The Different Types of Noise and How They Effect Data Analysis

Armin Bunde*

DOI: 10.1002/cite.202300031

Dedicated to Prof. Dr. rer. nat. Jörg Kärger on the occasion of his 80th birthday

Random (noisy) processes can be characterized by the way consecutive data are correlated. The data can be uncorrelated (white noise), short-range correlated (often called red noise), or long-range correlated (sometimes called pink noise). Here we describe the properties and applications of these different kinds of noise. We discuss, how they influence (i) the diffusion process, (ii) the occurrence of rare extreme events and (iii) the detection of an external trend that is superimposed on the noise; (ii) and (iii) are particularly relevant in the context of detecting anthropogenic global warming by data analysis.

Keywords: Detrended fluctuation analysis, Diffusion, Long-range correlations, Noise

Received: February 24, 2023; *revised:* July 10, 2023; *accepted:* July 10, 2023

1 Introduction

Noisy patterns and random processes are everywhere in our daily life. Our heartbeat and our breathing frequency are random, our gait lengths are random, the weather is random, catastrophic events are random, and in case we are active on the stock market, we also know that daily gains and losses are random. In all games of chance that have been known since primeval times consecutive events are independent, and the first mathematical studies of random processes were actually on dice games, for example by Christian Huygens on the gambler's ruin [1].

Probably the most important and best studied random process is diffusion. Diffusion can be modelled by a random walker who walks randomly in space. Applications range from the spreading of epidemics and innovations to the random movement of atoms and ions in complex materials. For recent reviews, see two books [2, 3] edited by our jubilee Jörg Kärger, who contributed significantly to this important field.

It was believed until last century that most random processes in nature indeed are like roulette or dice games. In recent years there is growing evidence that not all random processes in nature consist of independent steps, i.e., are *white* noise, but a large number of processes show correlations over time, some of them (like the climate) even correlations over many centuries. When the correlations have a finite range, one often refers to these processes as *red* noise, when the range is infinite, they are sometimes referred to as *pink* noise.

Here, we compare these three types of noise and discuss the consequences of the different kinds of correlations (i) on diffusion properties, (ii) on the occurrence of rare extreme events, and (iii) on the detection of external deterministic trends in noisy data. Item (iii) sounds a bit abstract but is important in the context of global warming when searching for anthropogenic effects in climate data. Here the challenge is to distinguish anthropogenic effects driven by man-made external trends from natural effects driven by long-term correlated noise.

2 Characterization of Noise

Consider a record y_i of discrete numbers, where the index i runs from 1 to L and all data are drawn from the same distribution, for example from a Gaussian. Let us assume that the length L of the record is very large and can be considered as infinite. Without loss of generality, we assume that the mean of the data is zero and the variance equal to one. (When records with periodic trends are considered, they must be first detrended).

For quantifying correlations in the record, it is convenient to look at the autocorrelation function $C(s) \equiv \langle y_i y_{i+s} \rangle \equiv \frac{1}{L-s} \sum_{i=1}^{L-s} y_i y_{i+s}$. By definition, $C(0) = 1$. When all data are independent of each other, $C(s)$ vanishes for $s > 0$. In this case the noise is called white noise.

In many cases, successive data y_i and y_{i+1} depend on each other. When the correlations have a finite range, the noise is

[1]Prof. Dr. Armin Bunde
(arminbunde00@googlemail.com)
[1]Justus-Liebig-Universität Gießen, Institut für Theoretische Physik, Heinrich-Buff-Ring 16, 35392 Gießen, Germany.

called short-range correlated or red noise [4]. This happens, for example, when $C(s)$ decays exponentially,

$$C(s) = p^s, \quad s = 0, 1, 2, \ldots \tag{1}$$

with $|p| < 1$. By definition, p is identical to the lag-1 autocorrelation $C(1)$. For $p > 0$, $C(s)$ can be written as $C(s) = \exp(-s/\xi)$, where $\xi \equiv \int_0^\infty C(s)ds = 1/|\ln p|$ denotes the correlation (persistence) length. For $p < 0$, $C(s)$ describes anticorrelations.

