NISS

On Exceedance Based Environmental Criteria

Part I: Basic Theory

M.R. Leadbetter

Technical Report Number 9
December, 1993

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

On Exceedance Based Environmental Criteria I: Basic Theory *

M.R. Leadbetter
Department of Statistics
University of North Carolina
and
National Institute for Statistical Sciences

Abstract

The current "Ex-Ex" criterion for ozone and two possible secondary criteria ("Area over threshold" and "SUM06") are discussed within a general framework of compliance criteria obtained as functions of excess values over threshold levels. Their basic statistical properties are obtained from the theory of [6] which obtains Poisson-type and normal approximations for such "exceedance statistics" above high and moderate threshold levels.

The roles of level height, and the clustering of exceedances are discussed along with the distributional results obtained, in relatively non technical terms. The Poisson and normal type results given provide a basis for calculation of probabilities of correct compliance classification. Numerical results will be presented in Part II based on 1980-90 ozone data from selected U.S. cities.

*Research supported by the EPA Cooperative Agreement CR 819638 01 0 with National Institute for Statistical Sciences, and Office of Naval Research Grant N00014-93-1-0043.

1 Introduction

Environmental compliance is typically determined from the size of some function of observed or measured values – here referred to as a compliance statistics (CS)- large values typically indicating lack of compliance. The CS may simply be an average of observed values (as for coal sulfur criteria), when no exceedance level for the observed variables enters the calculation of the CS itself. On the other hand the types of CS considered here are directly *exceedance based* in that they are obtained from the excess values of concentrations above a specified (high) threshold level, out of a total of n measured concentrations. Specifically the three cases considered are

(i) Expected exceedances (Ex-Ex), (the current ozone criterion) for which the CS is N_n, the number of exceedances of the level $u_n = .12$ ppm in a three year period.

(ii) Area over threshold (AOT), the CS being the sum A_n of values above a threshold in a specified period.

(iii) SUM06 proposed secondary criterion, where the CS is the sum S_n of total concentrations (rather than just excesses) during periods of threshold exceedances. Since this comprises both excess values above the level u_n and the height u_n itself during exceedance periods (see also Fig. 2), (iii) may be obtained simply from (i) and (ii), viz.

$$S_n = A_n + N_n u_n.$$

These three criteria may be set in a statistical framework. For the existing current ozone criterion (Case (i)) compliance is defined to mean an *expected* exceedance rate of no more than one per year (i.e. 3 per 3 years) of the .12 ppm standard. As noted above, this is tested by the *actual* number of exceedances N_n in 3 years, non compliance being declared if $N_n > 3$. This is thus a classical test for the mean of an observed r.v. Properties of the procedure (e.g. misclassification probabilities) are discussed in [1] along with possible modifications to the procedure.

The AOT and SUM06 criteria may be similarly regarded as tests for the expected values of the respective areas, based on the observed values in a given period.

The required statistical properties may be simply obtained from general theory of [5] by identifying each CS as a special case of exceedance measures considered there. More specifically, denote the n measured values by $X_1, X_2, \ldots, X_n$ and their excess values above the threshold u_n by $(X_i - u_n)_+$ ($= X_i - u_n$ if $X_i \geq u_n$ and zero if $X_i < u_n$) as indicated in Figure 1:

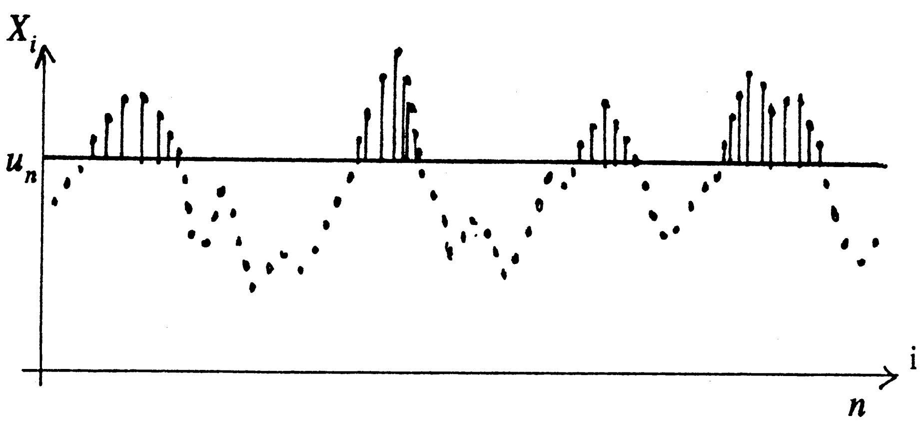

Figure 1: Excess Values $(X_i - u_n)_+$

Each of the above criteria can be expressed simply by the general mathematical form for the CS (considered in [5]. See also [6], [3])

$$Z_n = \sum_{i=1}^{n} \psi_n((X_i - u_n)_+) \tag{1.1}$$

for an appropriately chosen function ψ_n. Specifically it is easily checked that for each case $\psi_n(x) = 0$ for $x < 0$ and

