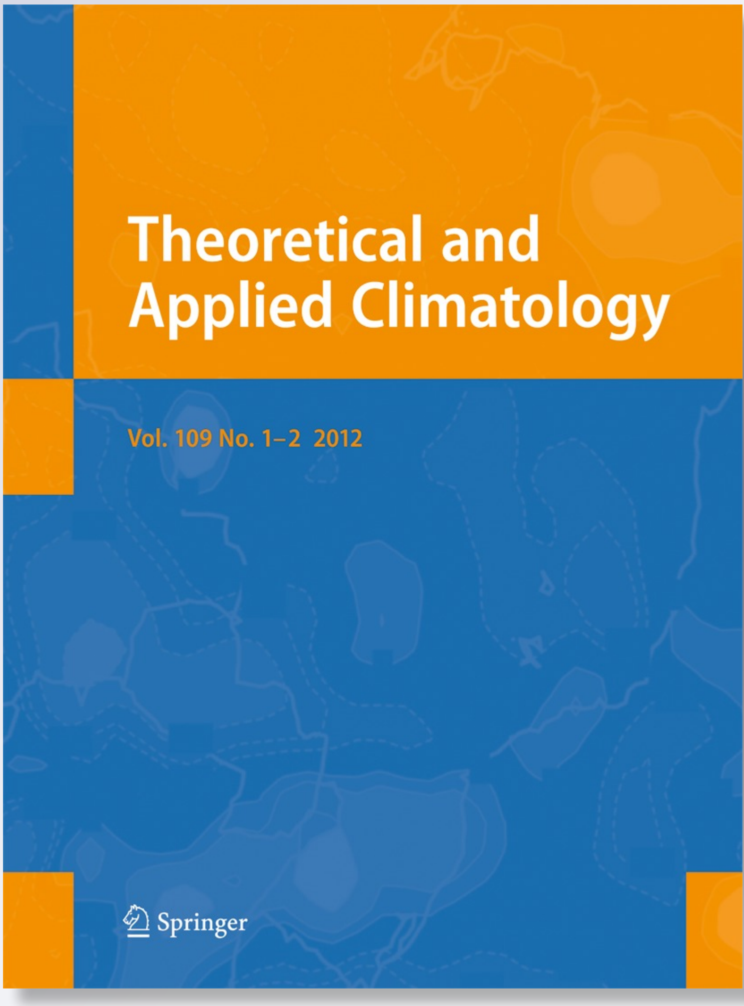
Effect of extreme value loss on long-term correlated time series

Naiming Yuan, Zuntao Fu, Huiqun Li & Jiangyu Mao

Theoretical and Applied Climatology

ISSN 0177-798X
Volume 109
Combined 1-2

Theor Appl Climatol (2012) 109:133-140
DOI 10.1007/s00704-011-0566-x



Theor Appl Climatol (2012) 109:133–140
DOI 10.1007/s00704-011-0566-x

ORIGINAL PAPER

Effect of extreme value loss on long-term correlated time series

Naiming Yuan · Zuntao Fu · Huiqun Li · Jiangyu Mao

Received: 10 May 2011 / Accepted: 23 November 2011 / Published online: 23 December 2011

Abstract Effects of extreme value loss on long-term correlated time series are analyzed by means of detrended fluctuation analysis (DFA) and power spectral density analysis. Weaker memory can be detected after removing of extreme values for the artificial long-term correlated data, indicating the emergence of extreme events may be closely related to long-term memory. For observational temperature records, similar results are obtained, but not in all stations. For example, in some stations, only extending of scaling range to smaller time scales occurs, which may be due to the asymmetric distribution of values in the record. By comparing our findings with previous works, clustered positions of the extreme events are recognized as an important property in long-term correlated records. Through a simple numerical test, close relations between extreme events and long-term memory are discovered, which is helpful for our understanding of the effects of extreme value loss on long-term correlated records.

N. Yuan · Z. Fu · H. Li
Laboratory for Climate and Ocean–Atmosphere Studies,
Peking University, Beijing, 100871, China

N. Yuan · Z. FU (✉) · H. Li
Department of Atmospheric and Oceanic Sciences,
School of Physics, Peking University, Beijing,
100871, China
e-mail: fuzt@pku.edu.cn