Of great importance in nature are noisy data, where the autocorrelation function decays by a power law (see, e.g., [5, 6]),

$$C(s) = \frac{(2-\gamma)(1-\gamma)}{2} s^{-\gamma}, \quad s > 0 \tag{2}$$

with γ between 0 and 1. Such correlations are named "long-range" since the mean correlation length ξ diverges in the limit of infinite record length L. These data are sometimes referred to as pink noise. With increasing γ the correlations decrease. In the limit of $\gamma = 1$ we have white noise.

In the past decades, long-range correlated noise has increasingly become the focus of interest, since a large number of well-known noisy data in nature are long-range correlated. River flows [5–12], atmospheric and sea surface temperatures [13–31], sea level heights [32–34], or wind fields [35] and midlatitude cyclones [36] are long-range correlated. In addition, there are indications that also the intervals between consecutive large earthquakes are long-term correlated [37]. In the human body, noncoding DNA sequences [38, 39], heartbeat intervals [40, 41] and gait lengths [42] show long-range correlations, and even in man-made noise long-range correlations can be observed: examples are the volatility in financial records [43, 44] and, perhaps surprisingly, human writing [45, 46].

Long-range correlated records can also be characterized by the power spectral density $S(f) = |y(f)|^2$, where $y(f)$ is the Fourier transform of $\{y_i\}$. With increasing frequency f, $S(f)$ decays by a power law

$$S(f) \sim f^{-\beta} \tag{3}$$

with $\beta = 1 - \gamma$ [4]. For uncorrelated data, $\beta = 0$, i.e., the spectrum is flat. Eq. (3) can be used to generate Gaussian distributed long-range correlated records with fixed γ [4]. The numerical procedure consists of two steps: (i) one generates uncorrelated Gaussian data and transforms them to Fourier space; (ii) one multiplies the result by $f^{-\beta/2}$ and transforms it back to time space. The final result is long-range correlated with $\gamma = 1 - \beta$. By definition, $S(f)$ follows Eq. (3).

Fig. 1 compares parts of an uncorrelated record (a) with three long-range correlated records, with $\gamma = 0.8$ (b), 0.4 (c), and 0.2 (d); all series have been generated by the algorithm described above. The red line is the moving average over 30 data points. For the uncorrelated data, the moving average is close to zero, while for the long-range correlated data sets, the moving average can have large deviations from the mean, forming some kind of mountain valley structure. Let us assume, for example, that the data in Fig. 1d represent annual temperatures, and let us further assume that we know only the first 250 data points (first 250 years) and want to predict what will happen in the next 50 years. If we would know nothing about the natural persistence of the record and assume tacitly that the record is a superposition of white noise describing the natural fluctuations and some linear anthropogenic trend, we were inclined to draw a trend line and to extrapolate the line into the future to estimate future warming. However, since the data are long-range correlated, this simple procedure (which actually has been used quite often in geoscience in the past century) will fail, and one needs more elaborate techniques to distinguish natural from deterministic trends. The figure shows also that, as a consequence of the strong persistence, the extremes are not distributed randomly in time, but occur in clusters. We will come back to both issues, clustering of extreme events and trend detection, in Sects. 6 and 7.

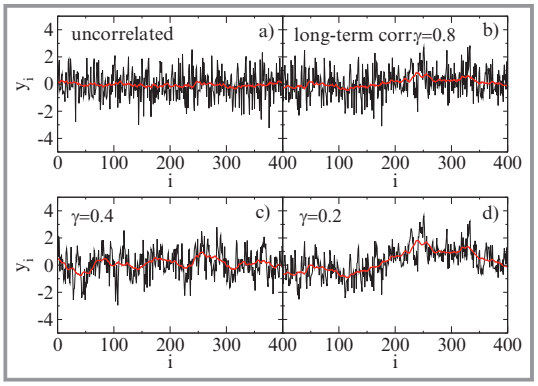

Figure 1. Comparison of an uncorrelated record with three long-range correlated records with $\gamma = 0.8$, 0.4, and 0.2. The full line is the moving average over 30 data points.

First, we study how diffusion depends on the type of noise considered and how the diffusion properties can be used to detect the type of noise.

