Atmósfera (1992), **5**, *pp. 47-56*

The extreme value distribution of rainfall data at Belgrade, Yugoslavia

UNKAŠEVIĆ MIROSLAVA

Hydrometeorological Institute of Republic of Serbia, Gandijeva 117, 11070 N. Belgrade, Yugoslavia

(Manuscript received March 18, 1991; accepted in final form, Sept. 10, 1991)

RESUMEN

Para las aplicaciones prácticas, la distribución original de las intensidades de lluvia y la de sus máximas anuales son ambas interesantes. La relación entre estas dos distribuciones no pude obtenerse a partir de la teoría clásica de los valores extremos debido a la viariación estacional y a la autocorrelación en los datos. Los resultados matemáticos para la distribución de los máximos en las secuencias m-dependientes se dan para ilustrar el efecto de dependencia local sobre la distribución de valores extremos. El número promedio de excedentes en un racimo es un parámetro importante en la relación entre la distribución original y la de valores extremos. Para lluvias caídas en los primeros 5 minutos, de datos provenientes de Belgrado, los cuantiles de los máximos anuales resultan sobre-estimados por casi 10 mmh^{-1} si el efecto de autocorrelación se ignora. Este sesgo puede fácilmente eliminarse, tomando en cuenta el arracimaje local de grandes intensidades pluviales en una temporada lluviosa.

ABSTRACT

For practical applications both the parent distribution of rainfall intensities and the distribution of their annual maxima are of interest. The relation between these two distributions cannot be obtained from classical extreme value theory because of seasonal variation and serial correlation in the data. Mathematical results for the distribution of maxima in m-dependent sequences are given to illustrate the effect of local dependence on the extreme value distribution. The average number of exceedances in a cluster is an important parameter in the relation between the parent and extreme value distribution. For 5-min rainfall data from Belgrade quantities of the annual maxima are overestimated by about 10 mm h^{-1} if the effect of serial correlation is ignored. This bias can easily be removed by taking local clustering of large rainfall intensities in a rainy spell into account.

1. Introduction

Statistical information about the occurrence of heavy rainfall can be given in different ways. For hydrological applications the data are often presented in the form of extreme value statistics (usually annual maxima). However, this is not the most suitable form of presentation for every user.

There is a relation between the parent and extreme value distribution. This relation is well-known for independent random variables with the same distribution (Gumbel, 1958). Unfortunately rainfall data over short intervals are neither independent nor identically distributed.

In most parts of the world there is an obvious seasonal variation in the occurrence of high rainfall rates. For instance, in Belgrade (Yugoslavia), 5-min intensities larger than 25 mm h^{-1} only occur during the summer period May-September. Often there is not only an annual cycle but also a diurnal cycle.

Figure 1 gives three hypothetical situations in which there is at least one 5-min interval with an average intensity greater than 10 mm h^{-1}. Examples were found in De Bilt (Buishand, 1984). In case A there is a single 5-min interval with a rainfall rate greater than 10 mm h^{-1}, whereas in the other cases there is some clustering of high rainfall rates due to dependence between the rainfall amounts of neighboring 5-min intervals. In case B there is a run of two adjacent intervals in which the rainfall rate exceeds the 10 mm h^{-1} threshold. Case C is more complicated since in this case there are two runs in which the threshold is exceeded.

A good insight into the relation between the parent and extreme value distributions of such data requires some knowledge of modern extreme value theory. In this paper a number of mathematical results for the distributions of maxima are discussed. These results are applied to 5-min rainfall data of Belgrade.

