern, but checks revealed that the counts given in Table 1 are correct. The fact that numbers of storms at the different stations are not independently distributed makes it difficult to establish precisely how unlikely such a result may be.

Cunnane (1979), in an analysis of British flood peak data, found that the departure from the Poisson assumption is in the direction of the variance being significantly greater than the mean. For the considerable majority of the daily rainfall records analysed herein this is also the case. Consequently, since for negative Binomial arrivals $E(m) < V(m)$, it may be ventured that this may provide a more suitable model for the annual distribution of the exceedances of q_0. Figure 3 shows two stations for which the Poisson assumption was rejected at the 5% level. The parameters of the negative Binomial distribution function were computed using the maximum likelihood procedure given in Johnson and Kotz (1969, p 132) and the values of e_0 so obtained varied from e_0 = 0,015 at East London to e_0 = 0,73 at Walvis Bay. We note that as $e_0 \rightarrow 0$ the Binomial approaches the Poisson distribution and departures from zero for the parameter e_0 were almost totally explained by the aridity of the climate. This implies that V(m) tends to increase relative to E(m) as mean annual rainfall decreases, which is not a particularly surprising result. Thus the assumption of negative Binomial arrivals would have advantage over that of a Poisson process only in the more arid regions. The Pretoria histogramme (Figure 3) appears suspiciously bimodal which suggests that the rate of arrival of storms could vary from year to year.

The relationship between T_0, the recurrence interval in the partial duration series sense and T_a the recurrence interval in the annual maximum sense has been derived (Equations 11b and 12b). Figure 4 shows these relationships graphically and illustrates that the interpretation of risk for a particular event only

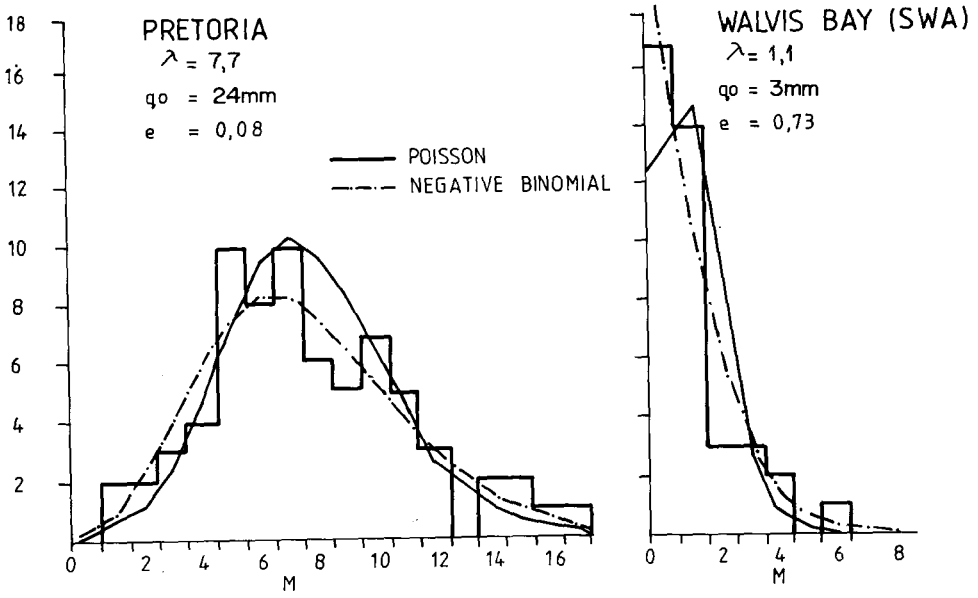

Figure 3
Empirical fit of Poisson and Negative Binomial models to storm arrival rates at a truncation level equal to the minimum annual maximum recorded historically.

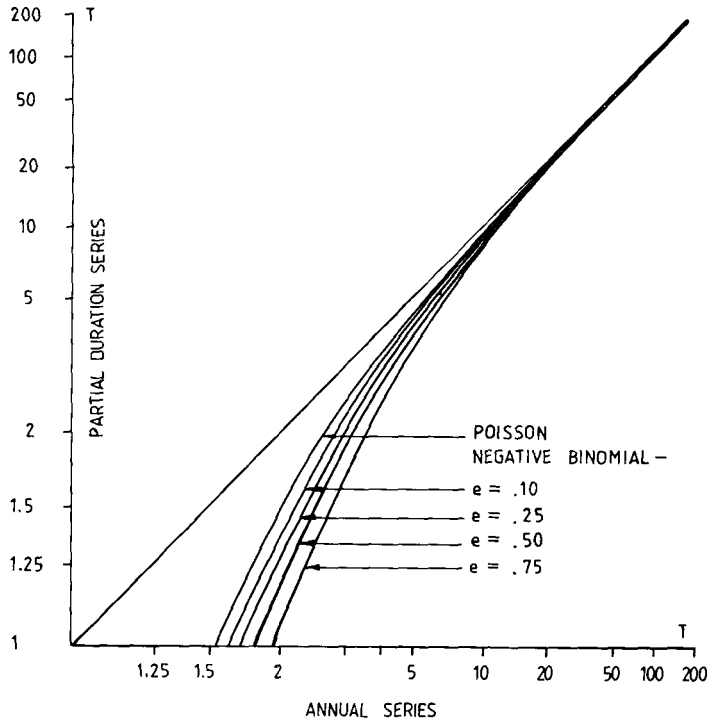

Figure 4
Relationship between recurrence interval (T) for events sampled from partial and annual Series and arrival rates assumed Poisson and Negative Binomial.