J. Mao
LASG, Institute of Atmospheric Physics, CAS, Beijing,
100029, China

1 Introduction

Recently, more and more studies show that many natural records exhibit long-term correlations, such as DNA sequences (Peng et al. 1992, 1994; Arneodo et al. 1995; Buldyrev et al. 1995), economic time series (Mantegna and Stanley 1995; Liu et al. 1997), heartbeat records (Peng et al. 1995; Bunde et al. 2000), as well as meteorological and climatological records (Koscielny-Bunde et al. 1998; Weber and Talkner 2001; kantelhardt et al. 2006; Chen et al. 2007; Rybski et al. 2008). Especially in the last decade, since a method called detrended fluctuation analysis (DFA) was developed (Peng et al. 1994), more studies on long-term correlation have been published Buldyrev et al. (1995); Koscielny-Bunde et al. (1998); Eichner et al. (2003). Using this method, we can characterize the fractal properties more reliably without being affected by nonstationarity and trends. Another interesting topic is the so-called extreme value statistics, which has also been studied extensively in recent years (Altmann and Kantz 2005; Livina et al. 2005; Eichner et al. 2006). Extreme events, such as very high temperatures, floods, droughts, or earthquakes, may cause disasters which can seriously affect our daily life. Thus, a question may come out: Are there any relations between extreme events and long-term correlations? Some work have been done to address this question, like the return intervals of rare events in records may have similar long-term correlations to that of the original records (Bunde et al. 2004, 2005; Eichner et al. 2007). However, in our paper, we mainly focus on how the extreme events, not the return intervals, affect the long-term correlations. According to the definition given by World Meteorological Organization (WMO), extreme events are defined when the absolute values exceed 2σ, where σ is standard deviation calculated from the whole time series. Also, climate extremes can be classified into two broad groups: (a) those based on simple climate statistics, which include extremes such as very low or very high daily temperatures, or heavy daily (monthly) rainfall amounts, etc. and (b) those driven by more complex mechanisms, such as drought, floods, or hurricanes, which do not necessarily occur every year at a given location (Easterling et al. 2000). In this paper, similar to the definition given by WMO and from Bunde et al. (2005), we define the extremes as values exceed a certain threshold in observational records.

The method we use here—DFA—has been considered robust. For positively correlated signals, even when up to 50% of the points are removed, the scaling behaviors will not be affected (Chen et al. 2002). Furthermore, according to Ma et al. (2010), even when the length of data loss segments is not fixed but random and follows a certain distribution, positively correlated signals still show practically no changes for up to 65% of data loss. However, we find that the loss of extreme values in signals may cause different results, which are interesting and imply close relations between extreme events and long-term correlations. To confirm this finding, we also use the power spectral density (PSD) analysis to test whether the long-term memory is affected or it is just a shortcoming of the method DFA. In this article, both artificial long-term correlated signals and observational records have been analyzed, respectively. For artificial signals obtained through a modified Fourier filtering technique (Makse et al. 1996), weaker long-term correlations have been found under the influence of extreme value loss. But for observational records, the situations become more complex. Some of the results are similar with that of the artificial data, but for some others, due to non-stationarity of the records and the resulting crossovers in the DFA curve, it is difficult to reach a reliable conclusion. Here we only show one possible case, of which the effect of extreme value loss is reflected in the extension of the scaling range. For further understanding of this phenomenon, we combine our findings with the previous work of return intervals and clustering of extreme events, and clustered positions of extreme values are found important to the long-term memory. Through a simple numerical test and from another perspective, we find that the emergence of clustered extreme values may be due to the long-term memory. For more details, please refer to Section 4.

The rest of this article is organized as follows: In Section 2, we will give a brief introduction of the analysis method and the data sets we used. Results based on the artificial data and the observational records are presented in Section 3. In Section 4, we make an extended discussion on the relations between extreme evens and long-term memory and conclude this paper.

2 Methodology and data

2.1 Methodology outline

2.1.1 Detrended fluctuation analysis

DFA is an important improvement of the classical FA for non-stationary signals based on the random walk theory. It is essential to distinguish trends from long-term fluctuations which are intrinsic to the data; thus, this method is reliable for the detection of long-term correlations. In this article, we employ the second-order DFA2, for our analysis (Kantelhardt et al. 2001). Suppose we have a fluctuating signal $T_i (i = 1, 2, 3, \cdots, N)$, one mainly considers the cumulated sum $Y_k = \sum_{i=1}^{k} \{T_i - \langle T \rangle\}$, where $\langle T \rangle$ is the mean of T_i and studies in non-overlapping time windows of length s. In each window, we calculate the local trend through second-order polynomial fitting and get the square fluctuation $F_s^2(j)$ as the variance of Y_k around this best quadratic fit, where j points to the j-th window. By averaging over all windows, we can get the root mean square fluctuations $F(s)$. For long-term correlated records, $F(s)$ scales as $F(s) \sim s^{\alpha}$, where the scaling exponent α is a self-similarity parameter which represents the signal's long-term power-law correlation properties. If $\alpha > \frac{1}{2}$, the signal is positively correlated; if $\alpha < \frac{1}{2}$, the signal is anti-correlated; if $\alpha = \frac{1}{2}$, there is no long-term correlation Peng et al. (1994). Notice the power-law auto-correlation function $C(s) \sim s^{-\gamma}$; the exponents α and γ are connected as $\gamma = 2(1 - \alpha)$ (Koscielny-Bunde et al. 1998; Kantelhardt et al. 2001).

