MASTER

ESTIMATION OF THE RETURN PERIOD FOR GUMBEL PROCESSES
WITH APPLICATIONS TO AIR QUALITY MEASURES

GLENN GALFOND

TECHNICAL REPORT NO. 36

MARCH 1980

PREPARED UNDER THE AUSPICES OF
SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Prepared For
THE U.S. DEPARTMENT OF ENERGY
UNDER CONTRACT NO. DE-AS02-76EV02874

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Estimation of the Return Period for Gumbel Processes
with Applications to Air Quality Measures

by

Glenn Galfond
Stanford University

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

TECHNICAL REPORT NO. 36

MARCH 1980

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH (SIMS)

PREPARED UNDER SUPPORT TO SIMS FROM

Department of Energy (DOE)

Rockefeller Foundation

Sloan Foundation

Environmental Protection Agency (EPA)

National Science Foundation (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

EB

Acknowledgements

I would like to express my deepest gratitude to Professor Paul Switzer for his encouragement and assistance, Professor Yash Mittal for her invaluable suggestions and guidance, and Professor Ray Faith for his help in computer simulations.

I would also like to express my appreciation to my fiancée and my parents for their encouragement and support.

Introduction

The return period for a pollutant concentration to exceed a given level is often of interest. For example, the return period for exceeding government standards is a useful measure of long term air quality*. Or, for medical reasons, one might be interested in the return period for exceeding some significant harm level. This paper examines several estimates of the return period based on extreme value theory and illustrates applications to air quality data.

If the random variable Y has distribution function $F(y)$, then we define

$$T(x) = 1/(1 - F(x))$$

as the return period for exceeding x . The return period $T(x)$ is interpreted as the number for which, on the average, one observes one value greater than x in $T(x)$ observations.

In order to estimate the return period, it is necessary to model the distribution of the observations. When each observation is the maximum of a large number of random variables (for example, the maximum hourly average concentration of ozone during the period of a month or a year) we shall model its distribution by the Gumbel distribution

$$F(x) = \exp[-\exp[-(x-u)/\alpha]] \quad \begin{matrix} -\infty < x < \infty \\ \alpha > 0 \\ -\infty < u < \infty \end{matrix} \quad (1)$$

*Roberts [5] goes so far as to suggest that air quality standards be written in terms of acceptable return periods for exceeding given levels rather than absolute levels never to be exceeded.

For a wide class of distributions, the convergence in distribution of the normalized maximum of n identically distributed random variables, under the assumption of independence or some dependent conditions, to the Gumbel distribution is well known (see, e.g., de Hann [1], Leadbetter [3]). Applications of this property have been presented by Gumbel [2], and more recently to air quality data by Roberts [5].

While the performance of various estimates of the parameters u and α in (1) have been studied extensively (see, e.g., Mann [4]), the behavior of estimates of the return period has not been satisfactorily examined. The first part of this paper compares the performances of the maximum likelihood estimate and the least squares type estimate (used by Gumbel and Roberts) for estimating return periods, and suggests a third estimate.

The application of these estimates to air quality data is illustrated in the second part of this paper. In particular, a technique for using nonstationary monthly maximum concentrations is given.

Estimating Return Periods

Least Squares Type Method.

Let $x_1, x_2, \dots, x_n$ be a sample from the Gumbel distribution (1) and $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ the order statistics of the sample. Then we have the approximation

$$F(x_{(i)}) = \exp[-\exp - [(x_{(i)} - u)/\alpha]] \approx \frac{i}{n+1}$$

or

$$\frac{x_{(i)}^{-u}}{\alpha} \approx -\ln -\ln\left(\frac{i}{n+1}\right) \equiv y_i .$$

The least square type estimates are obtained by fitting a line to the plot of $x_{(i)}$ versus y_i . Roberts, following the procedure used by Gunbel, suggests estimating u and α by fitting the line which minimizes the vertical and horizontal departures. This results in the estimates

$$\hat{\alpha}_1 = S/\sigma \quad \hat{u}_1 = \bar{x} - \hat{\alpha}_1 \bar{y} \quad (2)$$

where

$$S^2 = (\overline{x^2} - \bar{x}^2)n/(n-1)$$

$$\sigma^2 = (\overline{y^2} - \bar{y}^2)$$

$$\overline{x^2} = \sum_{i=1}^n x_i^2/n$$

$$\overline{y^2} = \sum_{i=1}^n y_i^2/n$$

$$\bar{x} = \sum_{i=1}^n x_i/n$$

$$\bar{y} = \sum_{i=1}^n y_i/n .$$

Then $T(x)$, the return period for exceeding a value x , is estimated by

$$\hat{T}_1(x) = 1/(1-\exp[-\exp[-(x-\hat{u}_1)/\hat{\alpha}_1]]) . \quad (3)$$

The Method of MLE.

In order to find the MLE of the return period it suffices to find the MLE's of α and u . Finding such estimates always involves numerical approximation since the equations to be solved have no closed form solution. It is shown in Appendix 1 that the likelihood equations can be written in the form

$$0 = \frac{1}{\alpha} \left[\sum_{i=1}^n x_i \frac{e^{-x_i/\alpha}}{\sum_{j=1}^n e^{-x_j/\alpha}} - \bar{x} \right] + 1 \quad (4i)$$

$$u = -\alpha \ln \left(\frac{\sum_{i=1}^n e^{-x_i/\alpha}}{n} \right) . \quad (4ii)$$

This form has the advantage that only (4i) need be solved by numerical approximation. In addition, it is shown in Appendix 1 that (4i) is monotone increasing in α .

By the properties of the MLE, the MLE at $T(x)$, the return period for exceeding x , is given by

$$\widehat{T_2}(x) = 1/(1 - \exp[-\exp[-(x - \hat{u}_2)/\hat{\alpha}_2]])$$

where $\hat{\alpha}_2$ and $\hat{u}_2$ are the solutions to (4i) and (4ii). It is strongly recommended that (4) be solved on a computer, rather than by hand, since small errors in the estimate of α may yield large errors in estimates of the return period.

