Extreme value statistics in records with long-term persistence

Jan F. Eichner,[1] Jan W. Kantelhardt,[2] Armin Bunde,[1] and Shlomo Havlin[3]

[1]*Institut für Theoretische Physik III, Justus-Liebig-Universität Giessen, 35392 Giessen, Germany*

[2]*Fachbereich Physik und Zentrum für Computational Nanoscience, Martin-Luther-Universität Halle-Wittenberg, 06099 Halle (Saale), Germany*

[3]*Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel*

(Received 7 September 2005; published 25 January 2006)

Many natural records exhibit long-term correlations characterized by a power-law decay of the autocorrelation function, $C(s) \sim s^{-\gamma}$, with time lag s and correlation exponent $0 < \gamma < 1$. We study how the presence of such correlations affects the statistics of the extreme events, i.e., the maximum values of the signal within time segments of the fixed duration R. We find numerically that (i) the integrated distribution function of the maxima converges to a Gumbel distribution for large R similar to uncorrelated signals, (ii) the deviations for finite R depend on the initial distribution of the records and on their correlation properties, (iii) the maxima series exhibit long-term correlations similar to those of the original data, and most notably (iv) the maxima distribution as well as the mean maxima significantly depend on the history, in particular on the previous maximum. The last item implies that conditional mean maxima and conditional maxima distributions (with the value of the previous maximum as condition) should be considered for an improved extreme event prediction. We provide indications that this dependence of the mean maxima on the previous maximum occurs also in observational long-term correlated records.

DOI: 10.1103/PhysRevE.73.016130 PACS number(s): 02.50.−r, 05.45.Tp, 89.75.Da

I. INTRODUCTION

Extreme events are rare occurrences of extraordinary nature, such as floods, very high temperatures, or earthquakes. In studying the extreme value statistics of the corresponding time series one wants to learn about the distribution of the extreme events, i.e., the maximum values of the signal within time segments of fixed duration R, and the statistical properties of their sequences. In hydrological engineering, for example, extreme value statistics are commonly applied to decide what building projects are required to protect riverside areas against typical floods that occur once in 100 years. Many exact and empirical results on extreme value statistics have been obtained in the past years, for reviews see, e.g., [1–6]. Most of these results, however, hold only in the limit $R \to \infty$ and are based on statistically independent values of the time series. Both assumptions are not strictly valid in practice. Since observational data are always finite, predictions for finite time intervals R are required, and—most importantly—correlations cannot be disregarded.

Figure 1 illustrates the definition of the series of maxima $(m_j), j=1,\ldots,N/R$ of original data $(x_i), i=1,\ldots,N$, within segments of size R for $R=365$, i.e., for annual maxima if (x_i) represents daily data. According to traditional extreme value statistics the integrated distribution of the maxima m_j converges to a Gumbel distribution (see Sec. II) for independently and identically distributed (i.i.d.) data (x_i) with Gaussian or exponential distribution density [1–3].

In recent years there is growing evidence that many natural records exhibit long-term persistence [7]. Prominent examples include hydrological data [8,9], meteorological and climatological records [10–14], turbulence data [15,16], as well as physiological records [17–19], and DNA sequences [20,21]. Long-term correlations have also been found in the volatility of economic records [22]. In long-term persistent records $(x_i), i=1,\ldots,N$ with mean $\bar{x}$, and standard deviation σ_x the autocorrelation function decays by a power law,

$$C_x(s) = \frac{1}{\sigma_x^2}\langle (x_i - \bar{x})(x_{i+s} - \bar{x})\rangle$$

$$\equiv \frac{1}{\sigma_x^2(N-s)}\sum_{i=1}^{N-s}(x_i - \bar{x})(x_{i+s} - \bar{x}) \sim s^{-\gamma}, \qquad (1)$$

where γ denotes the correlation exponent, $0 < \gamma < 1$.

Such correlations are named "long term" since the mean correlation time $T = \int_0^\infty C_x(s)ds$ diverges for an infinitely long series (in the limit $N \to \infty$). Power-law long-term correlations according to Eq. (1) correspond to a power spectrum $P(f) \sim f^{-\beta}$ with $\beta = 1 - \gamma$ according to the Wiener-Kinchin theorem. For studies of the effect of long-term persistence on the

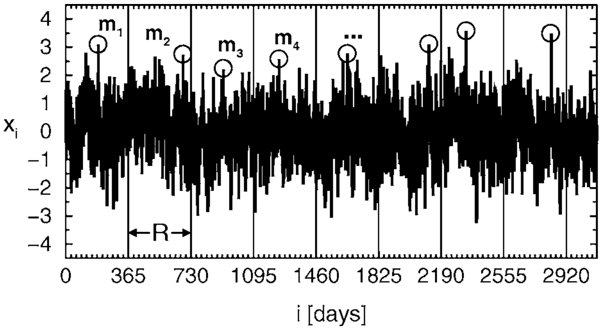

FIG. 1. Definition of maxima: A time series $(x_i), i=1,\ldots,N$, of, e.g., daily data is separated into segments of length $R=365$ days. The maximum values m_j (◯) in each segment, e.g., annual maxima, define another time series $(m_j), j=1,\ldots,N/R$.

statistics of return intervals (in time) between extreme events, see [23–25].

In long-term correlated records, the central assumption in the traditional extreme value statistics [1] is not fulfilled: extreme events cannot be viewed *a priori* as uncorrelated even when there is a long-time span between them. Recently, there have been some approaches to include correlations in the study of extreme value statistics. For the special case of Gaussian $1/f$ correlations in voltage fluctuations in GaAs films extreme value statistics have been demonstrated to follow a Gumbel distribution [26]. A somewhat different asymptotic behavior was observed in experiments on turbulence and in the two-dimensional XY model [27,28], see also [29]. Extreme value statistics have also been employed in studies of hierarchically correlated random variables representing the energies of directed polymers [30] and in studies of maximal heights of growing self-affine surfaces [31]. In the Edwards-Wilkinson model and the Kardar-Parisi-Zhang model for fluctuating, strongly correlated interfaces, an Airy distribution function has been obtained as an exact solution for the distribution of maximal heights very recently [32]. On the other hand, the statistics of extreme height fluctuations for Edwards-Wilkinson relaxation on small-world substrates are rather described by the classical Fisher-Tippet-Gumbel distribution [33]. Besides these recent results there is a theorem by Berman [34] (see also [2,3]) stating that the maxima statistics of stationary Gaussian sequences with correlations converges to a Gumbel distribution asymptotically for $R\to\infty$ provided that $C_x(s)\log_{10}(s)\to 0$ for $s\to\infty$, which holds for long-term correlations.

In this paper we focus on long-term correlated signals and show numerically that (i) the asymptotic convergence of the integrated maxima distribution to the Gumbel formula occurs also for long-term correlated Gaussian or exponentially distributed signals (x_i), (ii) for finite R, the deviation of the integrated maxima distribution from the asymptotics depends significantly on the initial distribution of the data (x_i) and their long-term correlation properties, (iii) the maxima series (m_j) exhibit long-term correlations similar to those of the data (x_i), and, most notably, (iv) the distribution density of the maxima, the integrated maxima distribution, and the mean maxima significantly depend on the history, i.e., the previous maximum m_0. The last item implies that the conditional mean maxima and conditional maxima distributions (with m_0 as condition) should be considered for improved extreme event predictions. We show that the conditional mean maxima for observational data [35,36] have a similar dependence on m_0 as for artificial long-term correlated data.

The paper is organized as follows: In Sec. II we briefly review the main results of traditional extreme value statistics. In Sec. III we study the maxima distribution density and the integrated maxima distribution for uncorrelated as well as long-term correlated data with Gaussian and exponential distribution for several values of R in order to test the convergence to the Gumbel distribution. In Sec. IV we investigate the correlation properties of the sequence of maxima. Sections V and VI report our results for the conditional maxima distributions and for the conditional mean maxima in artificial data as well as in real data. In Sec. VII we consider the centennial quantile, used in a hydrological risk estimation for centennial floodings, and discuss its interference by long-term memory. In Sec. VIII we present a brief summary and conclusions.

