An Alternative Approach for Extreme Value Statistics of Climatologic Dependent Variables in Small Samples

Z. Sen

With 1 Figure

Received January 22, 1997
Revised April 7, 1997

Summary

Extreme event analysis of meteorological variables by conventional statistical methods does not meet the necessary mathematical assumptions such as the independence of successive occurrences. An alternative method is proposed for analyzing extreme values of dependent variables by considering all the extreme values above a threshold value in finite samples. The probability for the maximum event in a given finite sequence of dependent meteorological variables is derived explicitly.

1. Introduction

Mankind is very sensitive to natural extreme phenomena and therefore scientific knowledge prior to their occurrences is of prime importance. Extreme values of various meteorological variables provide a key for significant decisions in planning, design and operation of numerous hydrological, agricultural and industrial activities. Consequently, their modeling and subsequent prediction with the minimum possible error is the main objective of existing extreme value statistics theory. (Gumbel, 1958; Fisher and Tippett, 1928). Construction of a convenient model which reflects the structural dependence and variation of observed phenomena is a first step in any prediction scheme. It seems that climate changes are associated not with higher intensity of extreme values but rather with a higher frequency of the occurrence of extreme values, (Flohn, 1989).

In general, the more persistent (dependent) the meteorological sequence, the more frequent the extreme events. To the best of author’s knowledge all of the present conventional methods of extreme value statistics presume independence and hence the question arises whether they are sufficient to obtain significant evidence on the extreme value predictions in the case of dependent processes. Gerstengarbe and Werner (1991) have shown various evidences that the conventional extreme value statistics are insufficient to meet all requirements and especially dependence of events leads to over-estimation in future predictions. Consequently, they suggest possible development of new methods.

The main purpose of this paper is to present a new method by considering the frequency and correlation structure in generation of extreme values. It is a departure from the classical extreme value statistics in that it presents theoretical foundations of extreme value evaluations in dependent meteorological variables. The method proposed is not only an empirical procedure for checking the applicability of classical extreme value statistics, but a procedure for checking independence of a sequence of random variables against a Markov model.
2. Conventional Extreme-Value Statistics

Let at any time interval, \( i \), the meteorological variable, \( X_i \), have a cumulative probability distribution (CPD) function \( F_i(x_0) = P(X_i \leq x_0) \), i.e., the probability of \( X_i \) being less than or equal to a threshold value, \( x_0 \). According to the addition rule of the probability theory for complementary events, \( P(X_i > x_0) = 1 - F_i(x_0) \).

The probability of global maximum meteorological variable, \( X_M \), to be less than or equal to \( x_0 \) during time period of \( n \) intervals can be written formally as

\[
P(X_M \leq x_0) = P(X_1 \leq x_0, X_2 \leq x_0, \ldots, X_n \leq x_0)
\]

The graphical representation of this expression is shown in Fig. 1 where there is no crossing of \( x_0 \) level during period \( n \). Equation (1) is the most general case that requires the multivariate probability distribution function (PDF) of meteorological variable at \( n \) distinct intervals and it is almost impossible practically to obtain such a distribution from historical data. Therefore, researchers have dealt with a simple form by assuming that successive occurrences of the meteorological variable are independent. In such a case the multiplication rule of probability theory leads to

\[
P_I(X_M \leq x_0) = \prod_{i=1}^{n} P(X_i \leq x_0)
\]

Herein, the index \( I \) denotes “independence”. In this expression the correlation structure of meteorological variables is assumed as completely independent. On the other hand, further simplification of Eq. (2) is possible if the meteorological random variables at each time interval are considered as statistically indistinguishable and identically distributed. In this case, Eq. (2) reduces to

\[
P_I(X_M \leq x_0) = P^n(X \leq x_0)
\]

where \( X \) is a stationary meteorological variable throughout period \( n \) and with the CPD function concept Eq. (3) becomes

\[
F_M(x_0) = P^n(x_0)
\]