Simulation of the Performance of $T_1(x)$ and $T_2(x)$

Simulations based on 2,000 repetitions were performed to estimate the distribution of $\widehat{T}_1(x)$ and $\widehat{T}_2(x)$ for sample sizes $n = 10, 15, 20, 30, 50$ and various values of the return period. In all simulations the observed random variables $x_1, x_2, \dots, x_n$ were iid with distribution function $F(x) = \exp[-\exp(-x)]$ (corresponding to $\alpha = 1, u = 0$). These simulations are not intended to tabulate the exact distribution of $\widehat{T}_1(x)$ and $\widehat{T}_2(x)$, but rather are intended to give an approximate evaluation of their performance. The results of the simulations are presented in Table 1 below.

Table 1: Simulated distribution of $\widehat{T}_1(x)$ and $\widehat{T}_2(x)$

$n = 10$

	Mean	Variance
$\hat{\alpha}_1$:	1.284	.168
$\hat{u}_1$:	-.065	.122
$\hat{\alpha}_2$:	.921	.061
$\hat{u}_2$:	.034	.123

$T(x)$	$\widehat{T}_1(x)$ Mean	$\widehat{T}_1(x)$ S.D.
3.333	3.5	2.2
5.0	5.3	6.0
10.0	12.2	34.7
20.0	32.9	211.7

$T(x)$	$\widehat{T}_2(x)$ Mean	$\widehat{T}_2(x)$ S.D.
3.333	4.7	4.5
5.0	8.8	15.8
10.0	30.9	147.8
20.0	140.8	1409.6

Table 1 (continued)

n = 15

	<u>Mean</u>	<u>Variance</u>
$\hat{\alpha}_1$:	1.222	.106
$\hat{u}_1$:	-.049	.079
$\hat{\alpha}_2$:	.949	.041
$\hat{u}_2$:	.025	.078

$T(x)$	$\widehat{T_1(x)}$ Mean	$\widehat{T_1(x)}$ S.D.
5.0	5.0	3.0
7.5	7.5	6.1
15.0	15.6	20.0
30.0	34.2	65.4

$T(x)$	$\widehat{T_2(x)}$ Mean	$\widehat{T_2(x)}$ S.D.
5.0	6.8	5.4
7.5	11.7	12.9
15.0	30.6	56.8
30.0	85.8	261.6

n = 20

	<u>Mean</u>	<u>Variance</u>
$\hat{\alpha}_1$:	1.177	.072
$\hat{u}_1$:	-.040	.054
$\hat{\alpha}_2$:	.965	.031
$\hat{u}_2$:	.017	.053

$T(x)$	$\widehat{T_1(x)}$ Mean	$\widehat{T_1(x)}$ S.D.
5.0	4.9	2.3
6.6667	6.5	3.8
10.0	9.8	7.8
20.0	20.4	27.0
40.0	44.5	96.5

$T(x)$	$\widehat{T_2(x)}$ Mean	$\widehat{T_2(x)}$ S.D.
5.0	6.1	3.2
6.6667	8.6	5.6
10.0	14.2	12.2
20.0	34.2	45.5
40.0	86.2	172.7

Table 1 (continued)

n = 30

	<u>Mean</u>	<u>Variance</u>
$\hat{\alpha}_1$:	1.133	.044
$\hat{u}_1$:	-.035	.036
$\hat{\alpha}_2$:	.971	.020
$\hat{u}_2$:	.008	.035

$T(x)$	$\widehat{T_1(x)}$ Mean	$\widehat{T_1(x)}$ S.D.
7.5	7.1	2.9
10.0	9.4	4.4
15.0	14.1	8.0
30.0	28.3	21.1
60.0	58.1	55.1

$T(x)$	$\widehat{T_2(x)}$ Mean	$\widehat{T_2(x)}$ S.D.
7.5	9.0	4.1
10.0	12.5	6.7
15.0	19.8	12.9
30.0	44.5	38.9
60.0	101.4	117.4

n = 50

	<u>Mean</u>	<u>Variance</u>
$\hat{\alpha}_1$:	1.091	.027
$\hat{u}_1$:	-.018	.022
$\hat{\alpha}_2$:	.986	.013
$\hat{u}_2$:	.010	.022

$T(x)$	$\widehat{T_1(x)}$ Mean	$\widehat{T_1(x)}$ S.D.
10.0	9.5	3.6
12.5	11.8	5.0
16.667	15.8	7.5
25.0	23.7	13.2
50.0	47.8	33.9
100.0	97.8	85.2

$T(x)$	$\widehat{T_2(x)}$ Mean	$\widehat{T_2(x)}$ S.D.
10.0	11.2	4.3
12.5	14.3	6.0
16.667	19.5	9.2
25.0	30.4	16.5
50.0	65.0	44.6
100.0	140.4	118.7

Observations.

Studying Table 1, one observes that for all sample sizes the least squares estimate of $T(x)$ has smaller bias and standard deviation than the MLE of $T(x)$. This seems, at first glance, contradictory to the fact that the MLE's of α and u perform better than the least squares type estimates. In particular, while both estimates of u behave similarly, the MLE $\hat{\alpha}_2$ has lower bias, and approximately half the variance of the least squares type estimate $\hat{\alpha}_1$.