II. EXTREME VALUE STATISTICS FOR I.I.D. DATA

In classical extreme value statistics one assumes that records (x_i) consist of i.i.d. data, described by distribution density $P(x)$, which can be, e.g., a Gaussian or an exponential distribution. One is interested in the distribution density function $P_R(m)$ of the maxima (m_j) determined in segments of length R in the original series (x_i) (see Fig. 1). Note that all maxima are also elements of the original data. The corresponding integrated maxima distribution $G_R(m)$ is defined as

$$G_R(m) = 1 - E_R(m) = \int_{-\infty}^{m} P_R(m')dm'. \tag{2}$$

Since $G_R(m)$ is the probability of finding a maximum smaller than m, $E_R(m)$ denotes the probability of finding a maximum that exceeds m. One of the main results of traditional extreme value statistics states that for independently and identically distributed (i.i.d.) data (x_i) with Gaussian or exponential distribution density function $P(x)$ the integrated distribution $G_R(m)$ converges to a double exponential (Fisher-Tippet-Gumbel) distribution (often labeled as type I) [1–3,37], i.e.,

$$G_R(m) \to G\left(\frac{m-u}{\alpha}\right) = \exp[-e^{-(m-u)/\alpha}] \tag{3}$$

for $R\to\infty$, where α is the scale parameter and u the location parameter. By the method of moments those parameters are given by $\alpha=\sigma_R\sqrt{6}/\pi$ and $u=m_R-n_e\alpha$ with the Euler constant $n_e=0.577\,216$ [3,38–40]. Here m_R and σ_R denote the (R-dependent) mean maximum and the standard deviation, respectively. Note that different asymptotics will be reached for broader distributions of data (x_i) that belong to other domains of attraction [3]. For example, for data following a power-law distribution (or Pareto distribution), $P(x)=(x/x_0)^{-k}$, $G_R(m)$ converges to a Fréchet distribution, often labeled as type II. For data following a distribution with finite upper endpoints, for example, the uniform distribution $P(x)=1$ for $0\leqslant x\leqslant 1$, $G_R(m)$ converges to a Weibull distribution, often labeled as type III. These are the other two types of asymptotics, that, however, we do not consider in this paper.

III. EFFECT OF LONG-TERM PERSISTENCE ON THE DISTRIBUTION OF THE MAXIMA

We begin by studying how the convergence of the integrated maxima distribution $G_R(m)$ towards the Gumbel distribution Eq. (3) is affected by long-term correlations in the signal (x_i). Regarding the distribution density $P(x)$ of the signal, we compare results for a Gaussian distribution, $P(x)=1/(\sqrt{2\pi})\exp(-x^2/2)$ $(-\infty<x<\infty)$ and an exponential distribution $P(x)=\exp(-x)$ $(0<x<\infty)$. Artificial long-term corre-

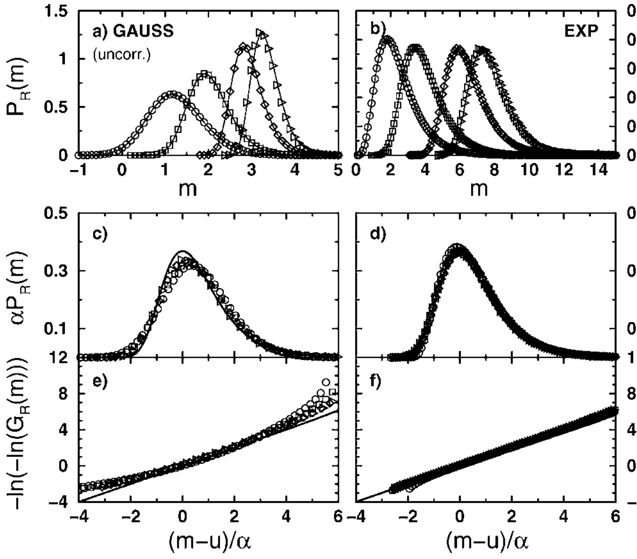

FIG. 2. Distributions of maxima in segments of length R for uncorrelated data with (a), (c), and (e) Gaussian and (b), (d), and (f) exponential distribution density $P(x)$. Pannels (a), and (b) show the distribution density function $P_R(m)$ of the maximum values for four segment sizes $R=6$ (circles), 30 (squares), 365 (diamonds), and 1500 (triangles). Pannels (c), and (d) show that a collapse of all four curves to a single curve is achieved in both cases, when the m axis is replaced by $(m-u)/\alpha$ and $P_R(m)$ is multiplied by the scale parameter α. The solid line is the Gumbel distribution density, Eq. (4). Pannels (e), and (f) show the corresponding integrated distribution $G_R(m)$ together with the Gumbel function Eq. (3). Note that exponential records converge to the Gumbel distribution much faster than Gaussian records.

lated signals following these distributions can be generated by the Fourier filtering method (see, e.g., [41]) and by the Schreiber-Schmitz iteration procedure [42,43], respectively. In the Schreiber-Schmitz procedure we employed 1000 iterations for each record of length $N=2^{21}\approx 2\times 10^6$. We found that our results do not depend on the number of iterations if more than 100 iterations are used. We studied 150 configurations for most plots.

Figures 2 and 3 compare the maxima statistics for uncorrelated and long-term correlated data [$\gamma=0.4$, see Eq. (1)], respectively. The results for Gaussian distributed data are shown on the left and for exponential distributed data on the right. In pannels (a) and (b) the unscaled distribution densities $P_R(m)$ of the maxima within segments of size R are shown for several values of R. Since Eqs. (2) and (3) yield that for $R\to\infty$

$$P_R(m) \to \frac{1}{\alpha}\exp\left[-e^{-(m-u)/\alpha} - \frac{m-u}{\alpha}\right], \tag{4}$$

the distribution densities $P_R(m)$ can be scaled upon each other if $\alpha P_R(m)$ is plotted versus $(m-u)/\alpha$, see Figs. 2(c), 2(d), 3(c), and 3(d). In Figs. 2(e) and 3(e) it is shown that the convergence towards Eq. (4) (continuous line) is rather slow in the case of a Gaussian distribution of the original data. In contrast, for an exponential distribution $P(x)$ the limiting Gumbel distribution is observed for both uncorrelated and long-term correlated data already for quite small segment sizes R. In Fig. 3(d) deviations occur only at very small R values ($R<10$) where scaling breaks down due to the sharp cutoff of the exponential distribution density $P(x)$ at $x=0$. In the long-term correlated case, where the correlation time T diverges, the fast convergence is particularly surprising, since the segment duration R can never exceed T. From a theoretical point of view, we expect a convergence towards the Gumbel limit only for very large R values. The reason for this fast convergence may be a rapid weakening of the correlations among the maximum with increasing values of R, as we will see in the next section (Fig. 5).