3 Diffusion

We focus on a simple random walk along a line. A random walker starts at X_0 and carries a long list of random numbers y_i, $i = 1, 2, \ldots, L$, with zero mean and variance one. In the ith time step, the walker steps to the right, when y_i is positive, and to the left, when y_i is negative. The step length is $|y_i|$. After s steps, the position of the random walker is X_s. Accordingly, after L steps the walker generated the sequence $X_1, X_2, \ldots, X_{L-1}, X_L$.

We would like to know what the typical distance is the random walker has traveled after s steps. To this end, we

consider also the positions X_1, X_2,…, X_{L-1}, X_{L-s} as starting points of a random walk, and obtain $L_s \equiv L+1-s$ random walk sequences of length s. Let us denote each of these segments by the index ν and the s random numbers in segment ν by $y_i^{(\nu)}$, $i = 1,\ldots, s$. Then the distance the random walker travelled in segment ν is given by

$$\Delta X_\nu(s) = \sum_{i=1}^{s} y_i^{(\nu)} \tag{4}$$

Since the random numbers have zero mean, the mean distance averaged over all segments ν is zero. In order to find the typical range covered by the random walker after s steps it is convenient to look at the mean square displacement, where $(\Delta X_\nu(s))^2$ is averaged over all segments ν,

$$\left\langle (\Delta X)^2(s) \right\rangle \equiv \frac{1}{L_s} \sum_{\nu=1}^{L_s} \left(\sum_{i,j=1}^{s} y_i^{(\nu)} y_j^{(\nu)} \right) = s + \sum_{i \neq j}^{s} \left\langle y_i y_j \right\rangle \tag{5}$$

Using the definition of the autocorrelation function $C(s)$ and the fact that the variance of the random numbers is 1, Eq. (5) can be written as [47]

$$\left\langle (\Delta X)^2(s) \right\rangle = s + 2 \sum_{m=1}^{s-1} (s-m) C(m) \tag{6}$$

Eq. (6) allows us to determine the mean square displacement for the different kinds of noise. For white noise, $C(m) = 0$ for $m \geq 1$, and thus $\langle (\Delta X)^2(s) \rangle = s$, which is the well-known diffusion law. For short-range correlated noise, $C(m) = p^m$, and Eq. (6) yields

$$\left\langle (\Delta X)^2(s) \right\rangle = s + \frac{2p}{(1-p)^2}(s - ps - 1 + p^s) \tag{7}$$

For small correlations, where p tends to zero, $\langle (\Delta X)^2(s) \rangle = s$, as expected. For strong correlations where the correlation length ξ is large, we can expand p into $p \equiv \exp(-1/\xi) \approx 1 - 1/\xi + (1/\xi)^2/2$. It is easy to see that in this case,

$$\left\langle (\Delta X)^2(s) \right\rangle \cong \begin{cases} s^2, & s \ll \xi \\ 2\xi(s-\xi), & s \gg \xi \end{cases} \tag{8}$$

Accordingly, for s well below ξ, the diffusion is ballistic (as expected), while well above ξ, diffusion is fast with a diffusion constant proportional to ξ.

In the presence of anticorrelations, p is negative. For strong anticorrelations, Eq. (7) yields

$$\left\langle (\Delta X)^2(s) \right\rangle \cong 1/2 \begin{cases} 1 + s/\xi + (1 - s/\xi)(-1)^s, & s \ll \xi \\ 1 + s/\xi, & s \gg \xi \end{cases} \tag{9}$$

Accordingly, for s well below ξ, $\langle (\Delta X)^2(s) \rangle$ oscillates between 1 and s/ξ. Well above ξ we observe regular diffusion with a very small diffusion constant $1/\xi$. Such a behavior can be observed in solid ionic conductors where ions are coupled by the Coulomb interaction. When an ion jumps to a neighboring site, it is out of equilibrium and the Coulomb forces tend to drive the ion back to where it came from. This behavior, which is described by Eq. (9), has been termed cage effect by Funke [48]. The ion is like in a cage and needs many foreward-backward jumps to escape the cage, and this leads to the well-known dispersion of the ionic conductivity in ionic glasses [48].