To explain this seemingly contradictory result one need examine the nature of $T(x)$. Using a one term Taylor expansion

$$\frac{-e^{-(\frac{x-u}{\alpha})}}{e^{-(\frac{x-u}{\alpha})}} \approx 1 - e^{-(\frac{x-u}{\alpha})}$$

so

$$T(x) = \frac{1}{1 - \exp\left[-e^{-(\frac{x-u}{\alpha})}\right]} \approx e^{\frac{x-u}{\alpha}}.$$

From the above we see that as α is decreased $T(x)$ is increased exponentially. Thus estimates of α which are moderately to the small side will lead to extremely large estimates of $T(x)$. Since the MLE of α is biased to the small side and the least square type estimate is biased to the large side, the MLE is more likely to produce small estimates for α even though it has a lower variance than the least square type estimate. When the MLE of α is multiplied by a constant to give it bias equal to the least square type estimate, the resulting estimate of $T(x)$ performs better than $\hat{T}_1(x)$. The optimal choice

of a constant is somewhat arbitrary. Therefore, a more unified approach is taken. The unifying criterion chosen is that of unbiasedness.

Nearly Unbiased Estimation of $T(x)$

From the above it is clear that, for reasonable sample sizes, the MLE of $T(x)$ does not enjoy the asymptotic properties of unbiasedness, and efficiency commonly associated with maximum likelihood estimates. The MLE's of u and α , however, do behave well. Mann [5] shows that even for small sample sizes $\hat{u}_2$ and $\hat{\alpha}_2$ are as good as, or better than, other commonly used estimates. Thus, it seems natural to work with $\hat{u}_2$ and $\hat{\alpha}_2$ for estimating $T(x)$. The constraint chosen for estimating $T(x)$ is that of unbiasedness. To this end we work with the parameters

$$\hat{m} = \frac{1}{\hat{\alpha}_2} \quad \text{and} \quad \hat{b} = \frac{-\hat{u}_2}{\hat{\alpha}_2}$$

which are the MLE's of $m = 1/\alpha$ and $b = -u/\alpha$. According to this reparameterization,

$$F(x) = \exp[-\exp[-(mx+b)]] \quad \begin{matrix} m > 0, & -\infty < x < \infty \\ -\infty < b < \infty \end{matrix}$$

and

$$T(x) = \frac{1}{1 - \exp[-e^{-(mx+b)}]}.$$

We shall assume that $\hat{m}$ and $\hat{b}$ are normal and unbiased (this is asymptotically correct; the validity of this approximation for small sample sizes will be examined later).

Given any level x for which $T(x)$ is to be estimated, we will be interested in

$$\hat{Y} = \hat{m}x + \hat{b}.$$

From our assumptions on $\hat{m}$ and $\hat{b}$ we have

$$\hat{Y} = \hat{m}x + \hat{b} \sim N(Y, \sigma_Y^2) \quad (5)$$

where $Y = mx + b$, $\hat{Y}$ is the MLE of Y , and σ_Y^2 is the variance of $\hat{Y}$. As done earlier we have

$$e^{-e^{-Y}} \approx 1 - e^{-Y}$$

so

$$T(x) = \frac{1}{1 - e^{-e^{-Y}}} \approx e^Y.$$

For values of $T(x)$ which we will be interested in, this approximation can be improved upon by adding $1/2$ to the right hand side yielding

$$T(x) = \frac{1}{1 - e^{-e^{-Y}}} \approx e^Y + .5 \quad (6)$$

To validate this approximation, Table 2 compares $T(x)$ to its approximation.

Table 2: Comparison of $T(x)$ with its approximation $e^Y + .5$

Y	$T(x) = 1/(1 - e^{-e^{-Y}})$	$e^Y + .5$
0.36651	2.000	1.943
0.90272	3.000	2.966
1.49994	5.000	4.981
2.25037	10.000	9.991
2.97020	20.000	19.996
3.90194	50.000	49.998
4.60015	100.000	99.999
5.29581	200.000	200.000

From (5) and (6) it is easily verified that

$$E \widehat{T}_2(x) \approx E(e^{\hat{Y}} + .5) = e^{Y + \sigma_Y^2/2} + .5.$$

Thus to obtain an unbiased estimate of $T(x)$ we need to know or estimate σ_Y^2 , the variance of $\hat{Y}$. To do this, we examine the variance/covariance structure of $\hat{m}$ and $\hat{b}$.

First let $x_1, x_2, \dots, x_n$ be a sample from the distribution $F(x) = \exp[-e^{-x}]$. Then the MLE's of $m (=1)$ and $b (=0)$ are $\hat{m}_0$ and $\hat{b}_0$, the solutions of

$$0 = \hat{m}_0 \left[\sum_{i=1}^n x_i \frac{e^{-\hat{m}_0 x_i}}{\sum_{j=1}^n e^{-\hat{m}_0 x_j}} - \bar{x} \right] + 1 \quad (7a)$$

$$\hat{b}_0 = \ln \left(\sum_{i=1}^n e^{-\hat{m}_0 x_i} / n \right). \quad (7b)$$

(See Appendix 1 for the likelihood equations.) From our assumptions,

$$\begin{pmatrix} \hat{m}_0 \\ \hat{b}_0 \end{pmatrix} \sim N \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right) \quad (8)$$

where the variance/covariance matrix for $\hat{m}_0$ and $\hat{b}_0$ can (and will) be found, so we shall assume they are known quantities. Using these quantities we can derive the variance/covariance matrix of $\hat{m}$ and $\hat{b}$ for arbitrary m and b , which in turn will lead to the value of σ_Y^2 .