2.1.2 Power spectral density analysis

PSD analysis is a conventional and well-known method to characterize the fractal properties of time series. It is convenient and useful when the time series is stationary, such as the artificial data we generate in our study. To determine the power spectral density $S(f)$, we first calculate the autocorrelation $C(s) = \langle [x(t+s) - \langle x \rangle][x(t) - \langle x \rangle] \rangle / \sigma^2$ of the signal, where s is the time lag and σ its variance. With $C(s)$, we can get $S(f)$ by Fourier transform. If the autocorrelation shows a scaling behavior for times larger than s, one also finds a scaling behavior of the power spectrum in the corresponding frequency region $f < 1/s$. For long-term correlated time series, with increasing frequency f, $S(f)$ decays by a power law, $S(f) \sim f^{-\beta}$, where β characterizes the fractal properties of signal Talkner and Weber (2000). Compared with the power-law autocorrelation function $C(s) \sim s^{-\gamma}$, it can be shown easily that $\gamma = 1 - \beta$. In this article, we apply this method to artificial signals to confirm our findings.

2.2 Data

In this study, we use a modified Fourier filtering technique Makse et al. (1996) to generate signals with different scaling behaviors. Nine groups of long-term correlated signals with length of 20,000 are generated, as well as a group of white noises for comparison. The values of the scaling exponents α are around 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95, respectively. To get a reliable numerical result, there are 1,000 samples in each group.

Meanwhile, we also use observational records for analyzing. Since many studies have shown that temperature records are positively long-term correlated (see Koscielny-Bunde et al. (1998), Fraedrich and Blender (2003), Lin et al. (2007), and Rybski et al. (2008) in this study), we only choose the daily mean temperature records over China. The data are obtained from China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn), and the latest 50 years (from 1961 to 2010) are chosen. Before analysis, we have standardized the data by subtracting the annual cycle and dividing by the seasonal standard deviation (Lennartz and Bunde 2011),

$$\tau_i = \frac{T_i - \langle T_i \rangle}{\langle (T_i - \langle T_i \rangle)^2 \rangle^{1/2}} \tag{1}$$

where T_i is the original temperature records and τ_i is the standardized data, which we use for analysis.

3 Results

At the very beginning, we would like to show one representative result from the artificial signals. Different from the way of removing segments in Chen et al. (2002); Ma et al. (2010), here we remove the extreme values in the records, and the way we select extreme values has been shown in Fig. 1e. We can see in Fig. 1a that if we remove the extreme values away from the artificial data, weaker scaling behavior can be found. To better confirm this findings, we also remove 5% values (not the extreme values) in the same signal for comparison. In Fig. 1f, the values are removed in the middle between two threshold values. From Fig. 1b, we can hardly see any changes of the DFA curves. To make sure that the weaker long-term memory shown in Fig. 1a is intrinsically originated, we also employ a conventional methods, PSD analysis for our study. In Fig. 1c, d, as expected, weaker long-term memory is found in Fig. 1c, while nearly no change is in Fig. 1d.

To make a more complete description of our findings, we repeat the experiment to artificial signals with different scaling DFA exponents. In Fig. 2, the scaling behaviors of all the long-term correlated signals become weaker after the removing of extreme values, but the difference varies. Obviously, the smaller the original exponent is, the less it changes, as shown in Fig. 2. When the signal is white noise, we even cannot find any differences. According to our definition of extreme events, we can always find extreme values, which are the top 5% (or 10%) largest (or smallest) in the record. The different changes we find in Fig. 2b may indicate that positions of extreme values are also very important to the long-term memory, which is related to the idea of the clustering of extreme events. We will give a more detailed discussion on this issue in Section 4.

According to Fig. 2b, the highest difference for 5% extreme value loss is only around 0.03. One may question that although there are differences between the two DFA curves, it seems too small. Can this difference be a statistical error? To exclude this possibility, we calculated the scaling exponents for signals with different percentages of extreme value loss. In Fig. 3, the more extreme values are removed, and the lower scaling exponents are obtained. Similar decreasing trend can also be found in the observational records. But if we remove the values randomly, no significant difference exists. Thus, we are more confident to say that the changes of the long-term correlations are due to the loss of extreme values.

To confirm our findings above, we use daily mean temperature records from stations over China for further analysis. For some stations, weaker scaling behaviors can arise after extreme values being removed, as shown in Fig. 4a–c; the more extreme values removed (up to 20%, as shown in Fig. 4), the smaller