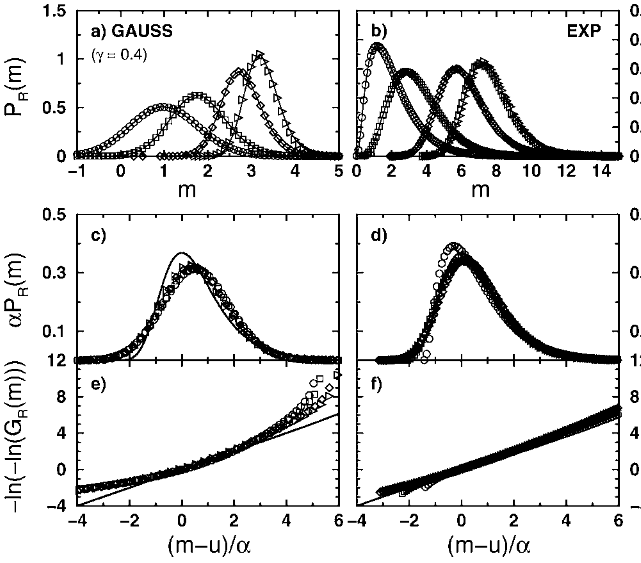

FIG. 3. Distributions of maxima in segments of length R for long-term correlated data with $\gamma=0.4$; for explanations of the plots and the symbols see Fig. 2. The curves in (a), and (b) appear broader than in Fig. 2, because in correlated data more small m values are considered in $P_R(m)$ than in uncorrelated data. In (d) the curve for $R=6$ (circles) differs most from the theoretical curve for uncorrelated data, an effect caused by the correlations together with a rather small R value: the left tail of $P_R(m)$ is strongly affected by the abrupt left end of the exponential. For larger R values this effect disappears and the Gumbel distribution is well approached. For the Gaussian data in (c), and (e), the Gumbel law (solid line) is again not well approached.

We conclude that the distribution $P(x)$ of the original data has a much stronger effect upon the convergence towards the Gumbel distribution than the long-term correlations in the data. Long-term correlations just slightly delay the convergence of $G_R(m)$ towards the Gumbel distribution (3). This can be observed very clearly in the plot of the integrated and scaled distribution $G_R(m)$ on the logarithmic scale in the bottom pannels of Figs. 2 and 3.

Figure 4 shows a direct comparison of the distribution densities $P_R(m)$ for uncorrelated and correlated Gaussian and exponentially distributed data for $R=365$ (corresponding to annual maxima). The distributions for the long-term correlated data exhibit a slight shift to the left and, in particular, a significant broadening of the left tail. The reason for this is that correlations cause some periods with many large values x_i and other periods with only relatively small values x_i.

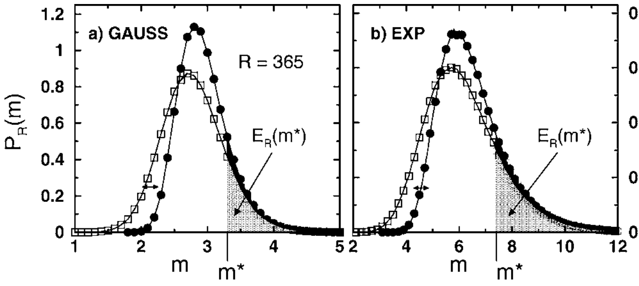

FIG. 4. Comparison of the distribution density $P_R(m)$ of the maxima for uncorrelated (circles) and long-term correlated (γ=0.4, squares) data for fixed R=365. For both (a) Gaussian and (b) exponentially distributed (x_i), the long-term correlations lead to a similar broadening of the left tail, while the right tail is hardly affected by correlations. In (a), the probability $E_R(m^*)$ of finding a maximum m larger than an (arbitrary but sufficiently large) m^*=3.3 is 0.14 for correlated data (gray area) compared with 0.16 for uncorrelated data, and in (b) 0.18 for correlated data compared with 0.2 for uncorrelated data for m^*=7.4. The difference in the probabilities $E_R(m^*)$ for large m^* is marginal.

When picking the annual maxima from the correlated data the periods where small x_i values dominate will yield rather small annual maxima compared with uncorrelated data; this leads to the broadening of the left tail of $P_R(m)$. The largest events are still identified as annual maxima, and hence the right tail of the distribution density is hardly affected by correlations. Figure 4 clearly illustrates that the probability of a maximum exceeding an arbitrary but sufficiently large value m^*, $E_R(m^*)$ [see Eq. (2)], is not significantly different for correlated and uncorrelated data, for both the $P(x)$ Gaussian and exponential.

IV. EFFECT OF LONG-TERM PERSISTENCE ON THE CORRELATIONS OF THE MAXIMA

The distributions of maxima considered in the previous section do not quantify, however, if the maxima values are arranged in a correlated or in an uncorrelated fashion, and if the clustering of maxima may be induced by long-term correlations in the data. To study this question, we have evaluated the correlation properties of the series of maxima (m_j), $j=1,\ldots,N/R$, of long-term correlated data with Gaussian and exponential distribution. Figure 5 shows representative results for the maxima autocorrelation function

$$C_m(s) = \frac{\langle (m_j - m_R)(m_{j+s} - m_R)\rangle}{\langle (m_j - m_R)^2\rangle}, \tag{5}$$

where m_R denotes the average maximum value in the series, and $\langle\ \rangle$ is the average over j similar to Eq. (1). The comparison with the scaling behavior of the autocorrelation function $C_x(s)$ of the original data (x_i) [see Eq. (1)] that follows a power-law decay, $C_x(s) \sim s^{-\gamma}$ with γ=0.4, reveals the presence of long-term correlations with a correlation exponent $\gamma' \approx \gamma$ in the maxima series. Hence, large maxima m are more likely to be followed by large maxima and small maxima are rather followed by small maxima, leading to clusters of large and small maxima. We note that a similar behavior has been observed for the series of return intervals between extreme events [23–25].

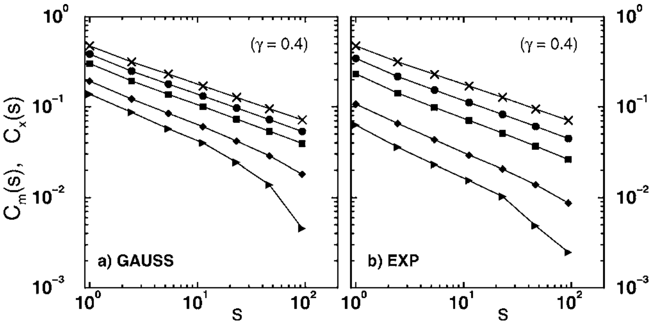

FIG. 5. Autocorrelation function $C_m(s)$ of the maxima (m_j) of (a) Gaussian and (b) exponentially distributed (x_i) for different R values, R=6 (circles), R=30 (squares), R=365 (diamonds), and R=1500 (triangles). The autocorrelation function $C_x(s)$ of the original data (x_i) (crosses) shows the slope $-\gamma$=−0.4.

Figure 5 shows also that the series of maxima for the same R values appear less correlated for exponential data than for the Gaussian data. Due to a wider distribution of the maxima in the exponential case (see Fig. 4) the autocorrelation function $C_m(s)$ is lower for the maxima of exponential data compared to the maxima of Gaussian data.

Figure 6 shows that the deviations of the autocorrelation function $C_m(s)$ from a power-law fit with slope γ=0.4 for large values of R and s are presumably caused by finite-size effects. They become significantly smaller as the length N of the series is increased. In the case of uncorrelated data, the series of maxima is also uncorrelated, $C_m(s)$=0 for $s>0$ (not shown).

V. CONDITIONAL MEAN MAXIMA

As a consequence of the long-term correlations in the series of maxima (m_j), the probability of finding a certain value m_j depends on the history, and, in particular, on the value of

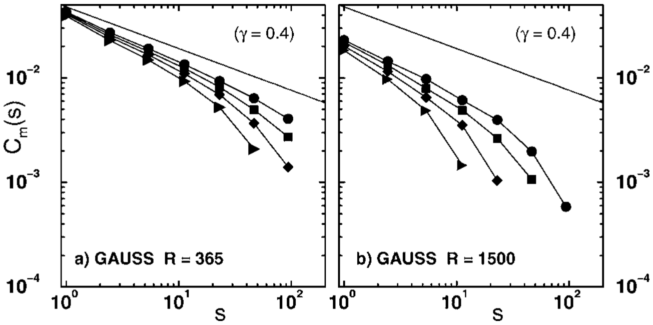

FIG. 6. Study of finite-size effects in the autocorrelation function $C_m(s)$ of sequences of maxima (m_j) for Gaussian distributed (x_i) with (a) R=365 and (b) R=1500. The set lengths are $N=2^{21}$ (circles), 2^{20} (squares), 2^{19} (diamonds), and 2^{18} (triangles). The descent of the slopes of $C_m(s)$ from the slope of the straight line ($-\gamma$=−0.4) with decreasing set length seems to be a finite-size effect.