(i) Ex-Ex ($Z_n = N_n$), $\psi_n(x) = 1$

(ii) AOT ($Z_n = A_n$), $\psi_n(x) = x$

(iii) SUM06 ($Z_n = S_n$), $\psi_n(x) = x + u_n$.

The three cases are illustrated in Figure 2 below:

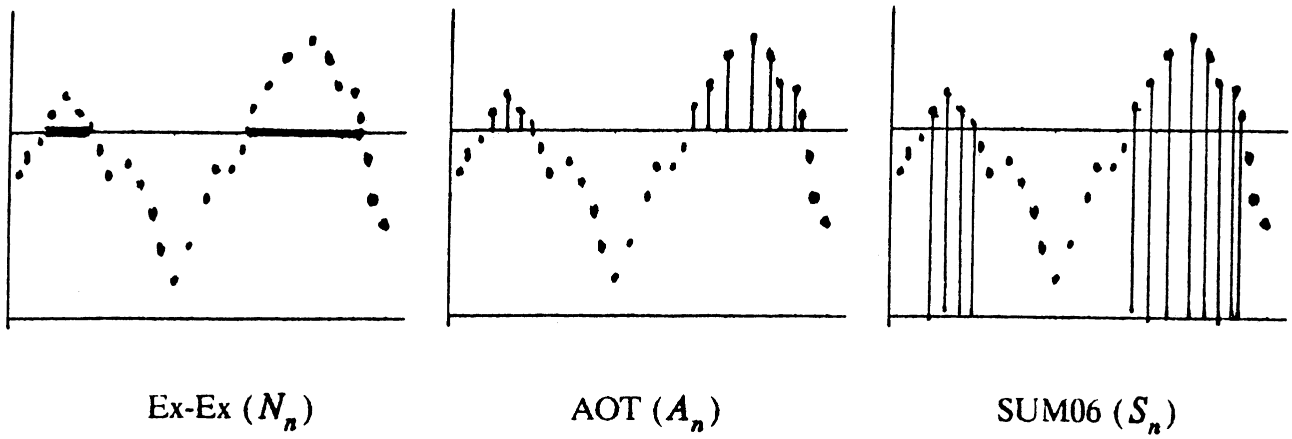

Figure 2: Contributions to values of CS $Z_n = N_n, A_n, S_n$.

The statistical properties of each CS (e.g. misclassification probabilities) may thus be obtained as special cases of asymptotic distributions of Z_n of the general form (1.1) given in [5] for high and moderate levels u_n. This theory makes only very general assumptions about the statistical nature of the environmental variables X_i and shows that two types of model for the CS are well founded:

(i) (Compound) Poisson (CP) models for "very high" thresholds

(ii) Normal models for "moderate" thresholds.

These will be described in Sections 5 and 6. As a rule of thumb one expects the CP models (i) when the exceedance events are rare (e.g. for very high levels relative to the bulk of observed values X_i). This is typically the case for compliant (or "moderately" non-compliant) situations with the Ex-Ex criterion. For the lower (moderate) levels used in the potential AOT and SUM06 criteria or for grossly non-compliant Ex-Ex cases, the normal models are appropriate.

Figure 3 below illustrates the two situations with a plot of a typical summer ozone record for Los Angeles. Levels of .27 ppm and higher exhibit the rare occurrence of exceedances for CP modeling, and lower levels (e.g. .2 ppm shown) lead to normal models. This will be discussed more explicitly in Sections 5 and 6, following indications in Sections 2 of the more precise meaning of "high" and "moderate" levels, the notion of exceedance clusters in Section 3, and general results of Section 4.