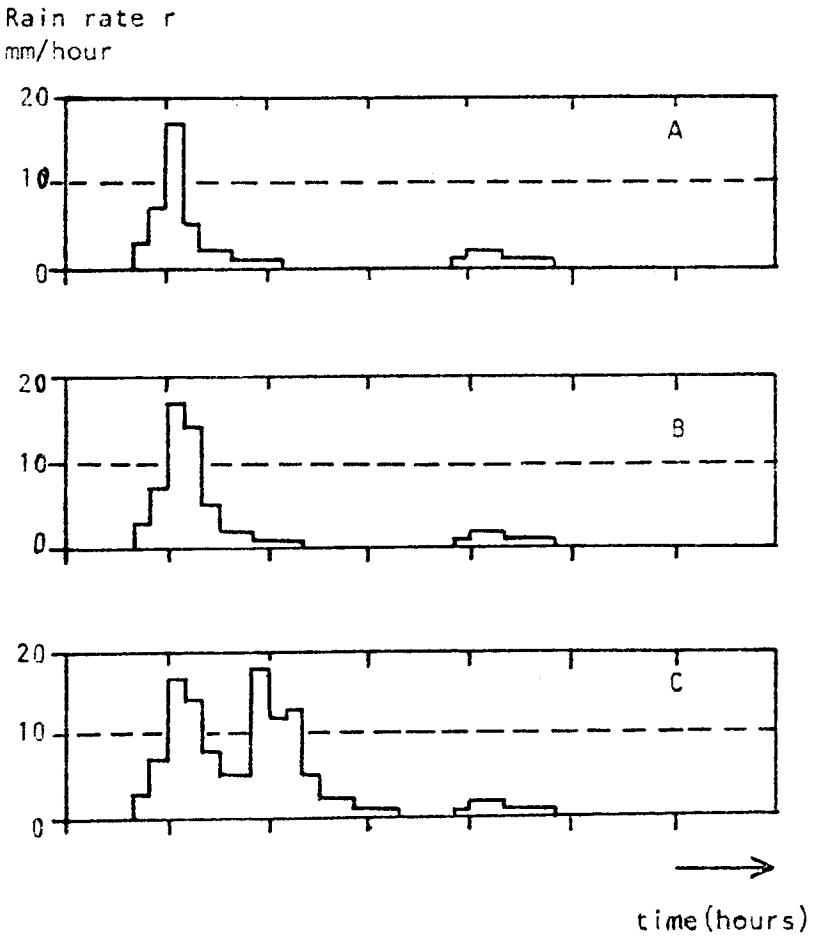

Fig. 1. Examples of rainfall events with 5-min rain rates exceeding a threshold of 10 mm h^{-1}

2. The distribution of maxima of independent random variables

Let $R_1, R_2, \ldots\ldots$ be a sequence of random variables and

$Z_n = \max(R_1, R_2, \ldots\ldots, R_n)$. In our application to the rainfall data, the R_i will refer to average intensities over consecutive 5-min intervals in a particular year.

In the first instance we assume that the R_i are independent with a common distribution function

$$F(r) = P_r(R_i < r), \quad i = 1, \ldots\ldots, n. \tag{1}$$

Then for the distribution of the maximum Z_n we obtain

$$P_r(Z_n < r) = P_r(R_1 < r, \ldots\ldots, R_n < r) = F^n(r). \tag{2}$$

For large n the distribution of Z_n is determined by the shape of the right tail of the parent distribution $F(r)$. Using the approximation

$$\ell_n F(r) \sim -[1 - F(r)],$$

for large r we obtain

$$P_r(Z_n < r) \sim e^{-n[1-F(r)]} \tag{3}$$

for the distribution of Z_n. The quantity $n[1 - F(r)]$ gives the expected number of R_i that exceed the threshold r.

In general, $F_i(r) = P_r(R_i < r)$ slowly varies with i and for large r the process of exceedances can be regarded as a nonhomogeneous Poisson process. Therefore we can write

$$a(r) = \sum_{i=1}^{n} [1 - F_i(r)] = n[1 - F(r)] \tag{4}$$

The approximation (3) becomes now as

$$P_r(Z_n < r) \sim e^{-a(r)} \tag{5}$$

The number of exceedances $S_n(r)$ of the level r has a binomial distribution that for large r can be approximated by a Poisson distribution with parameter $a(r)$. The approximation (5) then follows immediately from the fact that $P_r(Z_n < r) = P_r[S_n(r) = 0]$.

The accuracy of this Poisson approximation for Belgrade is shown in Figure 2, where f denotes frequency and k the number of heavy one-day rains (higher than 25 mm h^{-1} (Janc, 1987)).

In many applications the tail of the parent distribution can be approximated by an exponential distribution. This implies that for large r

$$a(r) \sim e^{-\tau_a(r-u_a)}. \tag{6}$$

Substitution of (6) in (4) gives

$$P_r(Z_n < r) \sim \exp[-e^{-\tau_a(r-u_a)}], \tag{7}$$

which is a Gumbel distribution with location parameter u_a and scale parameter τ_a. Note from (6) that u_a is the value of r for which $a(r) = 1$, i.e., u_a has an expectation of being exceeded just once in a sequence of length n.