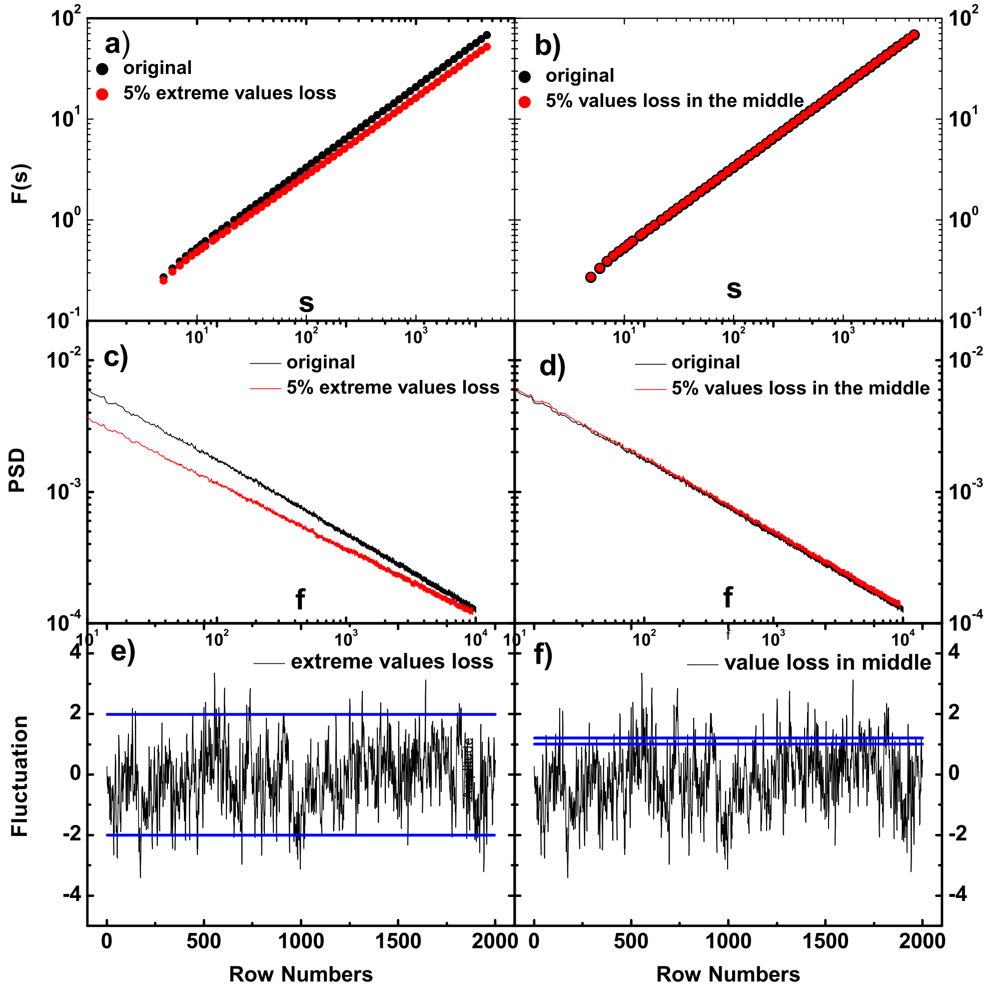


Fig. 1 DFA and power spectral density results of artificial records. Average results of one group, which contains 1,000 records of length 20,000, with the same Hurst exponent, are shown. **a** and **c**, where remarkable changes can be found in, represent the case when 5% extreme values are removed. While **b** and **d** are the results after the removal of 5% middle values, where nearly no changes can be seen. **e** and **f** show the two ways we remove values

exponents obtained. Furthermore, results of 20 observational records are shown in Fig. 5, where we can see a translation of the exponents values toward the left side of the axis, as the number of the extreme value loss becomes higher. However, not all the results from observational records keep in line with this tendency. Due to non-stationarity of the records and the resulting crossovers in the DFA curve, results from observational records are more complicated and difficult to reach a reliable conclusion. See Fig. 4d, which is one of the complicated cases we are interested to show. After the removal of extreme values, we cannot find weaker

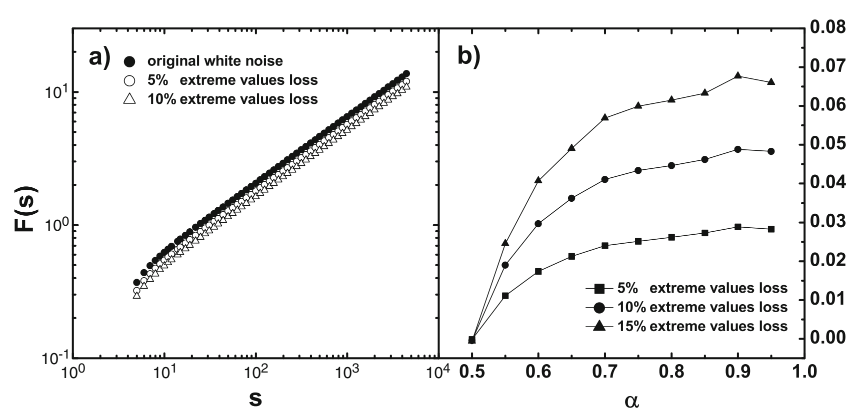


Fig. 2 **a** Results for the white noise, with 5% and 10% extreme value loss, respectively. No changes can be found. **b** Different changes of DFA exponents for original signals characterized with different scaling properties. We can see that the smaller the original exponent is, the less it changes