By taking derivatives of both sides this expression can be written in terms of probability density function (PDF) as

\[
f_M(x_0) = nP^{n-1}(x_0)f(x_0)
\]

Provided that the lower boundary, say \( \alpha \), of a meteorological variable is very small but never equal to zero, Eq. (4) gives a set of straight lines on double logarithmic paper as

\[
\log F_M(x_0) = n\log F(x_0) \quad \alpha < x_0 < +\infty
\]

where \( n \) can never be considered as zero. However, if the number \( n \) is large then utilization of Eq. (4) is extremely difficult. Hence, the study of the asymptotic properties of distribution as \( n \to \infty \) is of great importance since they may be used as an approximate estimate for finite but sufficiently large \( n \) values. Under rather general assumptions the distribution function of the appropriately standardized maximum \( M \) of a large number of independent random variables may be approximated by one of the following three distribution functions (Fisher and Tippett, 1928)

\[
F_1(x) = \exp[-\exp(-x)] \quad (-\infty < x < +\infty)
\]

\[
F_2(x) = \begin{cases} 0 & (x < 0) \\ \exp(-x^{-a}) & (a > 0), \quad (x > 0) \end{cases}
\]

and

\[
F_3(x) = \begin{cases} \exp[-(-x)^a] & (a > 0), \quad (x < 0) \\ 0 & \text{otherwise} \end{cases}
\]

in which \( F_i(x) \ i = 1,2,3 \) is the cumulative probability distribution function; \( 'a' \) is a constant value; and finally, \( x \) is the standard variable related to the main variable \( X \) as

\[
X = \mu + \beta x
\]

where \( \mu \) is the mean value and \( \beta \) is the frequency factor. The validity of expressions in Eq. (7) is possible under the following assumptions only:

(i) that there is a large number of observations,
(ii) that the meteorological variables are identically distributed,
(iii) that there is no serial correlation between successive occurrences of extreme values, in other words, they are independent from each other as the classical extreme value theories require.

On the other hand, it is a well known fact that e.g. daily, monthly, etc., meteorological variables are rarely independent and cause a number of problems in extreme value statistical analysis. These various problems are well documented in a study by Gerstengarbe and Werner (1991).

### 3. Alternative Development

In practice, extreme values of any meteorological phenomenon are extracted from a continuous record of observations by considering either a finite time period or a certain truncation level. In the former case extreme-values for longer durations are obtained from relatively shorter duration records. For instance, if hourly measurements of a meteorological variable are available for many years then the extreme value series can be obtained depending on the basic period adopted such as day, week, month or year. It is obvious that the number of extreme-values is equal to the number of non-overlapping basic periods within the record duration. A major disadvantage of such an approach is that the second largest value within a basic period is ignored although it might be greater than the extreme-value of some other periods.

However, a second procedure which can be used in any meteorological study is the consideration of all the values above a preselected threshold value, $x_0$, as shown schematically in Fig. 1. No doubt, application of this scheme may give rise to two or more extreme values that may be adjacent, and hence, highly dependent. Increase in the threshold value leads not only to a decrease in the number of extreme values but also to a decrease in their dependence of successive extreme values due to the increased duration between successive extremes. Hence, it is to be expect that with the increase of threshold value, conventional extreme value statistics becomes more applicable.

Let us consider, in general, different probabilistic characteristics that result from the truncation of a record at a given threshold level, $x_0$. As can be seen from Fig. 1 there are over- and under-crossings of the truncation level and hence, basic truncation variables emerge from further considerations. Two of these four basic truncation variables are, the relative values of exceeding, $(X > x_0)$, and non-exceeding, $(X \leq x_0)$ events. Two other variables are non-interrupted durations of exceedences, $L_e$, and non-exceedence, $L_n$. The probability occurrences of these events can be related to the basic probability distribution function, $f(X)$, of meteorological variable, $X$, as the probability or exceedence