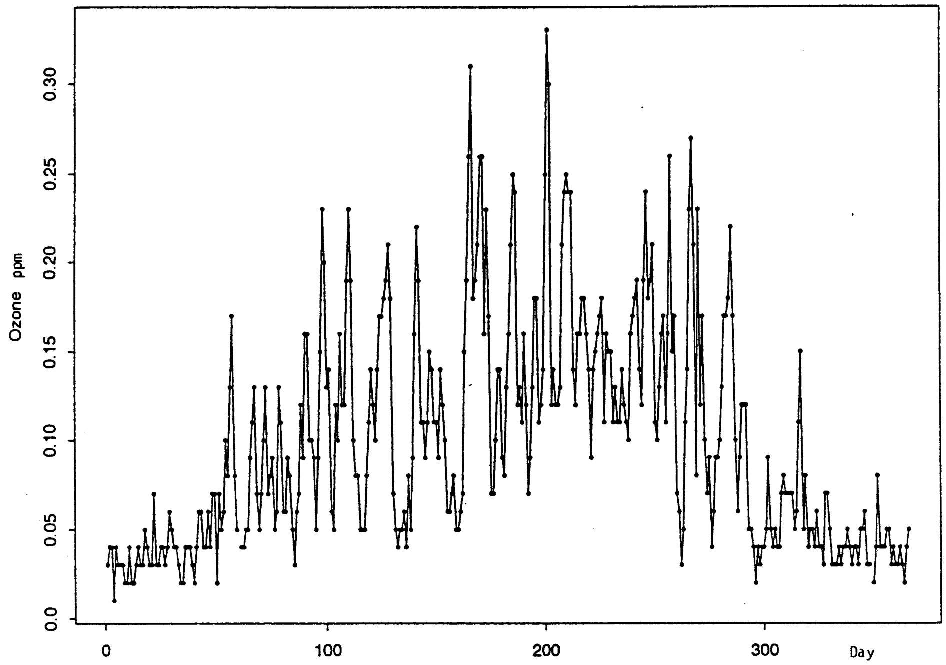

Figure 3: Ozone levels in Los Angeles, summer 1989

2 High and moderate levels

As noted, the CP and normal modeling cases are substantially distinguished by the heights of the threshold u_n. The precise distinction arises from different rates of increase of threshold u_n with n, the number of observed values, in underlying limit theorems. This is indicated very briefly here – full details may be found in [5].

Specifically if F denotes the distribution function (d.f.) of the environmental r.v.'s X_i, $F(x) = P\{X_i \leq x\}$, the levels u_n are regarded as *high* if the individual "exceedance probability" $(1 - F(u_n))$ is small and the expected number of exceedances c_n $= n(1 - F(u_n))$ has a "moderate" value. On the other hand if the exceedance probability $(1 - F(u_n))$ is small but the expected number of exceedances c_n is large, the level u_n

is regarded as "moderate". These correspond in the underlying limit theorems to the requirements $n(1-F(u_n)) \to \tau$ (some fixed finite τ) and $n(1-F(u_n)) \to \infty$, respectively.

From this it follows that high levels have relatively few exceedances (the expected number in fact approaching τ). On the other hand for moderate levels the expected number of exceedances is large, though small as a proportion of n, the total number of observed values. This is summarized as follows

Level	**Limit theorem requirement for threshold u_n**	**Practical implications**
Very high (CP Model)	$1-F(u_n) \to 0 \quad n(1-F(u_n) \to \tau < \infty.$	Small or moderate number of threshold exceedances.
Moderate (normal model)	$1-F(u_n) \to 0 \quad n(1-F(u_n)) \to \infty$	Large number of threshold exceedances but small as a proportion of number of number of observed X_i

This of course fits the situation illustrated in Figure 3.

3 Exceedance clusters

Any realistic model must be able to account for statistical **dependence** between nearby observed values X_i. Very often this involves high positive correlation between neighboring values, so that one high value tends to attract another, resulting in clusters of exceedances.

For very high levels clusters are often well defined in the obvious way from the first to the last exceedance in a group ("run clusters" of [2]). For lower levels or highly "oscillating" cases, the clusters are less obviously defined in this way, but a useful definition is that of a "block cluster". This is obtained by choosing a "block size" r_n and dividing the observed values $X_1, X_2, \ldots X_n$ into successive groups or "blocks" of length r_n, the block B_1 containing the first r_n values $X_1, X_2, \ldots X_{r_n}$, B_2 containing the next r_n, $X_{r_{n+1}}, X_{r_{n+2}}, \ldots X_{2r_n}$ and so on.