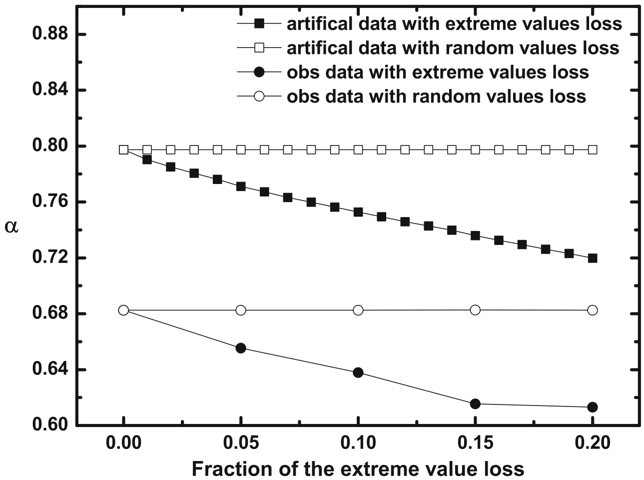


Fig. 3 Scaling exponents for signals with different percentages of extreme values removal and random values removal. The results for both the artificial data and the observational records are shown. More obvious changes can be found after more extreme values are removed. In contrast, no changes can be found if the values are removed randomly

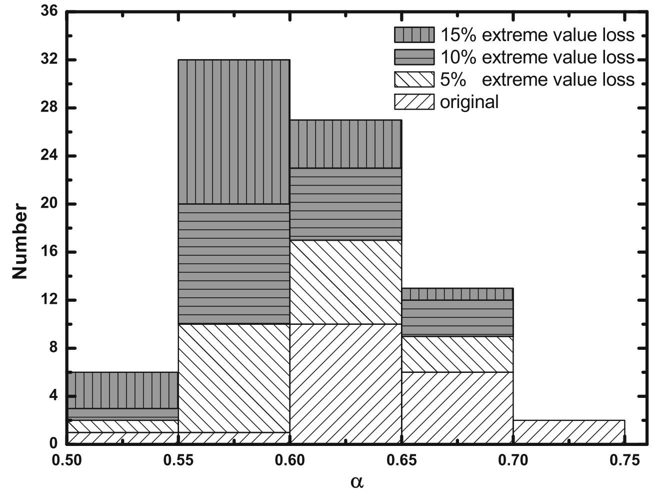
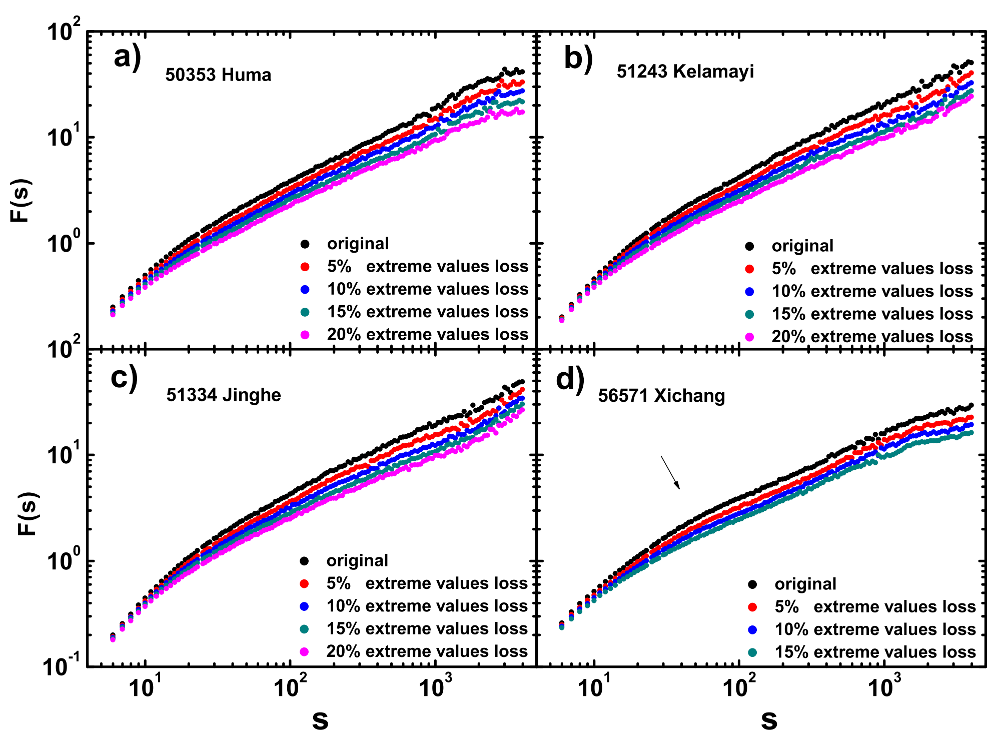


Fig. 5 Histogram of the Hurst exponents from 20 observational records. The more extreme values are removed, the smaller the exponents are; as shown in the figure, a remarkable translation of the histogram toward left side can be found

correlations, but extended scaling range. To understand this findings, we consider the original record and find the fluctuation is asymmetric, with more extremely low values exist. The removal of extreme events can reduce the number of extremely low values and make the record more uniform and symmetric. As a result, the scaling range of the DFA curve is extended to small time scales. However, here is only a rough discussion, and more studies on this issue are needed in the future.