Letting $Y_i = (x_i - b)/m$ we have

$$F_i(y) = P(Y_i < y) = P\left(\frac{x_i - b}{m} < y\right) = P(x_i < m y + b) = e^{-e^{-(my+b)}}$$

If one were to observe the Y_i and calculate the MLE's of m and b , from the likelihood equations we have

$$0 = \hat{m} \left[\sum_{i=1}^n Y_i \frac{e^{-\hat{m}Y_i}}{\sum_{j=1}^n e^{-\hat{m}Y_j}} - \bar{Y} \right] + 1$$

$$\hat{b} = \ln \left(\sum_{i=1}^n e^{-\hat{m}Y_i} / n \right)$$

Substituting $(x_i - b)/\hat{m}$ for Y_i and simplifying yields

$$0 = \frac{\hat{m}}{m} \left[\sum_{i=1}^n x_i \frac{e^{-\frac{\hat{m}}{m} x_i}}{\sum_{j=1}^n e^{-\frac{\hat{m}}{m} x_j}} - \bar{x} \right] + 1$$

$$\hat{b} = \ln \left(\sum_{i=1}^n e^{-\frac{\hat{m}}{m} x_i} / n \right) + \frac{\hat{m}}{m} b$$

Comparing the above to (7) leads to the solutions

$$\begin{aligned}\hat{m} &= m \hat{m}_0 \\ \hat{b} &= \hat{b}_0 + \hat{m}_0 b\end{aligned}\tag{9}$$

and (8) implies that for arbitrary m and b

$$\begin{pmatrix} \hat{m} \\ \hat{b} \end{pmatrix} \sim N \left(\begin{pmatrix} m \\ b \end{pmatrix}, \begin{pmatrix} m^2 \sigma_1^2 & m b \sigma_1^2 + m \sigma_{12} \\ m b \sigma_1^2 + m \sigma_{12} & \sigma_2^2 + 2b \sigma_{12} + b^2 \sigma_1^2 \end{pmatrix} \right).$$

Now we have

$$\begin{aligned}\sigma_Y^2 &= \text{var}(\hat{m} x + \hat{b}) \\ &= x^2 \text{var}(\hat{m}) + 2x \text{cov}(\hat{m} \hat{b}) + \text{var}(\hat{b}) \\ &= (mx + b)^2 \sigma_1^2 + 2(mx + b) \sigma_{12} + \sigma_2^2 \\ &= Y^2 \sigma_1^2 + 2Y \sigma_{12} + \sigma_2^2.\end{aligned}\tag{10}$$

Note that σ_1^2 , σ_{12} , and σ_2^2 are known quantities.

Under our model

$$E(\widehat{T_2(x)}) \approx E(c \hat{Y} + .5) = e^{Y + \sigma_Y^2/2} + .5 = e^{Y + (Y^2 \sigma_1^2 + 2Y \sigma_{12} + \sigma_2^2)/2} + .5 \tag{11}$$

where the quantity to be estimated is $T(x) \approx e^Y + .5$. A naive attempt to rid $\widehat{T_2(x)}$ of bias would be to modify it with an estimate of the bias.

This approach leads to the estimate

$$\widehat{T_3(x)} = e^{\hat{Y} - (\hat{Y}^2 \sigma_1^2 + 2\hat{Y} \sigma_{12} + \sigma_2^2)/2} + .5 \tag{12}$$

To see how good $\widehat{T_3(x)}$ is we need the following lemma.

Lemma 1. Let $\hat{Y} \sim N(Y, \sigma_Y^2)$. Then

$$E e^{a\hat{Y}^2 + b\hat{Y} + c} = \frac{e^{(aY^2 + bY + b^2\sigma_Y^2/2)/(1-2a\sigma_Y^2) + c}}{\sqrt{1-2a\sigma_Y^2}}$$

Proof: The proof of Lemma 1 is straightforward, one merely completes the square in the exponent to form a normal density.

For sample size n , each of σ_1^2 , σ_{12} , σ_2^2 , and σ_Y^2 are $O(\frac{1}{n})$.

Thus from (11) and (12) and Lemma 1 we have

$$\begin{aligned} E \widehat{T_2}(x) &= e^Y + .5 + O(\frac{1}{n}) \\ E \widehat{T_3}(x) &= e^Y + .5 + o(\frac{1}{n^2}). \end{aligned}$$

The above leads us to believe that $\widehat{T_3}(x)$ will have much smaller bias than the MLE $\widehat{T_2}(x)$. We shall see shortly that this is correct, but first the constants σ_1^2 , σ_{12} , and σ_2^2 must be estimated.

The Distribution of $\hat{m}_0$ and $\hat{b}_0$.

The asymptotic distribution of $\hat{m}_0$ and $\hat{b}_0$ is calculated in Appendix 2 using standard MLE theory. In addition, the distribution for sample sizes $n = 10, 15, 20, 30, 40, 50, 60, 80, 100$ was estimated by a simulation with 10,000 repetitions. The results are presented below.

Table 3: Estimation of the distribution of $\hat{m}_0, \hat{b}_0$

n	$E \hat{m}_0$	$E \hat{b}_0$	$n \sigma_1^2$	$n \sigma_{12}$	$n \sigma^2$
10	1.180	-.0280	1.290	.280	1.673
15	1.107	-.0122	.960	.273	1.458
20	1.075	-.0097	.843	.273	1.278
30	1.050	-.0081	.773	.264	1.244
40	1.037	-.0028	.745	.269	1.198
50	1.026	-.0030	.681	.260	1.183
60	1.023	-.0031	.656	.261	1.177
80	1.016	.0006	.665	.266	1.145
100	1.015	.0006	.642	.253	1.148
∞	1.	0	.60794	.25702	1.1087

To check the approximation of the distribution by normality, histograms were plotted. Since the ultimate test of this assumption is the performance of $\hat{T}_3(x)$, I chose not to test these histograms for normality. I merely summarize the results by stating that, with the exception of $n = 10$, all the histograms "looked" normal. For $n = 10$ the histogram of $\hat{m}_0$ was heavily skewed to the right.

Distribution of $\hat{T}_3(x)$

Table 3 illustrates that for small sample sizes $\hat{m}_0$ is biased while $\hat{b}_0$ is nearly unbiased. Since the derivation of $\hat{T}_3(x)$ was based on the assumption of unbiased estimation of m and b , and since errors in $\hat{m}$ and $\hat{b}$ lead to exponential errors in $\hat{T}(x)$, we will want to correct for the bias in $\hat{m}_0$.