For long-range correlated noise where $C(m)$ decays by a power law, Eq. (6) yields [47]

$$\left\langle (\Delta X)^2(s) \right\rangle = s^{2-\gamma} \equiv s^{2\alpha} \tag{10}$$

where α can be identified with the Hurst[1)] exponent [5, 13]. For $\gamma \to 1$, we recover the white noise result, for $\gamma \to 0$, diffusion tends to ballistic motion.

Eqs. (8)–(10) allow to determine the kind of noise by generating a random walk: (i) when $\langle (\Delta X)^2(s) \rangle$ is proportional to s already for small s, the data can be considered as white noise; (ii) when $\langle (\Delta X)^2(s) \rangle$ starts with an exponent close to 2 at small s and then crosses to normal diffusion with an exponent 1 for larger s, the data can be considered as short-range correlated. Finally, when (iii) $\langle (\Delta X)^2(s) \rangle$ shows power law behavior with an exponent above 1, the data are long-range correlated.

4 Hierarchical Detrended Fluctuation Analysis

In several important applications, for example in climate records, external deterministic trends (e.g., by global warming) may be superimposed to the natural fluctuations. Since the effects of trends and long-range correlations resemble each other (see Fig. 1), there is a general problem to distinguish between them. To detect long-range correlations in the presence of smooth polynomial trends of order $n-1$, the hierarchical detrended fluctuation analysis (DFAn) [49] has been introduced. DFAn is a simple modification of (4): In each segment ν of length s, we now consider $\Delta X_\nu(l) = \sum_{i=1}^{l} y_i^{(\nu)}$, where l runs from 1 to s, and fit $\Delta X_\nu(l)$ by a polynomial of order n. Then we determine the variance $V_\nu^{(n)}(s)$ of $\Delta X_\nu(l)$ around the polynomial, and average $V_\nu^{(n)}(s)$ over all segments ν of length s to obtain the detrended fluctuation function

$$F^{(n)}(s) = \left(\frac{1}{L_s} \sum_{\nu=1}^{L_s} V_\nu^{(n)}(s) \right)^{1/2} \tag{11}$$

1) Harald Hurst was a practical engineer who was interested in the long-range storage capacity of reservoirs. He discovered, in the middle of last century, long-range persistence when studying the fluctuations of the water level in the Nile River.

The fitted polynomial represents the local trend in each segment ν. When n-order polynomials are used, the analysis is called DFAn, $n = 0, 1, 2, \ldots$.

In trend-free data, $(F^{(n)}(s))^2$ scales with s in the same way as $\langle(\Delta X)^2(s)\rangle$ (see Eqs. (8)–(10)), irrespective of n. Trends of order $k-1$ in the original data are eliminated by DFAk, while they contribute to DFA($k-1$) and DFA($k-2$) etc. This allows to determine the correlation exponent γ in long-range correlated data in the presence of trends. For example, in the presence of a linear trend, DFA0 and DFA1 will exaggerate the asymptotic exponent $\tilde{\alpha}$, while DFA2, DFA3 etc. will show the same slope α characterizing the natural fluctuations of the record. In [21] the difference between $\tilde{\alpha}$ and α was used to estimate qualitatively the strength of anthropogenic trends in Central Asia.

An extension of DFAn is the *multifractal* DFAn [50], MF-DFAn, where instead of (11) the more general average

$$F_q^{(n)}(s) \equiv \left(\frac{1}{L_s} \sum_{\nu=1}^{L_s} \left(V_\nu^{(n)}(s) \right)^{q/2} \right)^{1/q} \tag{12}$$

is considered, with q between $-\infty$ and $+\infty$. For $q \ll -1$ the small fluctuations dominate the sum, while for $q \gg 1$ the large fluctuations are dominant. It is generally assumed that the q-dependent average scales with s as