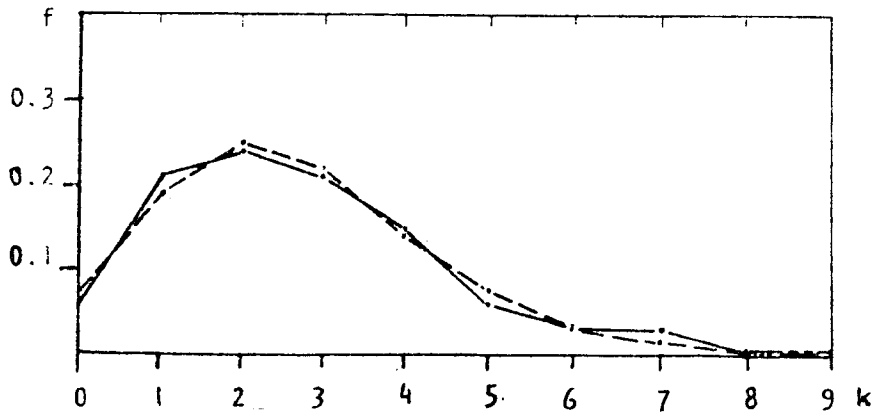

Fig. 2. Empiric (–) and theoretic (– – –) Poisson frequencies (K denotes number of heavy one-day rains).

3. Extreme value theory for sequences of dependent random variables

Correlation between successive values in a time series does not always put limitations on the applicability of the approximations given in the previous section (Leadbetter, 1983). For large n, Eq. (5) can still be used if the following conditions are satisfied:

1) The sequence $R_1, R_2, \ldots$ is a mixing sequence. In essence this condition requires that the various terms in the sequence may only be weakly dependent when their separation in time is large. For instance, in a mixing sequence,

$$P_r(R_1 < r,\ R_2 < r,\ R_k < r)$$

tends to

$$P_r(R_1 < r,\ R_2 < r)P_r(R_k < r) \text{ as } k \to \infty.$$

2) The random variables R_i and R_{i+k} are pairwise asymptotically independent in the right tail for every $k \neq 0$, i.e.,

$$P_r(R_{i+k} \geq r / R_i \geq r)$$

tends to zero as $r \to \infty$. In particular, if we consider exceedances of a high-level r, there will be no local clustering of such events (case A in Fig. 1).

Since sequences of rainfall amounts over certain time intervals are characterized by relatively

short memories, we may safely assume that mixing condition 1 is satisfied. However, departures from condition 2 may have a serious impact on the distribution of maxima.

To examine the distribution of run lengths in more detail, let us consider the event that a run with length greater than 2 starts at $t = i + 1$:

$$E_i = \{R_i < r,\ R_{i+1} \geq r,\ R_{i+2} \geq r,\ R_{i+3} \geq r\}. \tag{8}$$

A sequence of 1-dependent random variables is characterized by the property that the events $\{R_1 < r_1, \ldots, R_j < r\}$ and $\{R_{j+k} < r_{j+k}, \ldots, R_n < r_n\}$ are dependent if $k = 1$, but independent if $k > 1$.

For 1-dependent processes we obtain

$$P_r(E_i) = P_r(R_i < r,\ R_{i+1} \geq r,\ R_{i+3} \geq r) =$$

$$= P_r(R_i < r,\ R_{i+1} \geq r)P_r(R_{i+3} \geq r), \tag{9}$$

and thus

$$\frac{P_r(E_i)}{P_r(R_i < r,\ R_{i+1} \geq r)} \leq P_r(R_{i+3} \geq r). \tag{10}$$

The left-hand side of (10) gives the fraction of runs with a length greater than 2. Since, the right-hand side of (10) vanishes as $r \to \infty$, it follows that for large values of the threshold r there will be no run or hardly any runs with a length greater than 2. So for 1-dependent processes, exceedances of a high-level r occur either as single peaks (case A in Fig. 1) or paired in a run (case B in Fig. 1). To explain more complex situation like case C in Figure 1, we have to consider higher-order dependence models.