becomes significantly different at recurrence intervals of 5 years or less. Where design criteria are not therefore affected by relatively minor events the Poisson assumption for the evaluation of storm risk may be not unduly restrictive.

On the selection of q_o

The truncation level q_o should be selected with care because in practice one has to fit F_o. Although it is only necessary to fit the tail of a truncated or censored distribution, q_o should not be set too high otherwise there will be insufficient data points for the efficient estimation of F_o. The accuracy of the estimation of the required distribution could be improved by capitalizing on the fact that for any truncation level greater than q_o, the exceedances will also have the required distribution. Consequently q_o could be kept as low as reasonably possible and the resulting sample used to estimate the parameters of the exceedance distribution. The truncation level could then be raised to meet some of the other assumptions which would otherwise be violated if q_o were too low. However, one should be careful to balance improved accuracy of estimation with the need for the model to fit over the domain of real interest, that is the tail.

In the present study the selection of q_o as equal to the minimum annual maximum n-day rainfall for the 100 stations gave a mean value of 4,4 one day storms per year. The South African stations have a minimum record length of 40 years, those for South West Africa/Namibia, 20 years. A comparison of the value of T for a given estimate of q_a using the partial duration series where λ_o = E(m) and the annual *exceedance* series where λ_o = 1 (i.e. the n′ highest storm depths are taken where n′ is the number of years of record) is shown in Figure 5. The results are averaged for all 100 stations and show that the censored log-Normal model is reasonably robust towards the value of λ_o, whilst the exponential model consistently gave a higher estimate of T for each q_a when λ_o = 1.

The above procedure has the advantage that it is easy to automate which is essential if one is fitting models to a large number of stations. If, however, one is fitting models to only a few stations then it is feasible to select the appropriate truncation value more carefully. One could plot all the rainfall depths on suitable linearised graph paper and select as censoring point the smallest value for which the fit reasonably conforms to the lognormal distribution.

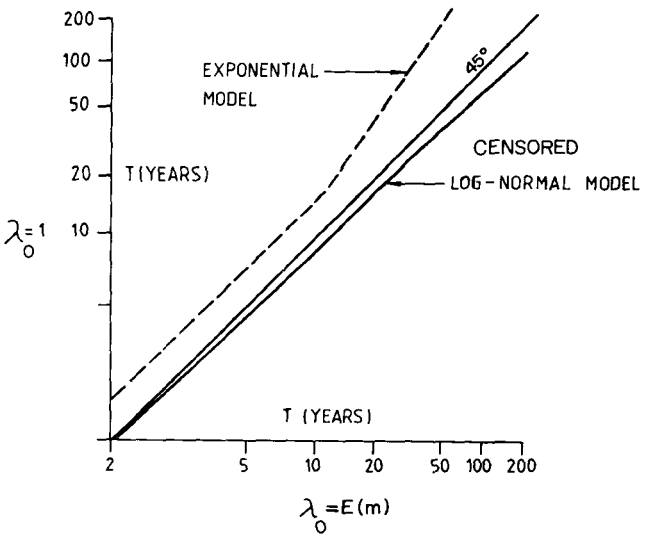

Figure 5
Mean relationship (100 stations) between recurrence interval estimates from partial series assuming censored log-Normal and exponentially distributed exceedances with λ_o = 1 and λ_o = E(m).

Comparative studies – annual maximum and partial duration series

The following empirical comparison of results in the estimation of extreme storm risk is confined to the more practical aspects of model choice. Since there is no theoretically sound reason for choosing any model above another, one is forced, in a practical situation, to assess the competing models in terms of empirical measures of fit, viability of assumptions, computational ease and accuracy and consistency of estimate with varying sample sizes.

For Southern African storm rainfall a number of estimation problems were encountered with some well-known models of annual maxima. For example, it was found that as a general model the log-Pearson Type III distribution is unsuitable because in almost half of the 100 cases considered the skew of the logarithms of the data is negative and the distribution therefore has an upper bound. This becomes a serious limitation when, as occurred, the bound is lower than the maximum observed events. Further problems were encountered with the General Extreme Value distribution with respect to the switch in boundedness depending on whether the data is distributed as Type II or Type III. Meteorological consideration apart, the existence of an upper or lower bound is purely a function of the skew of the data and given the large sampling variance of this moment, intolerable local variations in $\hat{q}_a(T)$ arose. In fact for annual maxima the use of extreme value theory is not automatically justified and even in the independent case convergence can be painfully slow.