From (9) and the approximate unbiasedness of $\hat{b}_0$ it follows that

$$E(\hat{m}) = E(m \hat{m}_0) = m E \hat{m}_0$$

$$E(\hat{b}) = E(\hat{b}_0 + \hat{m}_0 b) = E \hat{b}_0 + b E \hat{m}_0 \approx b E \hat{m}_0$$

Thus the unbiased estimates of m and b are

$$\hat{m}' = \hat{m}/E(\hat{m}_0) \quad (13i)$$

$$\hat{b}' = \hat{b}/E(\hat{m}_0) \quad (13ii)$$

Letting $\hat{Y}' = \hat{m}' x + \hat{b}'$, we have

$$\hat{Y}' \sim N(Y, (Y^2 \sigma_1^2 + 2Y \sigma_{12} + \sigma_2^2)/(E \hat{m}_0)^2) .$$

This leads to the estimate

$$\widehat{T_3(x)}' = e^{\frac{\hat{Y}' - (\hat{Y}'^2 \sigma_1^2 + 2\hat{Y}' \sigma_{12} + \sigma_2^2)/2(E\hat{m}_0)^2}{2}} + .5 \quad (14)$$

The distribution of $\widehat{T_3(x)}'$ was simulated for $n = 10, 15, 20, 30, 50$ and various values of $T(x)$. The simulation is based on 2,000 repetitions. The sample for each value of n is exactly the same sample as was used in the simulation of $T_1(x)$ and $T_2(x)$. The results are presented in Table 4 below.

Table 4: Mean and S.D. of $\widehat{T_3(x)}$

n = 10

T(x)	$\widehat{T_3(x)}$ Mean	$\widehat{T_3(x)}$ S.D.
3.333	3.2	1.5
5.0	4.6	2.9
10.0	8.6	6.8
20.0	14.3	12.7

n = 15

T(x)	$\widehat{T_3(x)}$ Mean	$\widehat{T_3(x)}$ S.D.
5.0	4.9	2.6
7.5	7.3	4.8
15.0	14.1	12.6
30.0	26.6	29.5

n = 20

T(x)	$\widehat{T_3(x)}$ Mean	$\widehat{T_3(x)}$ S.D.
5.0	4.9	2.1
6.6667	6.5	3.3
10.0	9.8	6.1
20.0	19.3	16.2
40.0	37.8	40.9

n = 30

T(x)	$\widehat{T_3(x)}$ Mean	$\widehat{T_3(x)}$ S.D.
7.5	7.5	3.0
10.0	9.9	4.5
15.0	14.9	8.0
30.0	29.5	20.3
60.0	58.4	49.5

n = 50

T(x)	$\widehat{T_3(x)}$ Mean	$\widehat{T_3(x)}$ S.D.
10.0	10.0	3.5
12.5	12.5	4.8
16.667	16.7	7.1
25.0	25.1	12.3
50.0	50.3	30.4
100.0	100.8	73.2

Observations.

With the exception of $n = 10$, $\widehat{T_3(x)}$ is nearly unbiased. Comparing $\widehat{T_3(x)}$ to $\widehat{T_1(x)}$ (see Table 5 below) it can be seen that $\widehat{T_3(x)}$ has smaller mean square error than $\widehat{T_1(x)}$ in almost all cases, and is especially better for sample sizes $n \leq 20$.

Table 5

n	T(x)	$\widehat{T_1(x)}$ MSE	$\widehat{T_3(x)}$ MSE
10	3.333	4.8	2.3
10	5.0	35.6	8.3
10	10.0	1210.0	50.0
10	20.0	45000.0	195.0
15	5.0	8.9	6.8
15	7.5	37.3	23.5
15	15.0	401.0	159.0
15	30.0	4300.0	833.0
20	5.0	5.2	4.2
20	6.6667	14.6	10.7
20	10.0	61.4	36.7
20	20.0	72.9	294.0
20	40.0	9330.0	1680.0
30	7.5	8.4	8.7
30	10.0	20.0	20.3
30	15.0	64.0	63.6
30	30.0	449.0	411.0
30	60.0	3040.0	2450.0
50	10.0	13.4	12.2
50	12.5	25.6	23.0
50	16.6667	57.7	50.1
50	25.0	177.0	152.0
50	50.0	1150.0	925.0
50	100.0	7260.0	5360.0

Applications

The applications section is concerned with three topics. First, application to air quality data is illustrated by applying the estimates $\widehat{T_1(x)}$ and $\widehat{T_3(x)}$ to the annual maximum hourly average concentration. Second, the performances of the estimates from the data are examined using the bootstrap method, and found to be in line with the theoretical results of the previous section. Finally, it is shown how one may greatly improve estimation of the return period by using monthly data, even when the pollutant is periodic over the year.

The data to be analyzed is found in Table 6. It contains the maximum hourly average concentration of carbon monoxide recorded in San Jose California from 1966 to 1978. From this data we wish to estimate the return period for exceeding the 25 ppm and 30 ppm levels.

Table 6: Maximum hourly average carbon monoxide concentrations (ppm) San Jose, California 1966-1978.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual max
1966	14	11	10	11	9	8	11	7	11	20	19	18	20
1967	29	17	8	6	9	4	6	8	9	26	21	20	29
1968	20	13	10	9	9	8	9	7	15	21	17	18	21
1969	15	12	10	6	9	5	7	7	12	23	26	22	26
1970	14	16	11	10	9	8	7	6	18	14	17	16	18
1971	21	22	9	8	8	7	6	8	15	18	22	21	22
1972	18	15	11	7	9	8	11	9	15	18	22	16	22
1973	19	14	10	10	8	10	9	10	15	19	16	23	23
1974	23	20	12	9	5	6	6	5	8	18	21	15	23
1975	16	14	8	8	7	5	7	12	14	16	17	23	23
1976	25	17	9	9	9	9	4	7	10	18	27	29	29
1977	16	15	10	10	6	4	8	10	6	15	24	18	24
1978	15	12	9	11	8	7	7	8	13	17	27	19	27
Average	19	15	10	9	8	7	8	8	12	19	21	20	24

Use of the Annual Maxima

For a wide class of distributions (including the lognormal, gamma, normal, logistic, and others) it has been shown that the normalized maximum of n identically distributed random variables converges in distribution to the Gumbel distribution. Clearly, the hourly average concentration of CO is not identically distributed over the year since it exhibits daily as well as seasonal nonstationarity. However, if we were to view the annual maximum as the maximum of, say, the thirty random variables X_i where X_i is the maximum hourly average over the i th days in each month (ignoring the fact that not all months have thirty days) then the idea of the annual maximum being the maximum of identically distributed random variables might seem more palatable. A good way to check our distributional assumptions is with an extreme probability plot.