3.1 Sequence with higher-order dependence

In an m-dependent sequence, events are independent if they are separated by more than m time units (e.g., an mth-order moving average process). For the distribution of the maximum Z_n of such a sequence it was shown by Newell (1964) that for large n

$$Pr_i(Z_n < r) \sim e^{-nP_r(D_i)}, \tag{11}$$

where D_i denotes the event

$$D_i = \{R_i \geq r,\ R_{i+1} < r, \ldots., R_{i+m} < r\}. \tag{12}$$

If $m = 1$ then (11) is reduced to

$$Pr(Z_n < r) \sim e^{-b(r)}, \tag{13}$$

where $b(r)$ denotes the expected number of runs above the threshold r. The quantity $nPr(D_i)$ gives the expected number of events D_1, D_2, ... that occur in a sequence of length n. If r is

large then $Pr(D_i)$ will be small and the occurrence times of events $D_1, D_2, \ldots$ approximately form a Poisson process. The right-hand side of (11) gives the probability that none of the events $D_1, D_2, \ldots D_n$ occur in this limiting Poisson process.

When $Pr(D_i)$ varies with season, then in general the limiting process will be a nonhomogeneous Poisson process; this leads to the following asymptotic distribution for the maximum:

$$Pr(Z_n < r) \sim \exp[-\sum_{i=1}^{n} Pr(D_i)]. \tag{14}$$

Let us now return to case C in Figure 1. Assume that the rainfall is a 12-dependent process (i.e., has a memory of 12×5 min $= 1$ h). If we choose a threshold of 10 mm h^{-1} then we have two runs, but the event D_i only occur at the end of the second run. In general, for the distribution of the maximum the number of clusters rather than the number of runs or individual exceedances has to be taken into account (Rootzen, 1978).

For the 1-dependent process, exceedances of a high-level r may occur in runs, but there is no local clustering of these runs. If we have m-dependent process with $m > 1$ then it is possible that runs also accur in bunches. In fact, there will be no clustering of runs if for large r

$$Pr(D_i) \sim Pr(R_i \geq r, \ R_{i+1} < r). \tag{15}$$

Then (13) can be used to obtain the asymptotic distribution of the maximum. In contrast with the 1-dependent process, runs with a length greater than 2 need not be rare events. But if for large r we have

$$Pr(D_i) < Pr(R_i \geq r, \ R_{i+1} < r), \tag{16}$$

then runs above the threshold r occur in clusters and (14) should be used to obtain the asymptotic distribution of Z_n.

An m-dependent sequence is a special case of a mixing sequence. The poisson limit for the point process of cluster positions of high-level exceedances remains valid for these more general sequences (Leadbetter, 1983). The nature of clustering of rare events can be derived from the probabilistic structure of the underlying stochastic process.

In case C (Fig. 1) there is strong local clustering of large values. Therefore, instead of (14) we can try the approximation

$$Pr(Z_n < r) \sim e^{-c(r)}, \tag{17}$$

where $c(r)$ denotes the expected number of rainy spells in a sequence of length n with at least one 5-min rain rate greater than or equal to the threshold r. Here a rainy spell is defined as an uninterrupted sequence of wet 5-min intervals bounded on each side by a dry 5-min interval. This is in fact a run for which the threshold r is equal to the smallest measurable rainfall amount (0.14 mm h^{-1} at Belgrade) in sequence.

Equation (17) is based on the assumption that for large r the number of clusters in a sequence of length n has a Poisson distribution with parameter $c(r)$. To be more specific about the distribution of Z_n we need to know the form of $c(r)$ for large r. If $c(r)$ is exponential, i.e.,

$$c(r) \sim e^{-\tau_c(r-u_c)}, \tag{18}$$

then we obtain

$$Pr(Z_n < r) \sim \exp[-e^{-\tau_c(r-u_c)}]. \tag{19}$$

Hence Z_n has a Gumbel distribution with location parameter u_c and scale parameter τ_c.

Note that is not necessary to describe the seasonal variation in the occurrence times of clusters of rare events to derive the distribution of the annual maximum Z_n. For this purpose we only need to know how the expected annual number of clusters $c(r)$ varies with r.

4. Application to 5-min data for Belgrade

Thirty years of data (1951-1980) were digitalized from the self recording rain gauge of Belgrade. From the 5-min data, estimates of $a(r)$ and $c(r)$ were obtained by counting the numbers of events and rainy spells with a rainfall intensity of at least r mm h^{-1} using Eqs. (6) and (18). The estimates $\hat{a}(r)$ and $\hat{c}(r)$ were plotted on a logarithmic scale (Fig. 3). From the figure it can be seen that even for very high rainfall intensities, $\hat{c}(r)$ still differs from $\hat{a}(r)$. The ratio $\hat{a}(r)/\hat{c}(r)$ is ~ 1.5 which means that there are on average 1.5 exceedances in a rainy spell. A consequence of this local clustering of high rain rates is that Eqs. (5) and (17) do not lead to the same result.