Figures 6 and 7 show the performance of selected annual and partial duration series models in terms of comparative mean magnitudes of $\hat{q}_a(T)$ and in the case of annual models only an empirical measure of fit. The log-Normal 2 and 3 parameter, Gumbel and log-Gumbel models were fitted to the 100 series of annual 1 day rainfall maxima by the maximum likelihood procedures given in Kite (1977). The Box-Cox transformation to normality is detailed in Chander *et al* (1978) and the fitting procedure for the mixture of two log-Normal models is given in Sing and Sinclair (1972). The parameters of the exponential and censored log-Normal models of the partial duration series are estimated by the maximum likelihood procedures given herein. (Equations 23 and 24).

The most obvious feature of Figure 6 is the systematic and considerable comparative overestimation of $q_a(T)$ by the log-Gumbel model. The reason for this is the fact that the inherent assumption of a linear relationship between the reduced variate and the log-log $(T/T - 1)$ function of recurrence interval does not hold in log space. This error is considerably increased on exponentiation back into real space (Boughton, 1980; Pitman, 1980).

By virtue of being fitted to the data by a constrained optimisation procedure to minimize deviations from the empirical Weibull plotting position, the mixture of two log-Normal models provides a yardstick with which to assess the performance of the competing models of annual maxima, that is if fit is to be the sole criterion of assessment. If this is so then the 3 parameter log-Normal and Box-Cox transformation to normality would appear to provide the best alternative estimates of $q_a(T)$. This is not surprising since both have an extra parameter over the log-Normal and Gumbel models. However, the tail behaviour in the case of the Box-Cox transformation may suggest overfitting and extrapolation beyond the range of the data may not be wise.

Of the two partial duration series models the censored log-Normal provides a mean estimate of $q_a(T)$ comparable to the best fitting annual models. The exponential model appears far less satisfactory. An investigation into the reasons for this confirmed

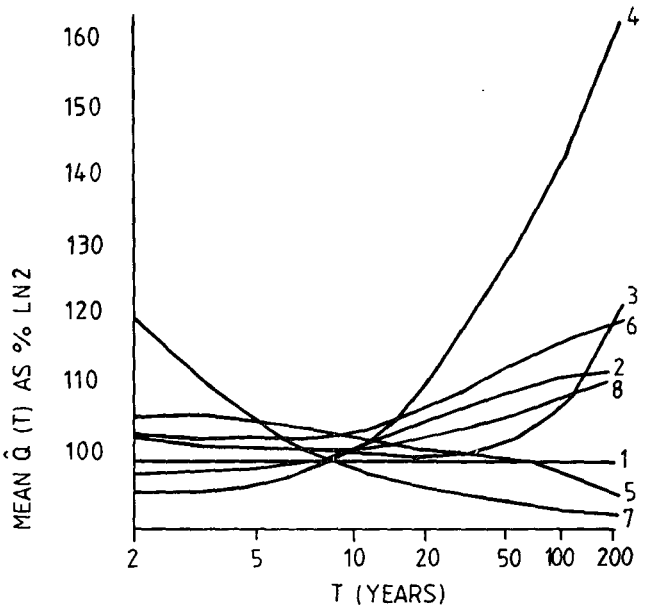

Figure 6
Mean estimate of Q(T) with 2-parameter for log-normal model of annual series as standard.
1. 2 Parameter log-Normal model of annual series.
2. censored log-Normal of partial series.
3. Box-Cox transformation to normality, annual series.
4. Log Gumbel model of annual series.
5. Gumbel model of annual series.
6. Log Pearson Type III model of annual series.
7. Exponential model of partial series.
8. 3-Parameter log-Normal model of annual series.

the result found for British flood data (NERC, 1975), namely that estimation of the scale parameter β is dominated by values near to the threshold itself rather than by the larger values. Consequently β is underestimated since it represents the mean amount by which the selected values greater than $\hat{q}_o$ exceed q_o.

An empirical measure of the goodness of fit of the annual models can be achieved by calculating the mean deviation between the estimated and observed events for the same probability of non-exceedance. This is then expressed as a percentage of the mean annual event.

The mean deviation may thus be expressed as (Prasad, 1970).

$$D = (100/K) \sum_{i=1}^{K} | (F_C - F_A)/F_m | \qquad (27)$$

where F_C, F_A and F_m represent the estimated, observed and mean events respectively and $K = 100$. F_A is calculated from the empirical Weibull plotting position and the results were averaged over 5% increments of non-exceedence. Figure 7 shows that the log-Gumbel model provides the worst fit on the average and the mixture of two log-normal distributions the best by virtue of the fitting procedure used. There seems little to choose between the other annual models although the criterion conveys little insight into their tail behaviour since there are relatively few observed events with $P > 99\%$.