Fig. 4 Results for the observational daily temperature records are shown. **a**–**c** are for Huma, Kelamayi, and Jinghe, respectively, where weaker long-term memory can be found after the removal of extreme events. **d** is the results for Xichang, which does not show a weaker long-term memory, but an extension of the scaling range



4 Discussion and conclusion

4.1 Discussion on the clustering of extreme events

The results above indicate that there should be close relations between long-term memory and extreme events. Combine with the interesting findings in previous work, that return intervals of extreme events in records will take similar long-term memory to that of the original records and long-term memory may lead to a pronounced clustering of extreme events, we can see that positions of the extreme events may be very important. Figure 6a shows the positions of extreme values we select to remove in one representative artificial long-term correlated signal, where one can find the clustering of extreme events easily. According to the work of Bunde et al. (2004, 2005) and Eichner et al. (2007), we also determine the probability density distribution of the return intervals by considering a group (1,000 samples) of artificial long-term correlated signals with length of 100,000 (five times as long as the artificial signals we use in Section 3, for better numerical results). Not only the return intervals of extreme values in both original data and shuffled data (which long-term memory is destroyed) but also the return intervals of middle values as shown in Fig. 1 are considered. See Fig. 6b. As expected, for long-term correlated records, larger probabilities can be found for larger return intervals as well as the smallest interval of 1, indicating a clustering of the extreme events. Removal of these extreme events means a damage to the old positions information, which may change the fractal properties and further induce weaker long-term memory. However, no significant clustering can be found for the middle values, since the probability density distribution is nearly the same with that of the shuffled signals. This could be one reason for the nearly unchanged long-term memory after the removal of middle values, as shown in Fig. 1b, d. However, we would like to mention that the existence of extreme values can also weaken the effect of the middle value loss. For further understanding, we make a simple numerical test here.

Consider a artificial long-term correlated signal of length 20,000 and a moving window of 1,000 points (a typical DFA scaling range when the records have 20,000 days), we estimate an quantity ε, which is proportional to the accumulations of the impacts from 1,000 points ahead. According to the power law autocorrelation function $C(s) \sim s^{-\gamma}$, ε is estimated as

$$\varepsilon_j = \frac{1}{1{,}000} \sum_{i=j}^{i=j+1{,}000} X_i (j + 1{,}001 - i)^{-\gamma},$$
$$j = 1, 2, 3, \cdots, 19{,}000 \qquad (2)$$

where X_i is the signal we use in this test, which is characterized with $\gamma = 0.4$. Thus, a new series ε_j^o (the shoulder mark o denotes the original long-term correlated record) of length 19,000 is obtained, which can represents the memory accumulations. For comparison, we also shuffle the signal to destroy its long-term memory, repeat the procedures above, and get anther series ε_j^s (the shoulder mark s denotes the shuffled record), as shown in Fig. 7a. As expected, we can find remarkable non-stationary with many clusterings of extremely high or low values, when the original signal is long-term correlated. However, surprisingly, the fluctuations of ε_j^s

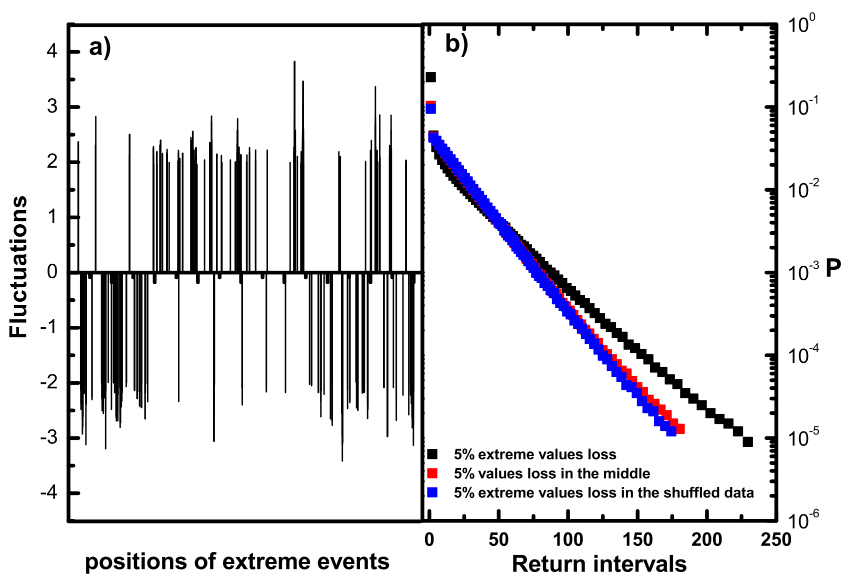


Fig. 6 **a** shows the positions of extreme values we removed from the data we study. **b** Probability density distributions of return intervals. *Black* for the return intervals of 5% extreme events in long-term correlated records; *red* for the return intervals of 5% middle values in long-term correlated records; *blue* for the return intervals of 5% extreme events in shuffled records