An extreme probability plot (see Gumbel [2]) is a plot of $X_{(i)}$, the i th order statistic, versus $y_i = -\ln(-\ln \frac{i}{n+1})$. If the underlying distribution is Gumbel, then the plot should look linear. This technique is particularly useful for determining outliers. Figure 1 is an extreme probability plot of the annual maxima from Table 6. The line was fitted visually. The line fits the data fairly well, so we accept the theory.

Least squares type fit: From (2) we have

$$\hat{m} = \frac{1}{\hat{\alpha}_1} = .299$$
$$\hat{b} = \frac{\hat{u}_1}{\hat{\alpha}_1} = -6.56$$

and

Figure 1

$$\hat{T}_1(x) = \frac{1}{1 - \exp[-\exp[-(.299x - 6.56)]]} .$$

Thus $\hat{T}_1(25) = 3.0$ and $\hat{T}_1(30) = 11.7$ are the least squares type estimates for $T(25)$ and $T(30)$.

$\hat{T}_3(x)'$: Using (4) we have the maximum likelihood estimates of m and b are

$$\hat{m} = .351$$

$$\hat{b} = -7.75 .$$

From Table 3 we interpolate linearly to get that, for $n = 13$

$$E \hat{m}_0 \approx 1.136$$

$$n \sigma_1^2 \approx 1.092 \quad \text{or} \quad \sigma_1^2 \approx .084$$

$$n \sigma_{12} \approx .276 \quad \text{or} \quad \sigma_{12}^2 \approx .021$$

$$n \sigma_2^2 \approx 1.544 \quad \text{or} \quad \sigma_2^2 \approx .119 .$$

Correcting $\hat{m}$ and $\hat{b}$ for bias, (13) yields

$$\hat{m}' = .351/1.136 = .309$$

$$\hat{b}' = -7.75/1.136 = -6.82$$

Thus $\hat{y}'_{25} = \hat{m}'(25) + \hat{b}' = .905$ and $\hat{y}'_{30} = \hat{m}'(30) + \hat{b}' = 2.45$. Finally,

(14) implies

$$\begin{aligned} \hat{T}_3(25)' &= .5 + \exp[\hat{y}'_{35} - (\hat{y}'_{35}{}^2 \sigma_1^2 + 2\hat{y}'_{35} \sigma_{12} + \sigma_2^2)/(2(1.136)^2)] \\ &= .5 + \exp[.818] \\ &= 2.8 \end{aligned}$$

Similarly $\hat{T}_3(30)' = 9.3$.

Comparison of $\widehat{T_1(x)}$ and $\widehat{T_3(x)}$

Samples of size thirteen were chosen at random from the empirical distribution of the annual maximum in Table 6. The mean and standard deviation were estimated using 1000 repetitions. The results were

	<u>mean</u>	<u>S.D.</u>
$\widehat{T_1(25)}$	3.4	1.24
$\widehat{T_3(25)}$	3.1	1.04
$\widehat{T_1(30)}$	19.7	59.9
$\widehat{T_3(30)}$	14.1	12.8

These results seem to agree with the simulations of the last section in that, for small sample sizes, $\widehat{T_3(x)}$ has much smaller variance in predicting large return periods. The variance of our estimate can further be reduced by considering the monthly data.

Using the Monthly Maxima

Though relatively easy to analyze, the annual maximum concentrations contain only a small portion of the information available. In order to utilize more of the available information, one may analyze the monthly maximum concentrations. Due to the seasonal nonstationarity of pollutant concentration, as well as the correlation between monthly maxima, the analysis becomes more complicated. An estimation procedure which deals with the nonstationarity, but not the correlation of the monthly maxima is detailed below. Its performance is evaluated via bootstrap simulation

and shown to be considerably better than the two previous estimates of the return period based on the annual maxima.

Since the monthly maximum concentration is the maximum of roughly thirty identically distributed random variables, it is natural to model its distribution with the Gumbel distribution. We will partition the year into N groups of months which behave similarly. Treating each group as an iid sample from a Gumbel distribution, we estimate

$$T_i(x) = \frac{1}{P(\max \text{ in group } i > x)}$$

by $\widehat{T_i(x)}$.

If, for a given year, the probability of exceeding x in any two groups of months simultaneously is small, then

$$P(\text{Yearly max} > x) \approx \sum_{i=1}^N P(\max \text{ in group } i > x)$$

and

$$T(x) = \frac{1}{P(\text{Yearly max} > x)} \approx \frac{1}{\sum_{i=1}^N P(\max \text{ in group } i > x)} = \frac{1}{\sum_{i=1}^N \frac{1}{T_i(x)}}. \quad (15)$$

We therefore estimate $T(x)$ by

$$\widehat{T(x)}^* = \frac{1}{\sum_{i=1}^N \frac{1}{\widehat{T_i(x)}}}.$$

To illustrate this procedure we analyze the data in Table 6. Considering the average monthly maximum concentration, it seems natural to form the three groups high, medium, and low concentration.

Group 1: January, October, November, December

Group 2: February, September

Group 3: March, April, May, June, July, August.

We graph the data for groups one and two in Figures 2 and 3 respectively to check the extreme value distribution assumption. Group three is not plotted since its contribution to the overall estimate will be seen to be negligible.