$$F_q^{(n)}(s) \sim s^{\eta(q)} \tag{13}$$

By definition, $\eta(2) = \alpha$. If $\eta(q)$ is independent of q, $F_q^{(n)}(t) \sim s^{\alpha}$, and both large and small fluctuations scale the same. In this case, a single exponent is sufficient to characterize the record, which then is referred to as *monofractal*. If $\eta(q)$ is not identical to α, the record is called *multifractal* (for a review, see [51]). In this case, the dependence of $\eta(q)$ on q characterizes the record. Temperature records can be considered as monofractal, while precipitation and river flow records with their pike-structure usually show weak to modest multifractality [7, 10, 25]. Multifractal data are more erratic than monofractal data and can show extreme events out of the blue ("black swans"). Pronounced multifractality occurs in financial data sets (see, e.g., [43, 44, 52] and references therein), for example in the daily returns of an asset.

5 Long-Range Correlations in Climate Records

Fig. 2 shows three representative temperature, river run-off, and precipitation records. The data below or above the periodic seasonal trend are in blue or red. The patches in Fig. 2a and 2b indicate correlations. In a correlation analysis, the periodic seasonal trend must be subtracted from the data.

Fig. 3 shows the results of a DFAn analysis, for six temperature, precipitation and river run-off records. For continental temperatures, the exponent α is around 0.65, while for island stations and sea surface temperatures α is considerably higher. There is no crossover towards uncorrelated

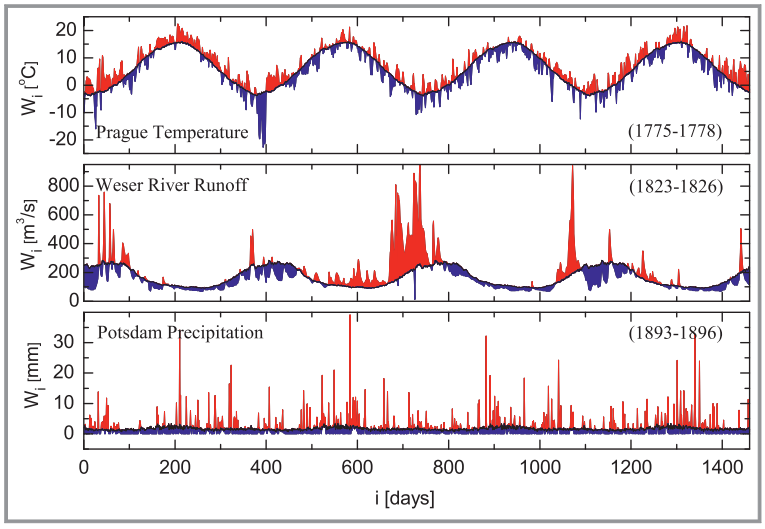

Figure 2. The figure shows 4-year-sections of a) the temperature record of Potsdam, b) the run.off record of the Weser River near Vlotho, and c) the precipitation record of Potsdam.

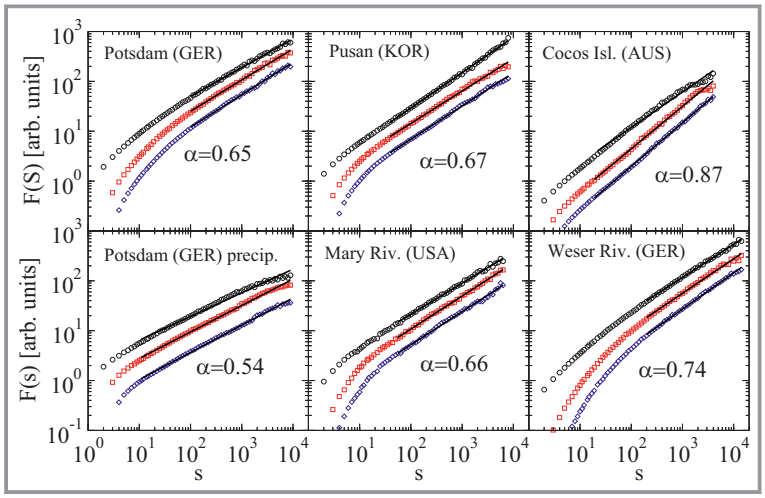

Figure 3. DFAn analysis of three temperature records (a–c), one precipitation record (d) and two run-off records (e,f). The symbols correspond to DFA0, DFA1, and DFA2 (from top to bottom). The straight lines are power law fits. The numbers denote the asymptotic slopes of the DFA2 fluctuation functions. The bending down of the fluctuation functions for very small s is an artefact of DFAn.