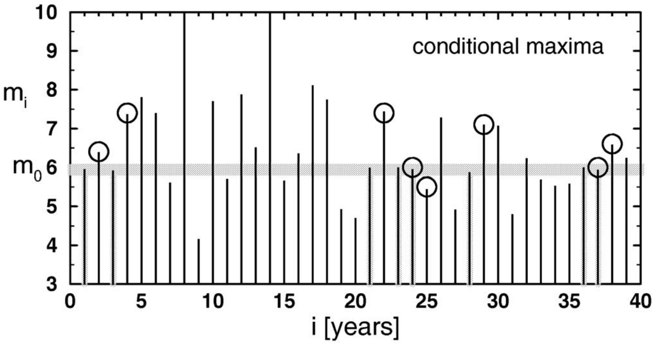

FIG. 7. Definition of the conditional maxima. In the sequence of (annual) maxima only those m values (indicated by circles) are considered, which directly follow a maximum of approximate size $m_0 \approx 6$ (gray band). The new sequence of m values is the sequence of the conditional maxima.

the immediately preceeding maximum m_{j-1}, which we will denote by m_0 in the following. This effect has to be taken into account in predictions and risk estimations. For a quantitative analysis we consider conditional maxima as illustrated in Fig. 7, where all maxima following an $m_0 \approx 6$ (within the gray band), i.e., the subset of maxima which fulfill the conditions of having a preceeding maximum close to m_0, are indicated by circles. The width Δm_0 sketched by the gray band around m_0 in Fig. 7 is set such that a sufficient number of approximately 700 conditional maxima is obtained for each record. The corresponding conditional mean maximum value $m_R(m_0)$ is defined as the average of all these conditional maxima. Note that $m_R(m_0)$ will be independent of m_0 for uncorrelated data.

Figure 8 shows the conditional mean maxima $m_R(m_0)$ versus m_0 for long-term correlated Gaussian and exponentially distributed data for four values of R. Of course, the mean maxima are larger for larger segment sizes R. This dependence is also observed for the unconditional mean maxima indicated by horizontal lines in Fig. 8. In addition to this trivial dependence, the conditional mean maxima significantly depend upon the condition, i.e., the previous maximum m_0, showing a clear memory effect. Evidently, this dependence is most pronounced for the small segment durations $R=6$ and 30. However, it is still observable for the large $R=365$ (most common for observational daily data) and even $R=1500$ (beyond common observational limits). Note that the results for Gaussian and exponentially distributed data agree only qualitatively: while the m_0 dependence of $m_R(m_0)$ is quite close to a linear dependence for the Gaussian data, there seems to be significant curvature for the exponentially distributed data, which is a remnant of the asymmetry of the exponential distribution.

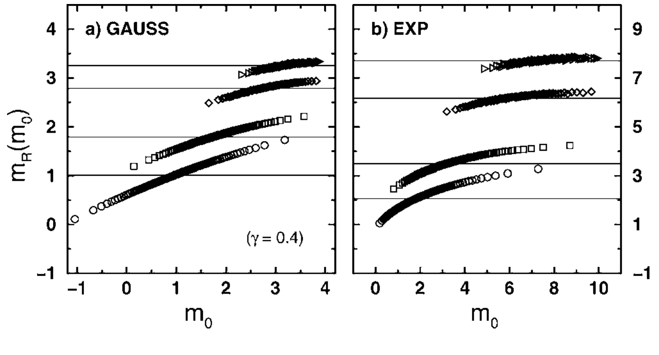

FIG. 8. Mean conditional maxima $m_R(m_0)$ for $\gamma=0.4$ and $R=6$ (circles), $R=30$ (boxes), $R=365$ (diamonds), and $R=1500$ (triangles) versus m_0 for (a) Gaussian and (b) exponential data. The straight lines represent the unconditional means m_R for a given R. The width Δm_0 for the condition m_0 was chosen such that approximately 700 m values were obtained for each m_0 in each of the 150 runs of $N=2^{21}$ data points. Both figures show the memory effect in the form of $m_R(m_0) > m_R$ for rather large m_0 (above m_R) and $m_R(m_0) < m_R$ for rather small m_0 (below m_R).

Next we test our predictions on real records which are known to exhibit long-term correlations. We have studied two data sets, (i) the annual data of the Nile river water level minima [8,35] and (ii) the reconstructed northern hemisphere annual temperatures by Moberg [36]. The Nile series is composed of 663 minimal water levels of the Nile river for the years 622 to 1284 (we use the last 660 data points), to the best of our knowledge, measured at Roda gauge near Cairo. Since the Nile data consist of annual minima, we study extreme minima instead of maxima. The northern hemisphere temperature reconstruction in degree Celsius after Moberg covers the period from 1 AD to 1979 AD (we use the last 1968 data points) and was last updated in February 2005. The correlation properties of both records have been shown elsewhere [8,24,44] to be characterized by $C_x(s) \sim s^{-\gamma}$ with $\gamma \approx 0.3$ [see Eq. (1)].

In order to get sufficient statistics for the conditional means $m_R(m_0)$, we have considered six m_0 intervals for each value of R and have set the width Δm_0 of the band around m_0 such that there are no gaps between the bands. Figure 9 shows the results for three values of R, $R=1$, 6, and 12 years. In all cases, the effect of the long-term correlations (persistence) on the conditional mean minima and maxima $m_R(m_0)$ is clearly visible for both records: the conditional means are smaller for smaller condition value m_0 and larger for larger condition value.

To prove that the dependence upon m_0 is indeed due to the long-term correlations in the records, we have also studied randomly shuffled surrogate data, where all correlations are removed. As shown by the open symbols in Fig. 9 the m_0 dependence completely disappears, indicating that the dependence was due to the correlations in the data.

VI. CONDITIONAL MAXIMA DISTRIBUTIONS

The quantity $m_R(m_0)$ is the first moment of the conditional distribution density $P_R(m|m_0)$, which is defined as the distribution density of all maxima m_j that follow a given maximum value m_0 ($m_{j-1} \approx m_0$, see Fig. 7). Figure 10 shows $P_R(m|m_0)$ for two values of m_0 and again for Gaussian as well as for exponentially distributed long-term correlated data sets with $\gamma=0.4$ and $R=365$. When compared with the unconditional distribution density $P_R(m)$, the long-term correlations lead to a shift of $P_R(m|m_0)$ to smaller m values for small m_0 and to larger m values for large m_0, respectively. The conditional exceedance probability

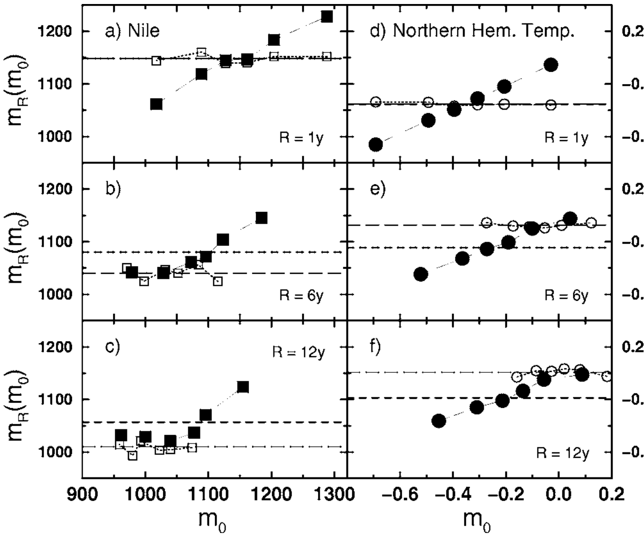

FIG. 9. (a)–(c) Mean conditional minima $m_R(m_0)$ for the annual data of the Nile river water level minima (squares) [35] and (d)–(f) mean conditional maxima for the reconstructed northern hemisphere annual temperatures after Moberg (circles) [36], for (a) and (d) $R=1$ year (b) and (e) $R=6$ years, and (c) and (f) $R=12$ years. The filled symbols show the results for the real data and the open symbols correspond to surrogate data where all correlations have been destroyed by random shuffling. The shuffled data have a different m_0 range due to the broadening of the left tail of $P_R(m)$ in the correlated data (see Fig. 4). The unconditional mean minima (a)–(c) and maxima (d)–(f) are indicated by dashed lines; the long-dashed lines correspond to the shuffled data.