A final and essentially pragmatic test of model performance is to consider the variance of the estimates of $q_a(T)$ in an obviously climatologically homogenous region. For this purpose 20 daily rainfall records with $N \geq 40$ years and within Pretoria municipal boundary were chosen. Record lengths varied from 43 to 71 years, q_o (minimum annual maximum one day rainfall) from 18 to 34 mm and the maximum recorded event from 132 to 207 mm.

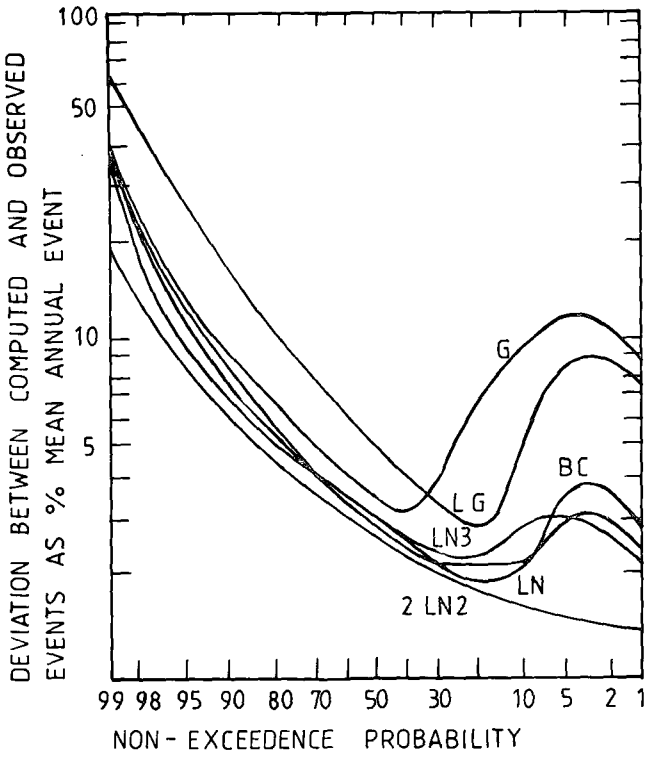

Figure 7
Mean deviation (100 stations) between computed and observed events expressed as a percentage of the mean annual event for annual models. (G: Gumbel, LG: Log-Gumbel, BC: Box-Cox tranformation to normality, LN3: 3-parameter log-Normal, LN: log-Normal, 2LN2: mixture of 2 log-Normal models).

In terms of the coefficient of variation of $\hat{q}_a(T)$ for various T Figure 8 reveals that both partial duration series models provide the greatest uniformity of estimate. Model inflexibility in the case of the two parameter models of annual series no doubt accounts for their providing greater consistency of estimate than the three parameter models. It may not, however, be strictly legitimate to make such comparisons as these or indeed employ tests of fit in order to assess models with different numbers of parameters.

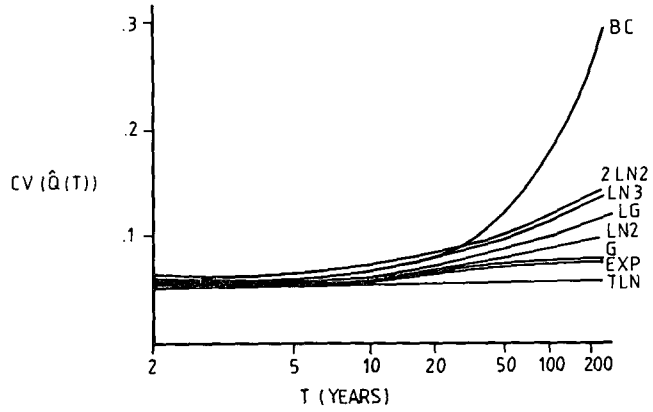

Figure 8
Coefficients of variation of 20 estimates of the T year daily event in an assumed homogenous storm region in Pretoria urban area from various annual and partial series models. (EXP: Exponential model of partial series, TLN: Censored log-Normal model of partial series).

Discussion

If the substantiation of the hypothesis that a partial duration series contains more information on extremes than the annual series is accepted then attention should be directed towards a comparison of the statistical efficiency of partial series. Using theoretical and experimental arguments both Yevjevich and Taesombut (1978) and Cunnane (1973) have concluded that for $\lambda_o > 1{,}5$ the asymptotic sampling variance of $\hat{q}_a(T)$ is much lower for a partial duration than for the corresponding annual series. They confined their attention to exponential models but one would suspect that for log-Normal models the figure may be lower based on the following argument. For an exponential model of annual series one estimates only one parameter (β) and for the partial series, two (β, λ_o). Therefore the ratio of parameters is 1:2. For the log-Normal case the ratio is 2:3, that is μ and σ against μ, σ and λ_o. The relative variation in $\hat{q}_a(T)$ when parameters are increased from one to two would be higher than that when increased from two to three. However, this point would require experimental substantiation.