behavior at larger time scales. For the precipitation data, α is close to 0.55, not being significantly larger than for uncorrelated records. In the Pusan record, the DFA0 curve bends up, which is a clear indication of an external trend. A more comprehensive analysis has shown that the distribution of α-values is quite broad for run-off, sea-surface temperature, and precipitation records, but quite narrow for continental temperatures where α is close to 0.65 [17]. This feature can be used as an efficient test bed for climate models [22, 53, 54] and paleo reconstructions [55].

The time window accessible by DFAn is typically 1/4 of the length of the record. For instrumental records, the time window is thus restricted to about 50 years. For extending this limit, one can either take reconstructed records or model data, which roughly range up to 2000 years. Both have, of course, large uncertainties, but it is remarkable that exactly the same kind of long-range correlations can be found in model data, thus extending the time scale where long-range memory exists to at least 500 years [20, 22]. Paleo reconstructions show stronger long-range correlations where the exponent α is exaggerated [55].

6 Clustering of Extreme Events

Next, we consider the rare events in a record. Understanding (and predicting) the occurrence of extreme and often hazardous events is one of the major challenges in science (see, e.g., [52]). An important quantity here is the time interval between large events (see Fig. 4), and by understanding the statistics of these return intervals one aims to better understand the occurrence of the extremes.

6.1 Return Intervals

Since extreme events are, by definition, very rare and the statistics of their return intervals poor, one usually studies also the return intervals between less extreme events, where the data are above some threshold q (not to be confused with the multifractal parameter introduced in Sect. 4) and where the statistics is better, and hopes to find some general "scaling" relations between the return intervals at low and high thresholds, which then allows to extrapolate the results to very large, extreme thresholds (see Fig. 4).

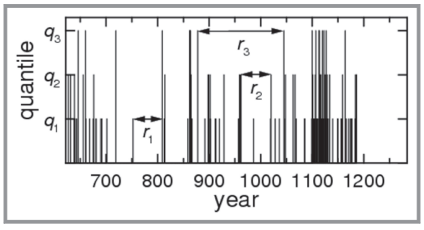

Figure 4. Illustration of the return intervals for three equidistant threshold values q_1, q_2, q_3 for the water levels of the Nile at Roda (near Cairo, Egypt). One return interval for each threshold (quantile) q is indicated by arrows.

Let us assume that only a small fraction f_q of the data is above q. Then the mean return interval R_q is identical to $1/f_q$. For uncorrelated data, the return intervals r are independent of each other. Also, for short-range correlated data the return intervals are uncorrelated, as long as R_q is well above the correlation length ξ.

For long-range correlated records, the return intervals are arranged in a long-range correlated fashion: long intervals are more likely to be followed by long intervals and short intervals by short intervals. This leads to a clustering of extreme events. To see this effect in the relatively short climate records, we follow [56] and distinguish only between two kinds of return intervals, "small" ones (below the median) and "large" ones (above the median) and determine the mean R_q^+ of those intervals that follow a large (+) interval and the mean R_q^- of those intervals that follow a small (–) return interval.

Fig. 5 shows both quantities (calculated numerically for long-range correlated Gaussian data) as a function of the correlation exponent γ. The lower dashed line is R_q^-/R_q, the upper dashed line is R_q^+/R_q. In the limit of vanishing long-range memory, for $\gamma = 1$, both quantities coincide, as expected. Fig. 5 also shows R_q^+/R_q and R_q^-/R_q for five climate records with different values of γ. One can see that the data agree very well, within the error bars, with the theoretical curves. The error bars have been obtained by .numerical simulations of records with the same correlation exponent γ and the same lengths as the climate records.