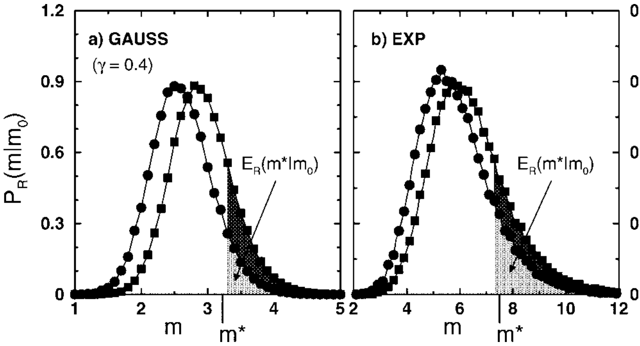

FIG. 10. (a) Conditional distribution density $P_R(m|m_0)$ of maximum values taken from correlated Gaussian data ($\gamma=0.4$) with $R=365$ and $m_0=2.06$ (circles) as well as $m_0=3.55$ (squares). (b) shows the same as (a) for exponentially distributed data with $m_0=4.10$ (circles) and $m_0=8.65$ (squares). The width Δm_0 around m_0 is set such that a sufficient number of approximately 700 conditional maxima is obtained for each of the 150 data sets considered here. The probability $E_R(m^*|m_0)$ to find a m value larger than an arbitrarily given m^* [see Eq. (2)] also depends on the history m_0. For example, in (a), $E_{365}(3.30|2.06)=0.08$ (gray area) is significantly smaller than $E_{365}(3.30|3.55)=0.20$ (black area plus gray area), and in (b) $E_{365}(7.35|4.10)=0.14<E_{365}(7.35|8.65)=0.24$.

$$E_R(m|m_0) = \int_m^\infty P_R(m'|m_0)dm' \qquad (6)$$

defines the probability of finding a maximum larger than m provided that the previous value was close to m_0. We find a strong dependence of $E_R(m|m_0)$ upon the condition m_0. Consequently, the difference between the unconditional probabilities $E_R(m)$ (see Fig. 4) and the corresponding conditional probabilities $E_R(m|m_0)$ depends strongly on m_0 in the presence of long-term correlations.

Next we quantify the effect of long-term correlations upon $E_R(m|m_0)$ for different conditions m_0 and different m values. Figure 11 shows the exceedance probability $E_R(m|m_0)$ for six m values versus m_0. The m values were chosen such that the corresponding unconditional probabilities are $E_R(m)=0.9$, 0.5, 0.3, 0.1, 0.05, and 0.01, respectively. For the Gaussian data and m corresponding to $E_R(m)=0.5$ the curve $E_R(m|m_0)$ varies by a factor of 2 depending on m_0, while the variation does not exceed a factor of 1.5 for the exponential data.

In general, the memory effect caused by the long-term correlations seems to be the strongest for intermediate m values. For $E_R(m)\leqslant 0.5$, the larger the m value (i. e., the lower the curve in Fig. 11) the smaller is the apparent effect of the correlations on the difference between the conditional probabilities $E_R(m|m_0)$ (symbols) and the unconditional probabilities $E_R(m)$ (straight lines). Hence, Fig. 11 may suggest that the memory effect will disappear for very large m values. This, however, is not true. Figure 12 shows the ratios of the conditional exceedance probabilities $E_R(m|m_0)$ and the unconditional exceedance probabilities $E_R(m)$. The figure clearly shows an increase of the memory effect for larger m

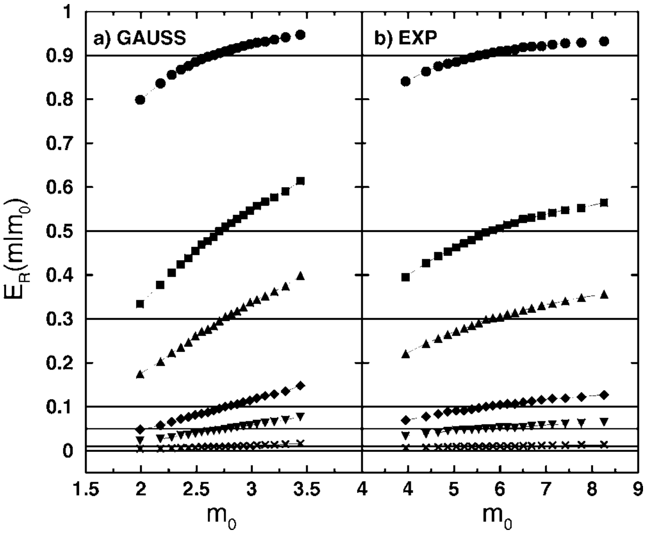

FIG. 11. Conditional exceedance probabilities $E_R(m|m_0)$ to find a maximum larger than a given m for (a) Gaussian and (b) exponentially distributed data with $\gamma=0.4$ and $R=365$ versus the condition m_0. The straight lines indicate the corresponding unconditional exceedance probabilities $E_R(m)$. The six m values were chosen such that $E_R(m)=0.9$ (circles, $m=2.15$ for Gaussian and 4.40 for exponential data), 0.5 (squares, $m=2.75$ and 5.95), 0.3 (triangles up, $m=2.95$ and 6.70), 0.1 (diamonds, $m=3.35$ and 8.00), 0.05 (triangles down, $m=3.55$ and 8.75), and 0.01 (crosses, $m=3.95$ and 10.40). Each point in the graph is based on a statistics of 500 conditional m values and averaged over 150 runs of $N=2^{21}$ data points.

values, i.e., for more extreme events. This increase seems weaker for exponentially distributed data than for Gaussian distributed data due to the less correlated maxium series of exponential data; however the tendency is the same. As Fig. 12 shows $E_R(m|m_0)$ can differ up to a factor of 2 from $E_R(m)$ when considering the history m_0 in the presence of long-term correlations (with $\gamma=0.4$). This effect has to be taken into account in predictions and risk estimations of large events.

VII. ESTIMATION OF CENTENNIAL EVENTS IN LONG-TERM CORRELATED RECORDS

In this section we discuss how the size of typical centennial events, i.e., typical maxima that occur once in 100 years, can be determined in practice. Such values are commonly used, e.g., in hydrological risk estimation for centennial floodings. If all distributions $P(x)$ and $P_R(m)$ of the considered time series were known exactly, i.e., for infinitely long records, two alternative definitions of typical centennial events would be possible as illustrated in Fig. 13(a) and 13(b).

The *first* definition of a typical centennial event [see Fig. 13(a)] considers the distribution density $P(x)$ (here: a Gaussian distribution of x_i, representing daily data) and determines the quantile $q_{36\,500}$ (dashed line) that is exceeded by only 1/36500 of all values of the distribution, i.e., on average $x_i > q_{36\,500}$ occurs once in 100 years=36 500 days. For this definition the distribution $P(x)$ must be known for very rare events. Moreover, since $q_{36\,500}$ is based only on the distribution of the values x_i, it is unaffected by possible correlations and clustering of centennial events (see [23,24]). This definition takes into account all events that exceed the quantile, regardless of whether they occur within the same 100 years period or not.