Situations exist where the number of observations less than q_o is actually known. This information can be used to great effect, particularly in the analysis of extreme daily rainfalls, where the degree of truncation is very high (± 90%). Cohen (1950) has shown that the variance formulas for σ and ξ, when the number of unmeasured observations (n_2) is known may be given by:

$$V(\hat{\sigma}) = \sigma^2 W(\xi)/n_1 \qquad (28a)$$

$$V(\hat{\xi}) = w(\xi)/n_1 \qquad (28b)$$

and where n_2 is unknown by:

$$V(\hat{\sigma}) = \sigma^2 W^*(\xi)/n_1 \text{ , and} \qquad (29a)$$

$$V(\hat{\xi}) = w^*(\xi)/n_1 \qquad (29b)$$

The explicit forms of the weighting functions W, w, W* and w* are given in Cohen (1950) and we will not repeat them here. Figure 9 shows a plot of these functions and illustrates that even for modest degrees of truncation by hydrometeorological standards the increased efficiency in parameter estimation by virtue of knowledge of n_2 is quite substantial. This gain in efficiency is illustrated for three South African one day rainfall records in Table 2.

TABLE 2
COMPARATIVE VARIANCES OF $\hat{\sigma}$ AND $\hat{\xi}$ FOR A LOG-NORMAL MODEL WITH n_2 KNOWN AND UNKNOWN

	Pretoria 513/404	Grahamstown 57/048A	Carnarvon 166/238A
n_1	536	494	172
n_2	5176	8072	1134
$\Phi(\xi)$	0,896	0,939	0,838
$\hat{\xi}$	1,260 4	1,550 1	1,028 7
$\hat{\sigma}$	0,871 3	1,014 7	0,808 3
$V(\hat{\sigma})$ n_2 known	0,001 48	0,001 6	0,002 6
$V(\hat{\xi})$ n_2 known	0,001 40	0,000 6	0,002 1
$V(\hat{\sigma})$ n_2 unknown	0,023 37	0,037 5	0,057 4
$V(\hat{\xi})$ n_2 unknown	0,205 2	0,245 0	0,423 3

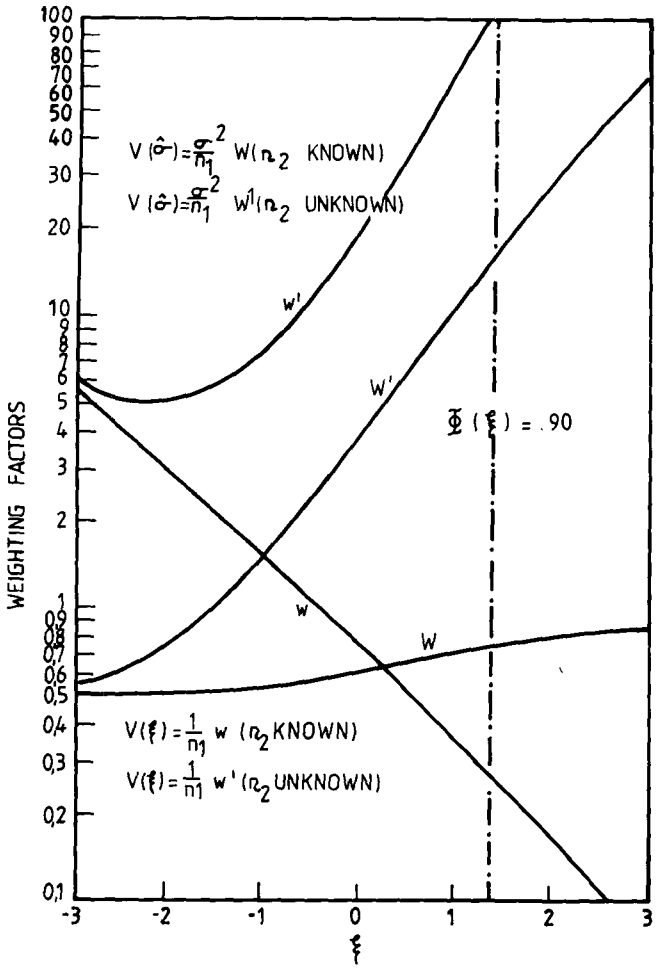

Figure 9
Weighting factors for estimates of parameter variance in a censored log-Normal model of partial series with n_2 known or unknown (after Cohen, 1950).

Although it would be difficult to theoretically justify the use of any probabalistic model of a partial duration series, parameter uncertainty can be reduced if the censored log-Normal model is used. In the analysis of annual maxima it is precisely this aspect of parameter uncertainty that is attracting most attention in contemporary studies with increasingly complex estimation procedures being recommended, based on an ever broadening base of theoretical argument.