This is illustrated in Figure 4, consisting of the portion of the ozone data values of Figure 3 lying above .2 ppm with a block size $r_n = 30$ days. For the (very high) level $u_n = .27$ ppm clusters occur in blocks 5, 6, 8 (of size 1, 2, 1 respectively). These block clusters happen to be the case as run clusters, which need not necessarily be the case, but typically become increasingly so at high levels.

On the other hand for the level $u_n = .20$ ppm, block clusters occur in all but the first block and often consist of more than one run cluster. This illustrates the contrast with the high level case (e.g. $u_n = .27$) where clusters are infrequent and identifiable as single exceedance runs above the threshold.

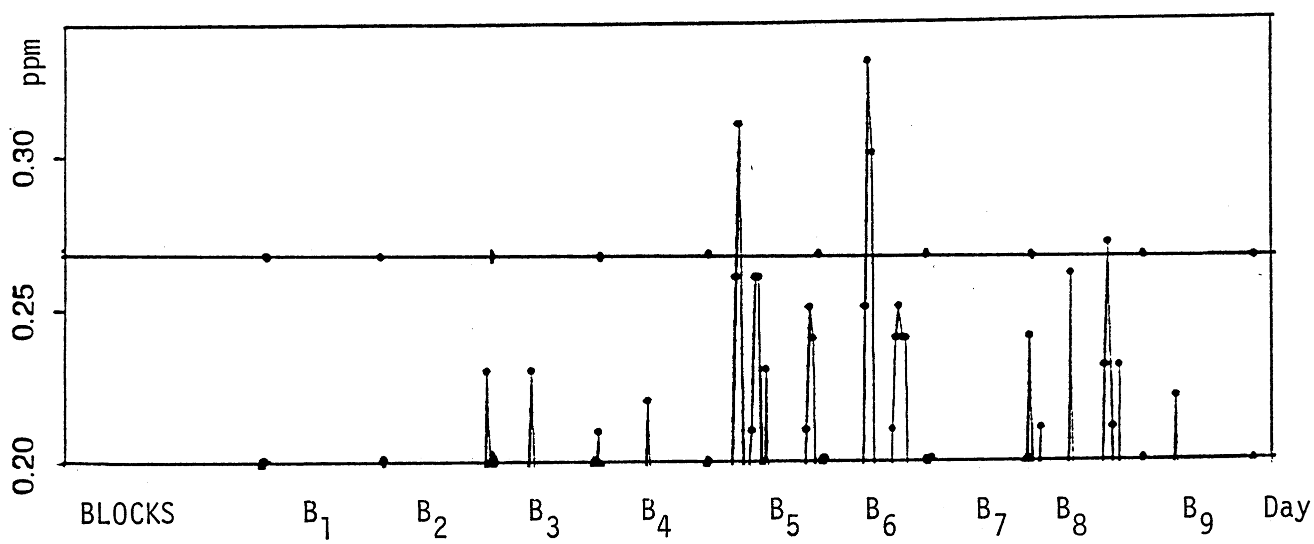

Figure 4: 1989 LA daily max 1 hour ozone levels above .2 ppm

The importance of the block clusters is that they tend to exhibit certain statistical independence properties even at moderate levels, which is not necessarily the case for run clusters. Of course at very high levels where the concepts coalesce, the run clusters are typically also block clusters and thus have the same independence properties.

In all cases therefore, block clusters provide the appropriate and tractable entities for statistical modeling. As will be seen in the following section the high and moderate level cases are contrasted by:

(i) In the high level case the cluster locations are described by Poisson occurrences, and their individual CS contributions (duration, area above threshold etc.) by independent random variables, with "general" distributions, whereas

(ii) For moderate levels exceedances can occur in many blocks and the sum of CS contributions is approximately normal.

Finally from a theoretical viewpoint a wide range of block size sequences are possible, subject only to a mild "growth" rate restriction in the limit theorem. In practice where the number n of observed values and level u_n are fixed it can be desirable to use several block sizes in performing statistical analyses (cf [4]).

4 Dependence, and a general result

As indicated, statistical dependence (e.g. serial correlation) is an essential ingredient in any realistic model for an environmental sequence X_i. This takes two forms – possible "long range dependence" between widely separated X_i values and "local" or "short range" dependence between nearby X_i and X_j. It is assumed that the former (long range) dependence falls off appropriately at long distances through a so called "strong mixing" condition discussed in detail in [5] while the local dependence may be quite high.