$$p(x_0) = P(X > x_0) = \int_{x_0}^{+\infty} f(X) dX$$

and probability of non-exceedence as

$$q(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(X) dX$$

respectively. Due to mutual exclusiveness $p(x_0) + q(x_0) = 1$. The most general measure of dependence between successive occurrences is achieved by the definition of conditional probability as follows

$$P(X_i > x_0/X_{i-1} > x_0) = P(X_i > x_0, X_{i-1} > x_0)/P(X_{i-1} > x_0)$$

This concept has already been used as a measure of dependence by Sen (1978). For independent case the joint probability in the numerator factorizes, and consequently, if the sequence is independent then

$$P(X_i > x_0/X_{i-1} > x_0) = P(X_i > x_0)$$

otherwise for dependent processes at least for some $x_0$ we obtain

$$P(X_i > x_0/X_{i-1} > x_0) \neq P(X_i > x_0)$$

Given a sequence of meteorological measurements Eqs. (12) and (13) or (11) can be verified at any desired threshold value. Hence, the truncation level which satisfies Eq. (12) approximately, say with 5% relative error, yields situation whereby the classical extreme-value statistics can be employed satisfactorily. Otherwise, if Eq. (13) is valid then the extreme values are dependent and therefore the theory of extreme values must be employed as described in the following section.
If there is a simple dependence in the form of first order Markov chain then dependent version of Eq. (2) from Eq. (1) becomes

$$P_D(X_M \leq x_0) = P(X_1 \leq x_0) \times \prod_{i=2}^{n} P(X_i \leq x_0/X_{i-1} \leq x_0)$$

(14)

Herein, the index D denotes “dependence”. The conditional probability on the right hand side of this expression constitutes the basis of autorun concept already proposed by Sen (1978) and it is equal to the first order autorun coefficient, \(r(x_0) = P(X_i \leq x_0/X_{i-1} \leq x_0)\). Of course, \(0 < r(x_0) < 1\). On the other hand, since \(q(x_0) = P(X_1 \leq x_0)\) Eq. (14) leads to,

$$P_D(X_M \leq x_0) = q(x_0) \prod_{i=2}^{n} r_i(x_0)$$

(15)

However, for stationary stochastic processes autorun coefficients become independent of time, and consequently, Eq. (15) takes its simplest form as

$$P_D(X_M \leq x_0) = q(x_0) r^{n-1}(x_0)$$

(16)

This is the fundamental expression in calculating the PDF of extreme values in finite dependent samples. Given a sequence of observations the extreme value PDF can be calculated by executing the following steps.

(i) Find the range value which is the difference between the maximum and minimum values within the series of observations of length \(n\).
(ii) Truncate the series within the range at different levels starting from the minimum value. To this end, the range can be divided into a fixed number of classes, not less than five, and accordingly subsequent truncation levels are set up.
(iii) Calculate for each truncation level the values of \(q(x_0)\) and \(r(x_0)\) and plot them versus truncation levels.
(iv) Fit smooth curves through the scatter points so as to find the empirical relationship between pairs \([x_0, q(x_0)]\) and \([x_0, r(x_0)]\) and if needed, also \([q(x_0), r(x_0)]\). Under the null-hypothesis of independence the plots of \([x_0, q(x_0)]\) and \([x_0, r(x_0)]\) should show identical graphs and \([q(x_0), r(x_0)]\) should show a straight line with angle 45 degrees. This step is similar to the theoretical distribution function fitting of the frequency distribution where there are different tests for its confidence such as the Chi-square and Kolmogorov tests.
(v) If the functions \(q(x_0)\) and \(r(x_0)\) are close to each other within the sampling errors then the underlying generating mechanism of the meteorological variable is of independent (white noise) type. In this case only the classical extreme value statistics can be applied. Otherwise, structural dependence exists and the following final step must be executed in order to obtain the maximum value distribution.
(vi) Substitute functions \(q(x_0)\) and \(r(x_0)\) into Eq. (16) so as to obtain the PDF of maxima.