Partial duration series meanwhile are receiving nothing like the same attention yet this approach offers a far more flexible tool than the simple assessment of risk in the annual maximum sense. Engineers and hydrologists appear tied to the concept of recurrence and ignore a considerable number of alternative and potentially far more useful means of risk evaluation. For example, let $Q_m(T)$ be the random variable describing the maximum storm which will occur in the next T years and let F_T be its distribution function. As a simple case assume that the arrival of storms is Poisson and that the exceedances are exponentially distributed. Then using similar methods as shown above it can be shown that:

$$Q_m(T) \leq q_o \qquad \text{with probability } \exp(-\lambda_o T) \tag{30a}$$

$$F_T(q) = \exp\{-\exp[q - q_o - \beta \ln(\lambda_o T)]/\beta\}\ ,\ q > q_o \tag{30b}$$

Once the parameters have been estimated the complete distribution of $Q_m(T)$ is known. Given $Q_m(T) > q_o$ the first two moments are approximately: (because of a small probability that no storms occur):

$$E(Q_m(T)) = q_o + \beta(\ln(\lambda_o T) + \gamma) \tag{31a}$$

$$V(Q_m(T)) = \beta^2\pi^2/6 = 1{,}644\,93\,\beta^2 \tag{31b}$$

where $\gamma = 0{,}57722$ (Euler's constant).

Now because the whole distribution of $Q_m(T)$ is known one can give (a) the probability that the biggest storm in the next T years will be greater than any given value, (b) for each given risk the corresponding maximum storm depth, and (c) the expected value and variance of the biggest storm in the next T years.

Consider a further extension of the partial duration series approach which could substitute for the common practice of comparing various competing models in terms of the asymptotic sampling variance of $\hat{q}_a(T)$. Instead we propose a fixed event magnitude and compute the variance of T, the random variable describing the intervals between storms. Given, say that storms arrive according to a Poisson process, the interarrival times between storms exceeding q_o will be exponentially distributed. (We suppose for the purpose of this argument that q_o is selected sufficiently large so that the effects of seasonality can be ignored.) This distribution of T is then given by:

$$F_T(t) = 1 - \exp(-t/\lambda_o) \tag{32}$$

where λ_o is the average number of storms per year. The mean and variance of T are λ_o and λ_o^2, respectively. The variance of $\hat{\lambda}_o$ is λ_o^2/n.

Consider the example of the Pretoria storm of 200 mm in one day in January 1978. Conventional analysis (Adamson, 1981) gives such a storm a recurrence interval of approximately 200 years. From data records, only one such storm occurred in 66 years, therefore from (32) $\hat{\lambda}_o = 66$ with estimated variance $66^2/1$ (!). The estimated distribution of T is exponential with parameter 66. Table 3 gives some of the percentage points for this distribution.

In this example, with the information available, one can with almost equal confidence estimate the return period of the 200 mm storm event to be either greater than 100 years or less than 15 years! The purpose of the example, extreme as it may be, is to illustrate the very limited amount of information on extreme risk conveyed by the conventional concept of the recurrence interval and the unjustified faith accorded to the concept for design purposes.

TABLE 3
THE ESTIMATED DISTRIBUTION OF T FOR A 200 mm ONE DAY STORM OVER PRETORIA FROM 66 YEARS OF HISTORICAL DATA

P (t ≤ T)	T
14%	10 year storm or less
20%	15 year storm or less
32%	25 year storm or less
45%	40 year storm or less
63%	66 year storm or less
78%	100 year storm or less
95%	200 year storm or less

Conclusions

The optimal assumed distribution for sequences of annual maxima or exceedances is probably an impossible objective to achieve and certainly a theorectically watertight case for any model could not be presented. If, for example, the partial duration series approach to extreme value analysis has any theoretical advantages over the more usual annual maximum method then it certainly would not be a straightforward task to prove it to be so. Attempts based on fairly rigid assumptions are relevant only insofar as such assumptions hold, which is rarely very far at all. Monte Carlo experiments are sample bound and have yet to be carried out on a scale such that any universally applicable conclusions can be drawn. There is, however, a tendency for annual maxima and the concept of recurrence interval to be accepted as the only yardstick for the assessment of design risk while models are applied with little if any investigation of their suitability or theoretical justification with respect to the problem at hand.

The present work has illustrated some results relating partial and annual series where the arrival process is distributed as Poisson and negative Binomial. The concept of recurrence interval has been defined in the annual and naive sense and a clear distinction made. An alternative to the usually adopted exponential distribution of exceedances is shown to be the censored log-Normal and estimation procedures are given. Empirical comparative studies between the two models of partial duration series and selected annual maximum models have shown the censored log-Normal model to have certain advantages and based on these it was chosen for a study of some 2 500 n-day rainfall records in South Africa and South West Africa/Namibia. (Adamson, 1981).

Finally the flexibility of the partial duration approach to risk analysis has been illustrated and some areas for further study outlined with examples.

Acknowledgment

We wish to thank one of the referees for his comments and suggestions and the Manager, Scientific Services, Department of Environment Affairs for permission to publish.

References

ABRAMOWITZ, M. and STEGUN, I.A. (1972) *Handbook of Mathematical Functions.* Dover Publications Inc. 9th Edition. New York.

ADAMSON, P.T. (1981) Southern African Storm Rainfall. Technical Report No. 102. Branch of Scientific Services. Department of Environment Affairs. Pretoria.