From the mixing condition one may obtain constants r_n to be used as block sizes. The blocks and clusters thus defined have useful approximate independence properties – described in the following informally stated result (see [5] for precise details and conditions).

Proposition 4.1 Under the strong mixing condition the contributions to the CS Z_n given by (1.1) from each block (i.e. $\sum_{j \in B_i} \psi_n(X_j - u_n)_+$) are approximately independent. Hence Z_n may be modeled as the sum of independent contributions from each block (i.e. each block cluster, since blocks without exceedances do not contribute). Thus the distribution of Z_n may be obtained (to a good approximation) from classical theory for sums of independent terms, namely the added contributions from each block (cluster).

The precise implications of the result for high and moderate levels are contained in the next two sections

5 High levels, CP models and the Ex-Ex criterion

For high levels u_n exceedances tend to occur in widely separated clusters. The expected cluster size (i.e. number of exceedances in a cluster) is "customarily denoted by θ^{-1}, $(0 < \theta \leq 1)$. If $n(1 - F(u_n)) \approx \tau$ the number C of clusters is an approximately Poisson r.v. with mean $\theta\tau$ by the theory of [5] i.e.

$$P(C = r) \approx e^{-\theta\tau}(\theta\tau)^r / r! \tag{5.1}$$

Again from [5] the contributions to the CS Z_n of (1.1) from each cluster are approximately independent with some distribution function G. Hence the total CS Z_n is the sum of the Poisson (mean $\theta\tau$) number of independent r.v.'s with common d.f. G. That is Z_n is **Compound Poisson** based on the Poisson mean $\theta\tau$ and the d.f. G and we write for brevity $Z_n = CP(\theta\tau, G)$. The distribution function for Z_n is easily written down in terms of $\theta\tau$ and G:

$$P\{Z_n \leq x\} = e^{-\theta\tau} \sum_{s=0}^{\infty} (\theta\tau)^s G_s(x)/s! \tag{5.2}$$

where G_s denotes the s-fold convolution of G with itself.

The above discussion applies especially to the Ex-Ex criterion since the threshold .12 ppm is high according to our definition, at compliant or near-compliance situations. In this case Z_n is modeled as $CP(\theta\tau, G)$ where now G is the distribution of cluster size and θ^{-1} is its mean. Z_n is integer valued and for an integer x the sum in (5.2) runs just from 0 to x.

It should be noted that the parameter θ and distribution G are typically unknown and require estimation. Some guidance concerning the general form of G is available from dependent central limit theory but the very high dependence possible within a cluster can invalidate the assumptions and it seems likely that quite general forms for G may be possible. Obvious estimates for G (and θ) are available, although extensive data may be required for their application (cf. [4])

Similar results hold for other criteria (e.g. AOT and SUM06 type) if used at high levels – the only difference being the replacement of G by the d.f. of the cluster contribution to the CS – i.e. the sum of values in the cluster period for the SUM06 case, and the sum of excess values for AOT, and of course θ^{-1} is still the mean cluster size and not now the mean of G. However the main application of these criteria is anticipated to be at lower levels. Indeed we believe that there are strong reasons in terms of "stability" to consider application of the Ex-Ex criterion at lower levels also. These will be discussed with the underlying normal theory in the next section.

Finally the Poisson properties of high level exceedances are sometimes ascribed to **independence** of the underlying X_i. However as seen above for dependent cases (the usual situation) the **cluster positions** now become Poisson and the d.f. G for the contribution of a cluster to the CS describes the feature of individual cluster structure relevant to that CS.

6 Moderate levels and normality; AOT, SUM06, and Ex-Ex at lower levels

As noted at lower levels u_n the expected number of exceedances $c_n = n(1 - F(u_n))$ is large and the "run clusters" are too frequent to exhibit Poisson occurrences through independence. However the block clusters are asymptotically independent, and this leads to normal models for the CS.

Specifically if r_n denotes the block size used before, it may be seen from Proposition 4.1 that the CS (1.1) has the same asymptotic distribution as it would if the contributions from individual blocks were independent. For these lower levels this distribution is **normal** under standard classical conditions (including an appropriate "Lindeberg condition"). More precisely the CS Z_n is approximately normal

$$Z_n \approx N(\mu_n, \sigma_n) \tag{6.1}$$

where μ_n and σ_n are its mean and standard deviation.