4. Evaluation

It is not possible to solve Eq. (16) analytically, therefore, numerical simulation solutions on digital computers are sought for a given set of correlation coefficient \((\rho : 0.2; 0.4 \text{ and } 0.6)\) for finite lengths \((n : 2; 10; 100)\). In the simulations, the meteorological variable is assumed to originate from a normal distribution function with zero mean and unit variance. The joint PDF of two successive meteorological variables is assumed to accord with a bivariate PDF. It is worth of mentioning at this stage that Cramer and Leadbetter (1967) expressed the first order autorun coefficient as

$$r(x_0) = q(x_0) + \frac{1}{2\pi q(x_0)} \int_0^\rho \exp[-x_0^2/2(1 + z)](1 - z^2)^{1/2}dz$$

(17)

in which \(z\) is a dummy variable. Substitution of Eq. (17) into Eq. (16) leads to an explicit analytical form for calculating numerically the probability of maximum event in a finite sample size for any given autocorrelation coefficient, \(\rho\).

First of all the expected value of maxima in various sample sizes have been obtained with the Monte Carlo simulation and the results are shown in Table 1 together with numerical values obtained by considering Eq. (17).
Table 1. *Maximum Value Simulation*

<table>
<thead>
<tr>
<th>Sample size</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical</td>
<td>0.895</td>
<td>1.212</td>
<td>1.587</td>
<td>1.914</td>
<td>2.087</td>
<td>2.290</td>
<td>2.535</td>
</tr>
<tr>
<td>Simulation</td>
<td>0.876</td>
<td>1.210</td>
<td>1.587</td>
<td>1.912</td>
<td>2.080</td>
<td>2.289</td>
<td>2.533</td>
</tr>
</tbody>
</table>

Table 2. *Regression Coefficients*

<table>
<thead>
<tr>
<th>ρ</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.2</td>
<td>3.10</td>
<td>1.65</td>
</tr>
<tr>
<td>0.4</td>
<td>2.20</td>
<td>1.25</td>
</tr>
<tr>
<td>0.6</td>
<td>1.65</td>
<td>0.65</td>
</tr>
</tbody>
</table>

As stated in the previous section, in order to render Eq. (15) into a more practical form it is necessary to relate the truncation level probability \( q(x_0) \) to \( r(x_0) \). This is achieved as a result of the numerical evaluation of Eq. (17) and the matching of the most suitable regression curve through scatter diagram gives, in general,

\[
r(x_0) = aq^{1-\rho}(x_0)[1 - q(x_0)]^b
\]  

(18)

Depending on the value of \( \rho \) the regression coefficients are given in Table 2.

Finally, substitution of Eq. (18) into Eq. (17) leads simply to

\[
P(X_M < x_0) = q(x_0)\{aq^{1-\rho}(x_0)[1 - q(x_0)]^b\}^{n-1}
\]  

(19)

Again the numerical evaluation of this expression gives the change of \( P(X_M < x_0) \) with \( x_0 \) and the following interpretations can be drawn from this general expression:

(i) as the sample length, \( n \), increase the probability of maximum meteorological event occurrence decreases,

(ii) as the correlation coefficient increases there are significant increases in the probability of maximum event. This is tantamount to saying that occurrence of extreme values in dependent variables is less severe than in the independent case.

5. Conclusions

The major drawbacks in the conventional extreme value statistics theory are that finite sample length and the dependence structure of the available historical series cannot be accounted for. Therefore, the use of this classical theory in the evaluation of extreme values leads to overestimation due to sample length and dependence.

An alternative methodology has been developed herein for calculating extreme value statistics from finite length series with dependent serial structure. The necessary formulation for such a situation is derived and then its validity has been shown with a Monte Carlo simulation study.

References


Author's address: Zekai Şen, Technical University of Istanbul, Department of Meteorology, Maslak 80626, Istanbul, Turkey.