BOUGHTON, W.C. (1980) A Frequency Distribution for Annual Floods. *Water Resources Research.* **16**(2) 347-354.

BOX, G. and JENKINS, G. (1976) *Time Series Analysis: Forecasting and Control.* Holden Day. San Francisco.

CALENDA, G., PETACCIA, A. and TOGNA, A. (1977) Theoretical Probability Distribution of Critical Hydrological Events by the Partial Duration Series Method. *Journal of Hydrology.* **33** 233-245.

CANFIELD, R.V., OLSEN, D.R. and HAWKINS, R.H. (1980) Use of Extreme Values in Estimating Flood Peaks from Mixed Populations. Hydraulic and Hydrology Series No. UWRL/H-80/01. Utah Water Research Laboratory. Utah State University. Logan. Utah.

CHANDER, S., SPOLIA, S.K. and KUMAR, A. (1978) Flood Frequency Analysis by Power Transformation. Proceedings, American Society of Civil Engineers. *Journal of the Hydraulics Division.* **104**(1) 1495-1504.

COHEN, A.C. Jnr. (1950) Estimating the mean and variance of Normal Populations from Singly Truncated and Doubly Truncated samples. *Annals of Mathematical Statistics* **21** 557-567.

CUNNANE, C. (1973) A Particular Comparison of Annual Maxima and Partial Duration Series Methods of Flood Frequency Analysis. *Journal of Hydrology* **18** 257-271.

CUNNANE, C. (1979) Note on the Poisson Assumption in Partial Duration Series Models. *Water Resources Research* **15** 489-494.

GABRIEL, K.R. and NEUMANN, J. (1957) On a Distribution of Weather Cycles by Length. *Quarterly Journal of the Royal Meteorological Society.* **83** 375-380.

GALAMBOS, J. (1978) *The Asymtotic Theory of Extreme Order Statistics.* Wiley Series in Probability and Mathematical Statistics. John Wiley. New York.

GUMBEL, E.J. (1958) *Statistics of Extremes.* Columbia University Press. New York.

GUPTA, A.K. (1952) Estimation of the Mean and standard Deviation of a Normal Population from a Censored Sample. *Biometrika.* **39** 260-273.

HOUGHTON, J.C. (1978) Birth of a Parent: The Wakeby Distribution for Modelling Flood Flows. *Water Resources Research.* **14**(6) 1105-1109.

JOHNSON, N.L. and KOTZ, S. (1969) Discrete Distributions. Wiley Series in Probability and Mathematical Statistics. John Wiley. New York.

JOHNSON, N.L. and KOTZ, S. (1970) *Continuous Univariate Distributions – 1.* Wiley Series in Probability and Mathematical Statistics. John Wiley. New York.

KITE, G.W. (1977) Frequency and Risk Analysis in Hydrology. Water Resources Publications. Fort Collins. Colorado.

NERC (Natural Environment Research Council) (1975) Flood Studies Report. Volume 1. Whitefriars. London. 550 pp.

PARZEN, E. (1962) *Stochastic Processes.* Holden Day Inc. San Francisco.

PITMAN, W.V. (1980) A Depth-Duration-Frequency Diagram for Point Rainfall in SWA/Namibia. *Water SA.* **6**(4) 157-162. Discusson. *Water SA* (1981) 7(4) 265-268.

PRASAD, R. (1970) Frequency Analysis of Hydrologic Information. Paper presented at the August Conference of the American Society of Civil Engineers. Hydraulics Division. Minneapolis Minnessotta.

SING, K.P. and SINCLAIR, R.A. (1972) Two Distribution Method for Flood Frequency Analysis. Proceedings. American Society of Civil Engineers. *Journal of the Hydraulics Division* **98**(1) 29-44.

TODOROVIC, P. and ZELENHASIC, E. (1970) A Stochastic Model for Flood Analysis. *Water Resources Research* **6**(6) 1641-1648.

YEVJEVICH, V. and TAESOMBUT, V. (1978) Information on Flood Peaks in Daily Flow Series. Proceedings. International Symposium on Risk and Reliability in Water Resources Vol. 1. University of Waterloo. Ontario Canada.

Index of Notation

d	Fisher disperson test statistic
exp	Exponential function
E	Expected value
e_o	Parameter of negative Binomial distribution
F_a	Distribution function of the largest storm within a year
F_o	Distribution function of storms exceeding q_o
F_T	Distribution function of the maximum storm that will occur within the next T years
N	Number of years
n	Total number of rain days
n_1	Number of storms exceeding q_o
n_2	$n - n_1$
$p_o(m)$	Probability function of the number of storms exceeding q_o in one year
p	Probability
$q_a(T)$	T year event: annual maximum sense
$q_o(T)$	T year event: naive sense
q_o	Truncation level
q_i, $i = 1, 2, \ldots, m_1$	Exceedance Series
Q_a	Random variable describing the largest storm event within a year
Q_o	Random variable describing a storm which exceeds q_o