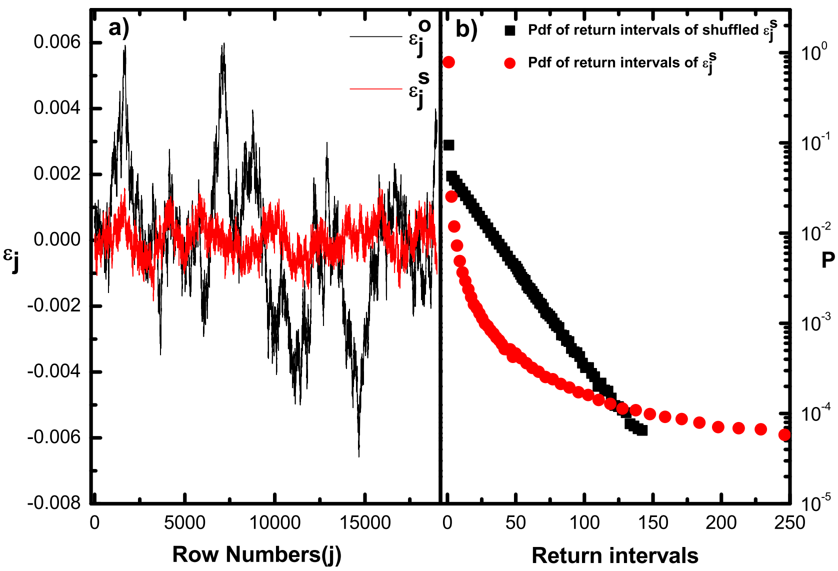


Fig. 7 New series ε_j obtained through the simple numerical test which represents the accumulation of memory. *Black line* (ε_j^o) shows the results from long-term correlated data, while *red line* (ε_j^s) refers to the results from shuffled data. Both series show remarkable clustering of extreme events. More clearer results are shown in **b** for the results from shuffled data, which is the probability density distributions of return intervals for both ε_j^s and shuffled ε_j^s

can also show clustering of extreme events, although the amplitude is smaller. See Fig. 7b, where the clustering of extreme values are shown more clearly. This indicates that, to some extent, the memory accumulations we estimated from 1,000 points ahead, or more precisely, the long-term memory, may induce clustering of extreme events. By thinking about this findings in a opposite direction, it may help us to understand why removal of clustered extreme events can affect long-term memories in the original signals. Thus, we should be very careful when we deal with records with missing points. Finding out reasons for the missing and estimating the probability that the missing values are extremely high (or low) are essential in the future studies.

4.2 Conclusion

According to the results in Chen et al. (2002) and Ma et al. (2010), it is known that the scaling behaviors obtained from DFA are robust, even when up to 50% of the points are lost in a considered record. In this article, not contradict with the former work, we have studied the effects of extreme value loss on long-term correlated time series. We found that the removal of extreme values can weaken the long-term memory in both the artificial and observational time series. The more extreme value loss, the weaker long-term memory is. With the same percentage of extreme value loss, bigger changes can be found when the long-term memory is stronger. However, situations for the observational records are more complex. In some cases, the effect of extreme value loss is reflected in the extension of scaling range, which may be due to the asymmetric distribution of the original records. In some other cases, due to non-stationarity of the records and the resulting crossovers in the DFA curve, we even cannot reach a reliable conclusion. Thus, better methods are needed for estimating the effect of extreme value loss in observational records.

By combining our findings with the previous work of return intervals and clustering of extreme events, we find close relations between extreme events and long-term memory. In long-term correlated records, extreme values are clustered. Removal of these clustered extreme values may change the fractal properties and further induce weaker long-term memory, which indicates positions of the extreme values seems important. Meanwhile, from the simple numerical test in Section 4.1, we can see that, to some extent, long-term memory may induce clustering of extreme events. These close relations require us to be very careful when dealing with records with long-term memory, where the extreme values may play an important role to any statistical properties we obtained.

Furthermore, we would like to mention an interesting work which has been published recently Mirzayof and Ashkenazy (2010). After a threshold-based dilution (data points that are smaller than the threshold are excluded) of exponentially distributed data, smaller DFA exponents are obtained, but still larger than 0.5, which indicates that long-term memory are partly preserved if only extreme values left in the record. Our findings that the removal of extreme values may weaken the long-term memory have contributed to this issue from a new perspective. All in all, extreme events have close relations with long-term memory.