Treating each group separately, the return period for exceeding the 30 ppm level is estimated using (14). The result is

Group	$\widehat{T_3(30)'} \text{ (in months)}$
1	40.8
2	89.8
3	76242

The above figures represent the number of observations in which one expects to observe one exceedence of 35 ppm. Since, for example, Group 1 has 4 observations per year, this means one would expect an exceedence of 30 ppm. for some month in group 1 every $40.8/4 = 10.2$ years. Similarly we convert the return periods in the other groups from months to years resulting in

Figure 2

Extreme Probability Plot for Group 1

Figure 3

Extreme Probability Plot for Group 2

Group	$\widehat{T_3(30)'} \text{ (in years)}$
1	10.2
2	44.9
3	12707

Then, according to (15), we estimate $T(35)$ by

$$\widehat{T(35)}^* = \frac{1}{\frac{1}{10.2} + \frac{1}{44.9} + \frac{1}{12707}} = 8.3 .$$

In order to compare this estimate with the previous estimates based on the annual maxima, a bootstrap simulation with 1000 repetitions was performed. For each repetition an entire year's data was selected at random from Table 6 thirteen times. These, and the previous results are summarized in Table 7.

Table 7: Bootstrap simulation, size 1000 for estimates of $T(30)$ based on data from Table 6.

	<u>value based on Table 6</u>	<u>sampling mean</u>	<u>sampling S.D.</u>
$\widehat{T_1(30)}$	11.7	19.7	59.9
$\widehat{T_3(30)'}$	9.3	14.1	12.8
$\widehat{T(30)}^*$	8.3	9.0	4.6

Summary and Precautions.

It is clear, in this case, that using the monthly maxima greatly reduces the variance, and probably the bias, in estimation of the return period. To assume this to be true in general would be fallacious. Among the factors that influence the relative performances of estimates based on the yearly and monthly maxima are

a) Distributional assumptions.

One would expect the annual maxima, being the maximum of a larger number of observations, to follow the Gumbel distribution more closely. It is therefore important to plot the data on an extreme probability graph in order to verify distributional assumptions. When using the monthly data, groups which have little influence on the overall estimate need not be considered.

b) Correlation between data points.

The possible effects of correlation between the monthly, as well as the yearly maxima has not been considered. One would expect the effect to be greater when analyzing the monthly maxima due to the higher correlation among data points.

c) Grouping of months.

How the months are grouped will affect the behavior of $\hat{T}(x)^*$.

The more months in each group, particularly in the group containing the highest concentrations, the lower the variance of $\hat{T}(x)^*$. Grouping months which do not have the same distribution, however, will bias $\hat{T}(x)^*$.

Appendix 1

a) The Likelihood Equations.

In order to find the MLE's of u and α , it suffices to find the MLE's of any one to one transformation of u and α . For the sake of simplicity the parameters

$$m = \frac{1}{\alpha} \quad b = \frac{-u}{\alpha}$$

are chosen.

Under this parameterization the density function is

$$f(x_1 \dots x_n) = \prod_{i=1}^n m e^{-(m x_i + b)} e^{-e^{-(m x_i + b)}} \quad m > 0, \quad -\infty < x_i < \infty.$$

The log-likelihood is (for positive m)

$$L(x_1 \dots x_n) = \log f(x_1 \dots x_n) = -e^{-b} \sum_{i=1}^n e^{-m x_i} - m \sum_{i=1}^n x_i - n b + n \log m$$

$$\frac{dL}{db} = 0 \Leftrightarrow 0 = e^{-b} \sum_{i=1}^n e^{-m x_i} - n \Leftrightarrow b = \log \left(\frac{\sum_{i=1}^n e^{-m x_i}}{n} \right) \quad (1.1)$$

$$\frac{dL}{dm} = 0 \Leftrightarrow 0 = e^{-b} \sum_{i=1}^n x_i e^{-m x_i} - \sum_{i=1}^n x_i + \frac{n}{m} \quad (1.2)$$

from (1.1) we have $e^{-b} = n / \sum_{i=1}^n e^{-m x_i}$ so substitution into (1.2) and simplification yield

$$\frac{dL}{db} = \frac{dL}{dm} = 0 \Leftrightarrow \begin{cases} b = \log \left(\sum_{i=1}^n e^{-mx_i/n} \right) & (1.3i) \\ 0 = m \left[\sum_{i=1}^n x_i \frac{e^{-mx_i}}{\sum_{j=1}^n e^{-mx_j}} - \bar{x} \right] + 1 & (1.3ii) \end{cases}$$

In terms of u and α the right hand side is equivalent to

$$u = -\alpha \log \left(\sum_{i=1}^n e^{-x_i/\alpha} / n \right) \quad (1.4i)$$

$$0 = \frac{1}{\alpha} \left[\sum_{i=1}^n x_i \frac{e^{-x_i/\alpha}}{\sum_{j=1}^n e^{-x_j/\alpha}} - \bar{x} \right] + 1 \quad (1.4ii)$$

b) Monotonicity.

To show that the right hand side of (1.4ii) is monotone increasing in α , we need only show the right hand side of (1.3ii) is monotone, decreasing in m since $m = 1/\alpha$. Let

$$f(m) = \bar{x} - \sum_{i=1}^n x_i \frac{e^{-mx_i}}{\sum_{j=1}^n e^{-mx_j}},$$

then we need to show $-m f(m) + 1$ is decreasing in m .

$$f'(m) = \sum_{i=1}^n x_i^2 \frac{e^{-mx_i}}{\sum_{j=1}^n e^{-mx_j}} - \left(\sum_{i=1}^n x_i \frac{e^{-mx_i}}{\sum_{j=1}^n e^{-mx_j}} \right)^2.$$

Since the above is of the form $E(Y^2) - (EY)^2$ where Y is the discrete random variable with density

$$P(Y_i = x_i) = \frac{e^{-mx_i}}{\sum_{j=1}^n e^{-mx_j}} \quad i = 1, \dots, n.$$

We know $f'(m) > 0$ whenever not all the x_i are equal, and thus is increasing in m . Because $f(0) = 0$, f is nonnegative for $m \geq 0$.