The *second* definition of a typical centennial event [see Fig. 13(b)] considers the distribution density $P_{36\,500}(m)$ of the centennial maxima m within periods of 100 years (histogram in the figure) and determines the mean centennial maximum value $m_{36\,500}$ as the first moment of this distribution (dashed line). Clearly, this definition includes the effects of correlations, but multiple exceedances of the threshold within one period of 100 years are not regarded in $m_{36\,500}$ by definition. Still, many centennial events must have occurred in the record (x_i) to allow the determination of $P_{36\,500}(m)$.

In real data, however, the number of values obtained from records with typical durations of 30, 50, or 100 years is not sufficient to study the distribution densities $P(x)$ or $P_R(m)$ in order to determine directly the size of centennial events by a calculation of $q_{36\,500}$ or $m_{36\,500}$. It is usually not known *a priori* if even one centennial event occurred within the observational period. This makes it very difficult to estimate the size of a typical centennial event. Therefore, a practical definition is needed. This *third* definition assumes that $P(x)$ is (most likely) in the domain of attraction of the Gumbel distribution. Then one applies the Gumbel fit formula Eq. (4) to the distribution density $P_R(m)$ of maxima m within (smaller) segments of size $R=365$ days and approximates the typical centennial event from this Gumbel fit. This procedure is illustrated in Fig. 13(c), where the histogram $P_{365}(m)$ of annual maxima and a fitted Gumbel curve (solid line) are shown. It is common practice in hydrology to estimate the size of a typical centennial event by calculating the threshold Q_{100} [dashed line in Fig. 13(c)], which is exceeded by only 1/100 of the fitted Gumbel distribution of annual maxima [38,40]. In terms of the integrated Gumbel distribution $G_{365}(m)$ [see Eqs. (2) and (3)] this definition corresponds to $G_{365}(Q_{100})=0.99$, which yields $Q_{100}=u-\alpha \ln(-\ln 0.99)$, i.e.,

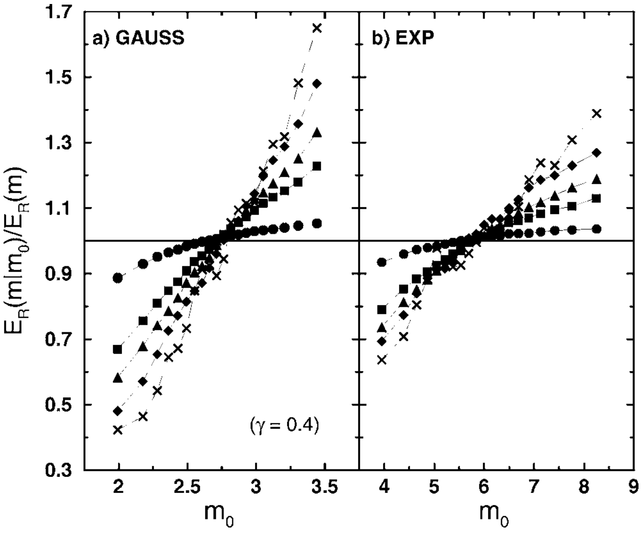

FIG. 12. Ratios of the conditional and unconditional exceedance probabilities $E_R(m|m_0)$ and $E_R(m)$ to find a maximum larger than m for (a) Gaussian and (b) exponentially distributed data with $\gamma=0.4$ and $R=365$. The symbols and the statistics are the same as in Fig. 11 except for the data corresponding to $E_R(m)=0.05$ (triangles down), which are not shown here to avoid overlapping symbols. The effect of long-term correlations seems to be strongest for the largest m (crosses): depending on m_0, $E_R(m|m_0)$ varies from 0.4 up to 1.7 for Gaussian data, i.e., by a factor greater than 4. For exponential data this factor is still greater than 2.

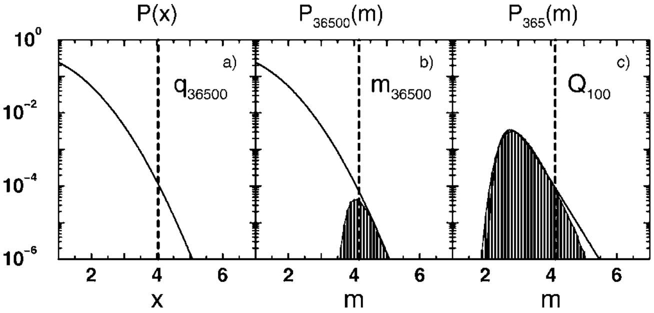

FIG. 13. Definition of centennial events: (a) the tail of a Gaussian distribution $P(x)$ (solid black line) with the 1/36 500 quantile (dashed line), (b) the Gaussian tail (solid black line) and the histogram $P_{36\,500}(m)$ of centennial maxima with the average $m_{36\,500}$ (dashed line), and (c) the histogram $P_{365}(m)$ of the annual maxima and the corresponding Gumbel fit (Eq. (4), solid line) with its 1/100 quantile Q_{100} [dashed line, Eq. (7)]. The deviations of the fit from the histogram are caused by the underlying Gaussian distribution, for which $R=365$ is not sufficiently large to reach the Gumbel limit.

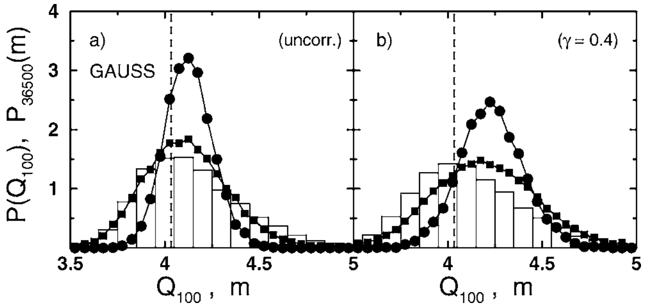

FIG. 14. Comparison of the quantile $q_{36\,500}$ (dashed vertical lines), the distribution density $P_{36\,500}(m)$ of centennial maxima (histograms), and the distributions of the Q_{100} values based on 30 years (squares) and 100 years (circles) for Gaussian distributed (a) uncorrelated and (b) long-term correlated (γ=0.4) data; see Fig. 13 and Eq. (7) for the definitions. Each Q_{100} value was calculated from a different segment of length 36 500 days (=100 years, circles) and 10 950 days (=30 years, squares). The histograms were obtained from 150 artificial records with $N=2^{21}$.

$$Q_{100} = m_{365} - [\ln(-\ln 0.99) + n_e]\frac{\sqrt{6}}{\pi}\sigma_{365} \tag{7}$$

or $Q_{100} \approx m_{365} + 3.14\sigma_{365}$ [38,40]. Here, m_{365} and σ_{365} denote the annual average and its standard deviation, respectively, which are easily accessible also in short records. The three definitions for centennial maxima, $q_{36\,500}$, $m_{36\,500}$, and Q_{100} (most regarded in hydrology), are similar, but differ slightly depending on the underlying correlation structure of the data, as we will show now.