Straight lines are fitted in Figure 3 for $r \geq 25$ mm h^{-1}. The fit is reasonable and therefore the approximations (6) and (18) can be applied. For $r < 25$ mm h^{-1} the points deviate from the fitted line. However, this has no serious consequences for the distribution of extreme values because the annual maximum is nearly always greater than 25 mm h^{-1}.

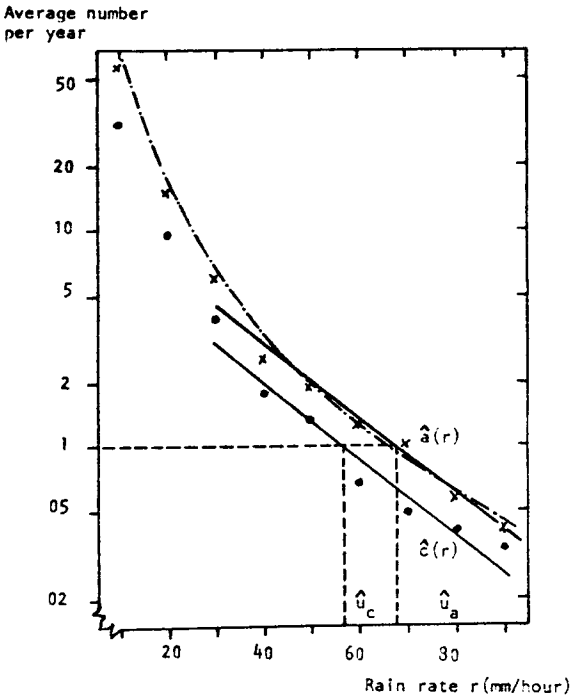

Fig. 3. Estimates of $a(r)$ (crosses) and $c(r)$ (dots) for 5-min rain rates at Belgrade. The quantities $\hat{a}(r)$ and $\hat{c}(r)$ refer to a sequence with a length of one year. The straight lines represent the exponential approximations (6) and (18), respectively; the curved line is based on the lognormal distribution (20).

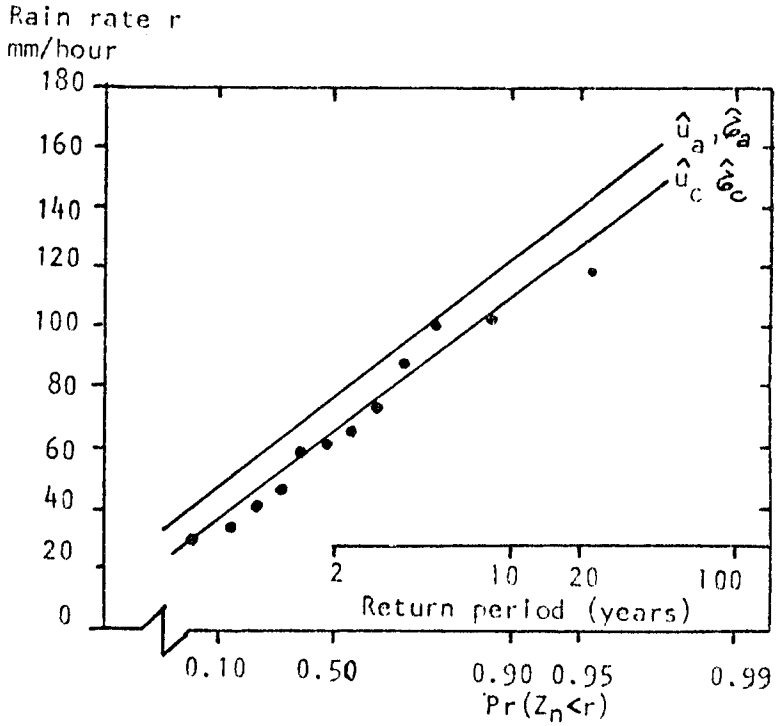

Fig. 4. Gumbel probability plot of annual 5-min rain rate maxima (dots) for Belgrade (30-year period). The Gumbel line with parameters $\hat{u}_a$ and $\hat{\tau}_a$ is based on the assumption that the data are independent, whereas the Gumbel line with parameters $\hat{u}_c$ and $\hat{\tau}_c$ is obtained with a model in which dependence is taken into account.