Acknowledgements Many thanks are due to supports from National Natural Science Foundation of China (No.40775040) and from the Knowledge Innovation Program of the Chinese Academy of Sciences (IAP09301). N. Yuan also likes to acknowledge financial support from the China Scholarship Council (CSC).

References

Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley, HE (1992) Long-range correlations in nucleotide sequences. Nature 356:168

Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL (1994) Mosaic organization of DNA nucleotides. Phys Rev E 49:1685

Arneodo A, Bacry E, Graves PV, Muzy JF (1995) Characterizing long-range correlations in DNA sequences from wavelet analysis. Phys Rev Lett 74:3293

Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Matsa ME, Peng CK, Simons M, Stanley HE (1995) Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys Rev E 51:5084

Mantegna RN, Stanley HE (1995) Scaling behaviour in the dynamics of an economic index. Nature 376:46

Peng CK, Havlin S, Stanley HE, Goldberger AL (1995) Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5:82

Makse HA, Havlin S, Schwartz M, Stanley HE (1996) Method for generating long-range correlations for large systems Phys Rev E 53:5445

Liu Y, Cizeau P, Meyer M, Peng CK, Stanley HE (1997) Correlations in economic time series. Physica A 245:437

Koscielny-Bunde E, Bunde A, Havlin S, Roman HE (1998) Indication of a universal persistence law governing atmospheric variability. Phys Rev Lett 81:729

Bunde A, Havlin S, Kantelhardt JW, Penzel T, Peter JH, Voigt K (2000) Correlated and uncorrelated regions in heart-rate fluctuations during sleep. Phys Rev Lett 85:3736

Talkner P, Weber RO (2000) Power spectrum and detrended fluctuation analysis: application to daily temperatures. Phys Rev E 62:150

Easterling DR, Meehl GA, Parmesan C et al (2000) Climate extremes: observations, modeling, and impact. Science 289:2068

Kantelhardt JW, Koscielny-Bunde E, Rego HHA, Havlin S (2001) Detecting long-range correlations with detrended fluctuation analysis. Physica A295:41

Weber RO, Talkner P (2001) Spectra and correlations of climate data from days to decades. J Geophys Res 106:20131–20144

Chen Z, Ivanov PCh, Hu K, Stanley HE (2002) Effect of nonstationarities on detrended fluctuation analysis. Phys Rev E 65:041107

Fraedrich K, Blender R (2003) Scaling of atmosphere and ocean temperature correlations in observations and climate models. Phys Rev Lett 90:108501

Eichner JF, Koscielny-Bunde E, Bunde A, Havlin S, Schellnhuber H-j (2003) Power-law persistence and trends in the atmosphere: a detailed study of long temperature records. Phys Rev E 68:046133

Bunde A, Eichner JF, Havlin S, Kantelhardet JW (2004) Return intervals of rare events in records with long-term persistence. Physica A 342:308

Altmann EG, Kantz H (2005) Recurrence time analysis, long-term correlations, and extreme events. Phys Rev E 71:056106

Bunde A, Eichner JF, Kantelhardt JW, Havlin S (2005) Long-term memory: a natural mechanism for the clustering of extreme events and anomalous residual times in climate records. Phys Rev Lett 94:048701

Livina VN, Havlin S, Bunde A (2005) Memory in the occurrence of earthquakes. Phys Rev Lett 95:208501

Eichner JF, Kantelhardt JW, Bunde A, Havlin S (2006) Extreme value statistics in records with long-term persistence. Phys Rev E 73:016130

kantelhardt JW, Koscielny-Bunde E, Rybski D, Braun P, Bunde A, Havlin S (2006) Long-term persistence and multifractality of precipitation and river runoff records. J Geophys Res 111:D01106

Chen X, Lin G, Fu Z (2007) Long-range correlations in daily relative humidity fluctuations: a new index to characterize the climate regions over China. Geophys Res Lett 34:L07804

Eichner JF, Kantelhardt JW, Bunde A, Havlin S (2007) Statistics of return intervals in long-term correlated records. Phys Rev E 75:011128

Lin GX, Chen X, Fu ZT (2007) Temporal–spatial diversities of long-range correlation for relative humidity over China. Physica A 383:585

Rybski D, Bunde A, von Storch H (2008) Long-term memory in 1000-year simulated temperature records. J Geophys Res 113:D02106

Ma QDY, Bartsch RP, Bernaola-Galvan P, Yoneyama M, Ivanov P.Ch (2010) Effect of extreme data loss on long-range correlated and anti-correlated signals quantified by detrended fluctuation analysis. Phys Rev E 81:031101

Mirzayof D, Ashkenazy Y (2010) Preservation of long range temporal correlations under extreme random dilution. Physica A 389:5573

Lennartz S, Bunde A (20011) Distribution of natural trends in long-term correlated records: a scaling approach. Phys Rev E 84:021129