Thus $mf(m)$ is the product of two nonnegative increasing functions and is therefore increasing. This implies $-mf(m) + 1$ is decreasing in m , as was to be shown.

The monotonicity of the likelihood equation guarantees a unique solution for m (or α) which is positive. In addition, it enables one to write simple algorithms for the solution of the likelihood equations.

Appendix 2

Asymptotic Covariance Matrix for $\hat{m}_0$ and $\hat{b}_0$.

For the Gumbel distribution,

$$F(x) = \exp[-e^{-(mx+b)}]$$

$$\log f(x) = -e^{-(mx+b)} - mx - b + \log m.$$

Differentiation yields

$$\frac{\partial^2}{\partial m^2} \log f(x) = -x^2 e^{-(mx+b)} - \frac{1}{m^2}$$

$$\frac{\partial^2}{\partial m \partial b} \log f(x) = -x e^{-(mx+b)}$$

$$\frac{\partial^2}{\partial b^2} \log f(x) = -e^{-(mx+b)} .$$

Let I be the asymptotic variance/covariance matrix for the maximum likelihood estimates $\hat{m}$ and $\hat{b}$, and I^0 be the value of I when $m = 1$ and $b = 0$. Then by definition, I^0 is the asymptotic variance/covariance matrix for $\hat{m}_0$ and $\hat{b}_0$ and

$$I_{11}^0 = I_{11} \Big|_{\substack{m=1 \\ b=0}} = E \left(\frac{\partial^2}{\partial m^2} \log f(x) \right) \Big|_{\substack{m=1 \\ b=0}}$$

$$I_{12}^0 = I_{12} \Big|_{\substack{m=1 \\ b=0}} = E \left(\frac{\partial^2}{\partial m \partial b} \log f(x) \right) \Big|_{\substack{m=1 \\ b=0}}$$

$$I_{22}^0 = I_{22} \Big|_{\substack{m=1 \\ b=0}} = E \left(\frac{\partial^2}{\partial b^2} \log f(x) \right) \Big|_{\substack{m=1 \\ b=0}} .$$

Note that when $m = 1$ and $b = 0$, $f(x) = e^{-e^{-x}} e^{-x}$, $E(x) = \gamma = .57722$ (Euler's constant), and $\text{Var}(x) = \pi^2/6$. Therefore, we have

$$I_{22}^0 = \int_{-\infty}^{\infty} e^{-x} e^{-e^{-x}} e^{-x} dx$$

letting $z = e^{-x}$ $|dz| = e^{-x}$,

$$I_{22}^0 = \int_{\infty}^0 z e^{-z} dz = -e^{-z}(z-1) \Big|_{\infty}^0 = 1.$$

Also from the above,

$$I_{12}^0 = \int_{-\infty}^{\infty} x e^{-x} e^{-e^{-x}} e^{-x} dx$$

letting

$$\begin{aligned} u &= x e^{-x} & du &= (e^{-x} - x e^{-x}) dx \\ dv &= e^{-e^{-x}} e^{-x} dx & \Rightarrow v &= e^{-e^{-x}} \end{aligned}$$

and integrating by parts,

$$\begin{aligned} I_{12}^0 &= x e^{-x} e^{-e^{-x}} \Big|_{-\infty}^{\infty} = \int_{-\infty}^{\infty} e^{-x} e^{-e^{-x}} dx + \int_{-\infty}^{\infty} x e^{-x} e^{-e^{-x}} dx \\ &= 0 - 1 + E(x) = \gamma - 1 = -.42278. \end{aligned}$$

Finally,

$$I_{11}^0 = \int_{-\infty}^{\infty} x^2 e^{-x} e^{-e^{-x}} e^{-x} dx + 1$$

letting

$$\begin{aligned} u &= x^2 e^{-x} & du &= (2x e^{-x} - x^2 e^{-x}) dx \\ dv &= e^{-e^{-x}} e^{-x} dx & \Rightarrow v &= e^{-e^{-x}} \end{aligned}$$

and integration by parts implies

$$\begin{aligned} I_{11}^0 &= 1 + x^2 e^{-x} e^{-e^{-x}} \Big|_{-\infty}^{\infty} - 2 \int_{-\infty}^{\infty} x e^{-x} e^{-e^{-x}} dx + \int_{-\infty}^{\infty} x^2 e^{-x} e^{-e^{-x}} dx \\ &= 1 + 0 - 2 E(x) + E x^2 \\ &= 1 - 2\gamma + \left(\frac{\pi^2}{6} + \gamma^2\right) \\ &= 1.82364 . \end{aligned}$$

Thus,

$$I^0 = \begin{bmatrix} 1.82364 & -.42278 \\ -.42278 & 1 \end{bmatrix}$$

and

$$I^{0-1} = \begin{bmatrix} .60794 & .25702 \\ .25702 & 1.1087 \end{bmatrix} .$$

Bibliography

- [1] De Haan, L. (1976). Sample extremes: an elementary introduction. Statistica Neerlandica 30, 161-172. ~~~
- [2] Gumbel, E. J. (1954). Statistical theory of extreme values and some practical applications. Nat'l Bur. Standards, Appl. Math. Series 33.
- [3] Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1979). Extremal and related properties of stationary processes. Statistical Research Report No. 1979-2, University of Umeå, Sweden.
- [4] Mann, N. R. (1968). Point and interval estimation procedures for the two-parameter Weibull and extreme-value distributions. Technometrics 10, 231-255. ~~~
- [5] Roberts, E. M. (1979). Review of statistics of extreme values with applications to air quality data. JAPCA 29, 733-740. ~~~