From the straight lines in Figure 3 we can get an estimate of the parameters u_a, τ_a and u_c, τ_c. The Gumbel lines with these parameters are shown in Figure 4. The upper line with parameters $\hat{u}_a$ and $\hat{\tau}_a$ should be the correct distribution of the annual maximum if the effect of dependence could be ignored. This is not the case, however. Although the differences between the two lines are not very large ($\sim$ 10 mm h^{-1}), the lower line, which is based on a reasonable dependence model, gives a better fit.

To correct for dependence we must have some idea about the average number of exceedances in a cluster, i.e., the ratio $a(r)/c(r)$ as an important parameter in the relation between the parent and extreme value distribution. The nature of clustering of rare events can be derived from the probabilistic structure of the under-lying stochastic process. As an alternative, information about clustering of large values can be obtained directly from data. When this quantity is known it is not difficult to derive the extreme value distribution from the parent distribution or the right tail of the parent distribution from the extreme value distribution.

It should be stressed, however, that the chosen extreme value distribution only gives information about the upper tail of the parent distribution. Although we may obtain the exponential tail from the extreme value distribution it is dangerous to extrapolate this result to lower rain rates.

From Figure 3 it is seen that the exponential distribution greatly underestimates the average number of exceedances $a(r)$ for rain rates less than 25 mm h^{-1}. In the literature the parent distribution of rainfall amounts over short intervals has often been approximated by a lognor-

mal distribution (Marshall, 1983). The curved line in Figure 3 is based on the two-parameter lognormal distribution

$$a(r) = n\left\{1 - \phi\left(\frac{\ell_n r - \mu}{\tau}\right)\right\}, \tag{20}$$

where $\phi(\ )$ stands for the standard normal distribution function. The parameters μ and τ can be obtained from u_a and τ_a using the relations (Singpurwalla, 1972)

$$\tau = 2\ell_n n)^{1/2}/(u_a \tau_a) \tag{21}$$

and

$$\mu = \ell_n u_a - \tau d_n, \tag{22}$$

where

$$d_n = (2\ell_n n)^{1/2} - \frac{\ell_n(\ell_n n) + \ell_n 4\pi}{2(2\ell_n n)^{1/2}}. \tag{23}$$

This two-parameter lognormal distribution gives a reasonable fit for values of r between 10 and 90 mmh^{-1} (Fig. 3). Extrapolations outside this range require further study of the shape of $a(r)$.

5. Conclusions

The classical theory of extreme values is based on the assumption that data are independent and identically distributed. Even when the data are correlated the classical theory may give the correct asymptotic distribution of maxima. However, the theory should be extended if rare events occur in clusters. Then the distribution of maxima is not directly related to the average number of exceedances of some high-level r as in the independent case, but to the number of clusters with large values. This result holds not only for stationary sequences but also for sequences with a seasonal component.

It is, of course, not always obvious how a cluster of rare events should be defined, but for 5-min rates at Belgrade it is sensible to look at the number of rainy spells in which the level r is exceeded. This results in a better approximation to the distribution of annual maxima than a model in which the distribution of extremes is related to the average number of exceedances of some threshold r.

REFERENCES

Buishand, T., 1984. The effect seasonal variation and serial correlation on rainfall data. *J. Climate Appl. Meteor.*, **24**, 154-160.

Gumbel, J., 1958. Statistic of Extremes, Columbia University Press, 375 pp.

Janc, N., 1987. Statistička obrada jakih kiša na području Beograda. *Zborni meteor. i hidrol. radova*, **14**, 34-37.

Leadbetter, R., 1983. Extremes and local dependence in stationary sequences. *Z. Wahr. verw. Geb.*, **65**, 291-306.

Marshall, R., 1983. A spatial-temporal model of storm rainfall. *J. Hydrol.*, **62**, 53-62.

Newell, G., 1964. Asymptotic extremes for m-dependent random variables. *Ann. Math. Statist.*, **35**, 1322-1325.

Rootzen, H., 1978. Extremes of moving averages of stable processes. *Ann. Probab.*, **6**, 847-869.

Singpurwalla, N. D., 1972. Extreme values from a lognormal law with applications to air pollution problems. *Technometrics*, **14**, 703-71.