$Q_m(T)$	Random variable describing the largest storm that will occur in the next T years
T	Recurrence interval (annual maximum sense)
T_o	Recurrence interval (naive sense)
V	Variance
W, w, W*, w*	Weighting functions for the censored log-Normal distribution
y(T)	Reduced Gumbel variate
α_o	Parameter of negative Binomial distribution
β	Scale parameter of exponential distribution
γ	Euler's constant = 0,57722
ξ_o	Standardized point of truncation
λ_o	Mean number of storms per year
μ	Mean
σ	Standard deviation
$\phi(t)$	Density function of a standard normal variable
$\Phi(x)$	Distribution function of a standard normal variable
Z(p)	$= \Phi^{-1}(p)$

Appendix

Sample solution for a censored log-Normal model of partial duration series

To illustrate estimation procedures for a singly censored log-Normal model of a partial duration series, we consider the example of point storm rainfalls at gauge 8/751, Swellendam (Lat. 34° 01′, Long. 20° 26′) for which there are 47 years of daily data available. The solution provides estimates of the 2, 5, 10, 20, 50, 100 and 200 year one-day rainfall depths.

Step 1

a) Find maximum one day rainfall depth for each year. (N = 47).
b) Find the minimum, q_o, of these maxima (q_o = 33).
c) Compute total number of days, n, for which rainfall depth exceeds trace (0,25 mm). (n = 4 035).
d) Record the rainfall depths $q_1, q_2, \ldots, q_{n1}$ which exceed q_o, where n_1 is the observed number of such events. Set $n_2 = n - n_1$. (n_1 = 205; n_2 = 3830).
e) Define a "storm day" as one upon which rainfall depth exceeds q_o and compute the average number of storm days per year. $\lambda_o = (n_1/N)$. (λ_o = 4,361 702).
f) Compute:

$$v_1 = \sum_{n=1}^{n_1} (\ln q_i - \ln q_o)/n_1 \qquad (v_1 = 0{,}383\ 505)$$

$$v_2 = \sum_{i=1}^{n_1} (\ln q_i - \ln q_o)^2/n_1 \qquad (v_2 = 0{,}261\ 102)$$

$$v = v_2/v_1^2 \qquad (v = 1{,}775\ 285)$$

Step 2

a) Set $x_1 = Z([n_2 + 0{,}1]/n)$ and $x_2 = Z(n_2/n)$

where the function Z is the inverse normal distribution function.

Define $f(x) = [1 - x\,Y(x)]/Y(x)^2 - v$

where $Y(x) = [n_2\,\phi(x)]/[n_1\,\Phi(x)]$

The values of Z, ϕ and Φ can be found in standard tables or from the algorithms given at the end of the Appendix. For K = 3, 4 . . . set

$$x_{K+1} = x_K - f(x_K)/\{[f(x_K) - f(x_{K-1})]/[x_K - x_{K-1}]\}$$

until $|x_{K+1} - x_K| < 10^{-6}$

The estimated point of truncation, $\hat{\xi}_o$, is given by $\hat{\xi}_o = x_{K+1}$ (ξ_o = 1,637 27).

b) Compute the estimates

$$\hat{\sigma} = v_1/\xi_o \qquad (\hat{\sigma} = 0{,}913\ 67)$$

$$\hat{\mu} = \ln q_o - \hat{\sigma}\,\hat{\xi}_o \qquad (\hat{\mu} = 2{,}000\ 58)$$

Step 3

For each recurrence interval, T (years), required set . . .

$$p = (1 - \Phi(\xi_o)) \ln (1 - 1/T)/\lambda_o + 1 \quad ,$$

$$\hat{q}_a(T) = \exp(\mu + \sigma\,Z(p))$$

The computed estimates for this example are given in Table A1.

TABLE A1

T (years)	p	Z(p)	$\hat{q}_a(T)$ mm
2	0,991 930	2,406 102	67
5	0,997 402	2,794 938	95
10	0,998 773	3,029 294	118
20	0,999 403	3,240 421	143
50	0,999 765	3,497 184	181
100	0,999 883	3,679 215	213
200	0,999 942	3,852 947	250

The following approximations for ϕ, Φ and Z are from Abramovitz and Stegun (1972) pp. 932-933.

$$\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$$

$$\Phi(x) = 1 - \phi(x)(b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5)$$

where $t = 1/(1 + b_o x)$, b_o = 2,231 641 9,
b_1 = 0,319 381 530, b_2 = 0,356 563 782,
b_3 = 1,781 477 937, b_4 = 1,821 255 978,
b_5 = 1,330 274 429

$$Z(p) = t - (c_o + c_1 t + c_2 t^2)/(1 + d_1 t + d_2 t^2 + d_3 t^3)$$

where $t = \ln(1/p)^{1/2}$, c_o = 2,515 517,
c_1 = 0,802 853, c_2 = 0,010 328,
d_1 = 1,432 788, d_2 = 0,189 269,
d_3 = 0,001 308