Figure 14 compares the quantile $q_{36\,500}$, the distribution density $P_{36\,500}(m)$ of centennial maxima, and the distribution of Q_{100} values for correlated and uncorrelated Gaussian distributed data. To obtain a similar statistical basis for both, $P_{36\,500}(m)$ and the distribution of Q_{100}, segments of length 36 500 days (100 years) should be considered for each of the Q_{100} values. This was done for the data shown by filled circles in Fig. 14. However, since real observational records are often shorter, Fig. 14 also shows the distribution of the Q_{100} values based on segments of 10 950 days (30 years, squares). One can see that $P_{36\,500}(m)$ and also the corresponding (actual) mean centennial maximum $m_{36\,500}$ are less affected by the considered long-term correlations than the estimated centennial events Q_{100} based on Gumbel fits, while the quantile $q_{36\,500}$ is independent of correlations. Note also that the shift of the Q_{100} histogram due to the correlations is larger in contrast to the small shift of the $P_{36\,500}(m)$ histogram. For $P_{36\,500}(m)$ we observe mainly a broadening caused by the correlations (see Fig. 4). The shift of the Q_{100} histogram to the right is probably caused by the influence of the scale parameter in the fit formula, i.e., by the standard deviation σ_{365} that appears in Eq. (7). In addition, for the quantity Q_{100} a broadening of the histogram is observed for the correlated data, leading to a less accurate estimation of the typical centennial event. However, the distribution of the Q_{100} values still remains significantly narrower than $P_{36\,500}(m)$, which indicates that—for single records comprising just about 100 years—typical centennial events can be approximated somewhat more reliably using Q_{100} instead of the single maximum m picked from the record. The estimations of centennial events via Q_{100} based on just 30 years [Fig. 14(c) and 14(d)] are similarly reliable as those via the maximum picked from a 100 years series.

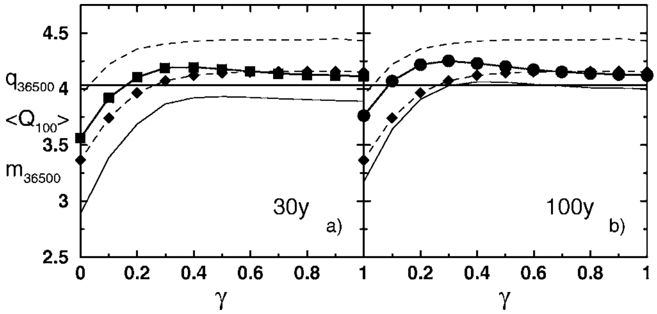

FIG. 15. Dependence of actual centennial events $m_{36\,500}$ and estimated centennial events Q_{100} on the correlation exponent γ of the data; small values of γ indicating strong long-term correlations. The average (actual) centennial maxima (diamonds) and the corresponding upper fluctuations (range of one standard deviation indicated by dashed lines) is compared with the average Q_{100} value (squares and circles) and its lower fluctuations (solid lines). The Q_{100} values were obtained from segments of size (a) 10 950 days (30 years, squares) and (b) 36 500 days (100 years, circles). The solid horizontal line in both figures is the 1/36 500 quantile $q_{36\,500}$=4.034.

In order to compare the dependence of the actual and estimated sizes of centennial events and the accuracy of their estimation in long-term correlated records quantitatively, we have studied histograms like those in Fig. 14 for artificial data characterized by different correlation exponents γ. Figure 15 shows, as a function of γ, the constant value $q_{36\,500}$ [as defined in Fig. 13(a)], the (actual) mean centennial maximum $m_{36\,500}$ (diamonds) with the corresponding standard deviation $\sigma_{36\,500}$ and the mean estimated centennial maximum $\langle Q_{100}\rangle$ with its standard deviation. Again, each single value Q_{100} is calculated for a segment of length 10 950 days (30 years) in Fig. 15(a) [squares] and 36 500 days (100 years) in Fig. 15(b) [circles]. The quantity $q_{36\,500}$ is (trivially) independent of the long-term correlations, because it is based only on the distribution density $P(x)$, and can hardly be calculated in practice, because $P(x)$ required for the calculation is usually not known. In contrast to $q_{36\,500}$, $m_{36\,500}$, and even more $\langle Q_{100}\rangle$ are significantly affected by strong long-term correlations ($\gamma<0.3$). Both values decrease with increasing correlations (decreasing γ). Due to the strong long-term correlations large maxima tend to cluster, i.e., there are epochs where considerably more large maximum values occur than in weakly correlated and uncorrelated data [24]. As a consequence, there exist also epochs, where the maximum values are considerably lower than those in weakly or uncorrelated records. With increasing correlations ($\gamma<0.3$) these periods of small maxima become more pronounced and more frequent, forcing the average centennial maximum $m_{36\,500}$ and also $\langle Q_{100}\rangle$ (which is based on annual maxima) to drop below the quantile $q_{36\,500}$. The corresponding standard deviations (dashed lines and solid lines), which characterize the widths of the histograms of $P_{36\,500}(m)$ and $P(Q_{100})$ and thus carry the information regarding the accuracy of the estimations for short data, increase with decreasing γ. For $\gamma > 0.4$ the mean centennial maximum $m_{36\,500}$ and $\langle Q_{100}\rangle$ are roughly constant.

The mean estimated centennial maximum $\langle Q_{100}\rangle$ tends to overestimate the size of the centennial events for $\gamma \leqslant 0.6$. However, the systematic deviation from $m_{36\,500}$ is even smaller for $\langle Q_{100}\rangle$ based on just 30 years of data [Fig. 15(a)] than for 100 years of data [Fig. 15(b)]. In addition, for weakly correlated data, the corresponding standard deviations are lower than $\sigma_{36\,500}$, indicating more reliable estimations of centennial maxima with Q_{100} for short records. In conclusion, we find that the quantity Q_{100} most regarded in hydrology is a very reliable predictor of typical centennial events in short records, but the values tend to be systematically larger than the $m_{36\,500}$ values for data with strong long-term correlations. However, in real hydrology data such as river runoff data the correlations are hardly stronger than $\gamma=0.3$ [45,46]. So the Q_{100} value is still a good estimator for centennial floods.

Finally, we want to study the effects of the long-term memory on Q_{100}. While there is hardly any memory to be expected in $m_{36\,500}$ because of the very large R value (see Figs. 5 and 8), Q_{100} should still show some dependence on the history, because it is based on m_{365} and σ_{365}. Figure 16 shows, for long-term correlated Gaussian distributed data, the conditional average $\langle Q_{100}(Q_{100}^{(0)})\rangle$ and the conditional $m_{36\,500}(m_0)$, i.e., the average Q_{100} value following a Q_{100} of size $Q_{100}^{(0)}$ and analogous for $m_{36\,500}(m_0)$. While the conditional average Q_{100} (filled circles) depends on the history, the conditional $m_{36\,500}$ (squares) fluctuate around the unconditional mean (solid line). Although the conditional effect on Q_{100} is relatively small, it remains measurable and can help to improve predictions of extreme events within given periods of time.

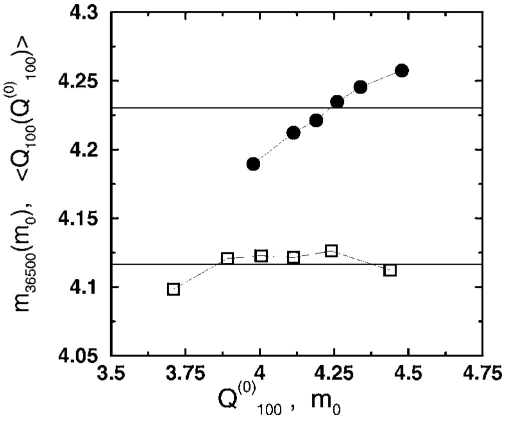

FIG. 16. Conditional average Q_{100} (filled circles) and conditional $m_{36\,500}$ (squares) for Gaussian long-term correlated ($\gamma=0.4$) data. The $\langle Q_{100}(Q_{100}^{(0)})\rangle$ values show a correlation based memory effect indicated by the upward trend, while the $m_{36\,500}$ values seem to have no memory due to the large R value. The horizontal lines correspond to the unconditional average values 4.23 and 4.12 for $\langle Q_{100}\rangle$ and $m_{36\,500}$, respectively. The Q_{100} values were calculated from segments of a length of 36 500 days.