Criteria based on expected values (e.g. Ex-Ex, Expected AOT or SUM06) involve inference concerning μ_n. This may be done through a modification to the above asymptotic normality obtained ([5]) by replacing σ_n by an estimate s_n defined by

$$s_n^2 = \sum_{i=1}^{k_n}\Big(\sum_{j\in B_i} \phi_n(X_j - u_n)_+ - r_n m_n\Big)^2 \tag{6.2}$$

where

$$m_n = n^{-1}\sum \phi_n(X_j - u_n)_+ \tag{6.3}$$

The more specific results for the individual cases are:

(i) Ex-Ex

As noted the CS $Z_n^{(1)} = N_n$, the number of exceedances of u_n is usually modeled as a Poisson r.v. for very high levels u_n. However its behavior at more moderate levels is of interest (a) as a component of the SUM06 criterion and (b) in its own right for possible implementation of Ex-Ex at lower levels with enhanced criterion "stability".

The limiting approximation (6.1) for the distribution of N_n becomes (again writing $c_n = n(1 - F(u_n)$ for the expected number of exceedances),

$$N_n \approx N(c_n, \sigma_n) \tag{6.4}$$

with $\sigma_n^2 = var(N_n)$. More usefully σ_n may be replaced in this by its estimate s_n where

$$s_n^2 = \sum_{i=1}^{k_n} N_n^2(B_i) - N_n^2/k_n \tag{6.5}$$

where $N_n(B_i)$ is the number of exceedances in the ith of the k_n blocks (of length r_n used for defining clusters).

This modified form of (6.4) clearly enables estimation (and testing) for the expected number of exceedances c_n.

(ii) AOT

The AOT criterion may be couched in a similar way to the Ex-Ex in restricting the **expected area** rather than expected exceedances above the threshold. The expected area is given by

$$\beta_n = n\mathcal{E}(X_i - u_n)_+ = n\int_0^\infty (1 - F(x + u_n))dx$$

Then the AOT CS A_n has the approximately normal distribution

$$A_n \approx N(\beta_n, \sigma_n) \tag{6.6}$$

where now σ_n^2 is the variance of A_n and may be replaced in (6.6) by its estimate

$$\sum_{i=1}^{k_n} A_n^2(B_i) - A_n^2/k_n \tag{6.7}$$

$A_n(B_i)$ being the contiribution $(\sum_{j\in B_i}(X_j - u_n)_+)$ to the total AOT statistic A_n arising from the ith block B_i.

(iii) SUM06

As noted earlier the SUM06 criterion involves the CS $S_n = A_n + u_n N_n$, with expected value

(6.8) $$\gamma_n = \beta_n + u_n c_n.$$

Again this criterion may be regarded as a statistical test, $\gamma_n \leq c$ indicating compliance for appropriately chosen c. Corresponding to (6.6) we have

(6.9) $$S_n \approx N(\gamma_n, \sigma_n)$$

where now $\sigma_n^2 =$ var S_n and which may be replaced by

(6.10) $$\sum_{i=1}^{k_n} S_n^2(B_i) - S_n^2/k_n$$

in which correspondingly $S_n(B_i)$ is the contribution to the total CS S_n arising from the block B_i.

References

[1] Curran, T., Leadbetter, M.R., "Comments on the statistical properties of the Ex-Ex ozone criterion" in preparation

[2] Leadbetter, M.R., "On high level exceedance modeling and tail inference" University of North Carolina Center for Stochastic Processes Technical Report No. 388, March 1993, to appear in *J. Stat. Planning and Inference*

[3] Leadbetter, M.R., Rootzén, H., "On central limit theory for families of strongly mixing additive random functions" Festschrift in honour of G. Kallianpur, Springer N.Y. 1993 p. 211-223

[4] Leadbetter, M.R., Rootzén, H., Weissman, I. and de Haan, L., "On clustering of high values in statistically stationary series" *Proc. 4th Int. Mtg. on Stat. Climatology*, N.Z. Met. Service, 1889, 217 - 222

[5] Rootzén, H., Leadbetter, M.R., de Haan, L., "On the distribution of tail array sums for strongly mixing stationary sequences" in preparation

[6] Rootzén, H., Leadbetter, M.R., de Haan, L., "Tail and quantile estimation for strongly mixing stationary sequences", University of North Carolina Center for Stochastic Processes Technical Report No. 292, April 1990