VIII. SUMMARY AND CONCLUSIONS

In summary, we have studied the effect of long-term correlations in a time series upon extreme value statistics. Considering series of maxima within segments of size R of the original data, we have shown numerically that the maxima distribution functions still converge to the same type of Gumbel distributions as for uncorrelated data for increasing R.

For finite values of R, however, some deviations occur especially for originally Gaussian distributed data. Our extensive numerical simulations suggest that contrary to the common assumption in extreme value statistics, the maxima time series turn out to be *not* independently, identically distributed numbers. The series of maxima rather exhibit long-term correlations similar to those in the original data. Most notably we find that the maxima distribution as well as the mean maxima significantly depend on the history, in particular on the value of the previous maximum. In addition, we have shown that long-term memory can lead to a slight systematic overestimation of centennial events if the approximation Q_{100} is considered. In general, Q_{100} is a surprisingly reliable approximation for centennial events especially in short records.

Nevertheless, further work is needed to test if our findings are similar in other (non-Gaussian) initial distributions. Our preliminary results indicate that the effects on Q_{100} might be reversed for exponentially distributed data. In addition, we suggest that memory via the conditional mean maxima and conditional maxima distributions as well as conditional Q_{100} values should be considered for an improved risk estimation in long-term correlated data. It is also plausible that multiscaling, which occurs, e.g., in many hydrological time series, might have an even more significant impact on risk estimation and the prediction of extreme events like floods. Further work is definitely required to study the effects of multiscaling in the time series upon extreme value statistics.

ACKNOWLEDGMENTS

This work has been supported by the Bundesministerium für Bildung und Forschung and the Israel Science Foundation.

[1] E. J. Gumbel, *Statistics of Extremes* (Columbia University Press, New York, 1958).

[2] J. Galambos *The Asymptotic Theory of Extreme Order Statistics* (John Wiley and Sons, New York, 1978).

[3] M. R. Leadbetter, G. Lindgren, and H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes* (Springer, New York, 1983).

[4] *Extreme Value Theory and Applications*, edited by J. Galambos, J. Lechner, and E. Simin (Kluwer, Dordrecht, 1994).

[5] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events*, edited by I. Karatzas and M. Yor (Springer, Berlin, 1997).

[6] H. v. Storch and F. W. Zwiers, *Statistical Analysis in Climate Research* (Cambridge University Press, Cambridge, 2001).

[7] *The Science of Disasters-Climate Disruptions, Heart Attacks, and Market Crashes*, edited by A. Bunde, J. Kropp, and H.-J. Schellnhuber (Springer, Berlin, 2002).

[8] H. E. Hurst, R. P. Black, and Y. M. Simaika, *Long-term Storage: An Experimental Study* (Constable & Co. Ltd., London, 1965).

[9] B. B. Mandelbrot and J. R. Wallis, Water Resour. Res. **5**, 321 (1969).

[10] E. Koscielny-Bunde, A. Bunde, S. Havlin, and Y. Goldreich, Physica A **231**, 393 (1996).

[11] J. D. Pelletier and D. L. Turcotte, J. Hydrol. **203**, 198 (1997).

[12] E. Koscielny-Bunde, A. Bunde, S. Havlin, H. E. Roman, Y. Goldreich, and H.-J. Schellnhuber, Phys. Rev. Lett. **81**, 729 (1998).

[13] P. Talkner and R. O. Weber, Phys. Rev. E **62**, 150 (2000).

[14] J. F. Eichner, E. Koscielny-Bunde, A. Bunde, S. Havlin, and H.-J. Schellnhuber, Phys. Rev. E **68**, 046133 (2003).

[15] M. F. Shlesinger, B. J. West, and J. Klafter, Phys. Rev. Lett. **58**, 1100 (1987).

[16] R. R. Prasad, C. Meneveau, and K. R. Sreenivasan, Phys. Rev. Lett. **61**, 74 (1988).

[17] C.-K. Peng, J. Mietus, J. M. Hausdorff, S. Havlin, H. E. Stanley, and A. L. Goldberger, Phys. Rev. Lett. **70**, 1343 (1993).

[18] A. Bunde, S. Havlin, J. W. Kantelhardt, T. Penzel, J.-H. Peter, and K. Voigt, Phys. Rev. Lett. **85**, 3736 (2000).

[19] J. W. Kantelhardt, T. Penzel, S. Rostig, H. F. Becker, S. Havlin, and A. Bunde, Physica A **319**, 447 (2003).

[20] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[21] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).

[22] Y. H. Liu, P. Cizeau, M. Meyer, C.-K. Peng, and H. E. Stanley, Physica A **245**, 437 (1997).

[23] A. Bunde, J. F. Eichner, S. Havlin, and J. W. Kantelhardt, Physica A **330**, 1 (2003).

[24] A. Bunde, J. F. Eichner, J. W. Kantelhardt, and S. Havlin, Phys. Rev. Lett. **94**, 048701 (2005).

[25] E. G. Altmann and H. Kantz, Phys. Rev. E **71**, 056106 (2005).

[26] T. Antal, M. Droz, G. Györgyi, and Z. Racz, Phys. Rev. Lett. **87**, 240601 (2001).

[27] S. T. Bramwell, P. C. W. Holdsworth, and J.-F. Pinton, Nature (London) **396**, 552 (1998).

[28] S. T. Bramwell, K. Christensen, J.-Y. Fortin, P. C. W. Holdsworth, H. J. Jensen, S. Lise, J. M. Lopez, M. Nicodemi, J.-F. Pinton, and M. Sellitto, Phys. Rev. Lett. **84**, 3744 (2000).

[29] K. Dahlstedt and H. J. Jensen, J. Phys. A **34**, 11193 (2001).

[30] D. S. Dean and S. N. Majumdar, Phys. Rev. E **64**, 046121 (2001).

[31] S. Raychaudhuri, M. Cranston, C. Przybyla, and Y. Shapir, Phys. Rev. Lett. **87**, 136101 (2001).

[32] S. N. Majumdar and A. Comtet, Phys. Rev. Lett. **92**, 225501 (2004); J. Stat. Phys. **119**, 777 (2005).

[33] H. Guclu and G. Korniss, Phys. Rev. E **69**, 065104(R) (2004).

[34] S. M. Berman, Ann. Math. Stat. **35**, 502 (1964).

[35] Data obtained from: http://sunsite.univie.ac.at/statlib/S/beran. See also: B. Whitcher, S. D. Byers, P. Guttorp, and D. B. Percival, Water Resour. Res. **38**, 1054 (2002).

[36] Data obtained from: http://www.ndcd.noaa.gov/paleo/recons.html. See also: A. Moberg, D. M. Sonechkin, K. Holmgren, N. M. Datsenko, and W. Karln, Nature (London) **433**, 613 (2005).

[37] R. A. Fisher and L. H. C. Tippett, Proc. Cambridge Philos. Soc. **24**, 180 (1928).

[38] V. te Chow, *Handbook of Applied Hydrology* (McGraw-Hill Book Company, New York, 1964).

[39] A. J. Raudkivi, *Hydrology* (Pergamon Press, Oxford, 1979).

[40] P. F. Rasmussen and N. Gautam, J. Hydrol. **280**, 265 (2003).

[41] H. A. Makse, S. Havlin, M. Schwartz, and H. E. Stanley, Phys. Rev. E **53**, 5445 (1996).

[42] T. Schreiber and A. Schmitz, Physica D **142**, 346 (2000).

[43] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).

[44] D. Rybski, (private communication).

[45] E. Koscielny-Bunde, J. W. Kantelhardt, P. Braun, A. Bunde, and S. Havlin, e-print physics/0305078, J. Hydrol. (to be published).

[46] J. W. Kantelhardt, E. Koscielny-Bunde, D. Rybski, P. Braun, A. Bunde, and S. Havlin, J. Geophys. Res. **111**, 1106 (2006).