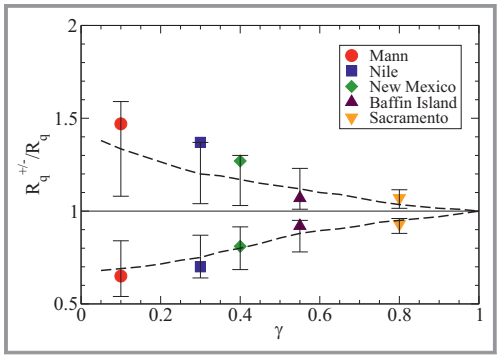

Figure 5. Mean of the (conditional) return intervals that either follow a return interval below the median (lower dashed line) or above the median (upper dashed line), as a function of the correlation exponent γ, for five long reconstructed and natural climate records. The theoretical curves are compared with the corresponding values of the climate records (from right to left): The reconstructed run-offs of the Sacramento River, the reconstructed temperatures of Baffin Island, the reconstructed precipitation record of New Mexico, the historic water levels of the Nile and one of the reconstructed temperature records of the Northern hemisphere (Mann record) (after [56]).

Therefore, long-range memory can explain the clustering of extremes that has been observed for example in the floods in Central Europe during the Middle Ages, or in the historic water levels of the Nile River, which are shown in Fig. 4 for 663y. A similar clustering was observed for

extreme floods, winter storms, and avalanches in Central Europe (see, e.g., Fig. 2 in [9], Figs. 4.4, 4.7, 4.10, and 4.13 in [57], and Fig. 66 in [58]).

6.2 Distribution of the Return Intervals

To quantify the risk of encountering an extreme event one needs to study the distribution of the return intervals $P_q(r)$ and the related exceedence probability $S_q(r) \equiv \int_r^\infty P_q(r')dr'$. In uncorrelated data, the probability that $l = r - 1$ consecutive data points are below q is $(1 - f_q)^l \equiv (1 - 1/R_q)^l \equiv \exp(\ln(1 - 1/R_q)l \approx \exp(-l/R_q) \approx \exp(-r/R_q)$ for $R_q \gg 1$. Accordingly, $P_q(r) = (1/R_q)\exp(-(r/R_q)$, which satisfies $\int_0^\infty drP_q(r) = 1$. Thus, the probability of finding a return interval larger than r is simply

$$S_q(r) = \exp\left(-r/R_q\right), \quad R_q \gg 1 \tag{14}$$

Short-range correlated data behave in the same way, as long as R_q is well above ξ, and a sequential ordering cannot occur.

The situation is different (but surprisingly not much more complicated) in long-range correlated records. It has been shown by extensive numerical calculations [56] that $S_q(r)$ can be well approximated by

$$S_q(r) \equiv \exp\left[-b_\gamma\left(r/R_q\right)^\gamma\right] \tag{15}$$

where the parameter $b_\gamma = \left[\int_0^\infty dx\exp(-x^\gamma)\right]^\gamma$ is determined by the condition $R_q = \int_0^\infty dr\left(rP_q(r)\right)$. In the limit $\gamma \to \infty$, S_q reduces to (14). The form of $S_q(r)$ indicates that return intervals both well below and well above their average are considerably more frequent for long-range correlated data than for uncorrelated data, which also is a manifestation of the clustering of the extremes.

The characteristic stretched exponential behavior of $S_q(r)$ is difficult to see in the comparatively short observational climate records but can be observed in long historic and reconstructed climate records [56]. Stretched exponential behavior appears in many other contexts, e.g., in materials science [59] (where it is known as Kohlrausch-Williams-Watts behavior) and in earth sciences [4]. In Sect. 6.4, we will take a short detour into linguistics and into the mathematical analysis of literary texts, where stretched exponentials can be observed nicely.

6.3 Prediction of Extreme Events: The Hazard Function

Next we ask: Given that the last "extreme" event above q occurred t time steps ago, what is the probability $W_q(t; \Delta t)$ that in the next Δt time steps an extreme event occurs (see Fig. 6)? This probability is of great importance since it specifies, at any time t, the risk associated with an extreme event. It can be easily verified that W (which is also called "hazard function"), is related to P and S by

$$W_q(t;\Delta t) = \frac{\int_t^{t+\Delta t} P_q(r)dr}{\int_t^\infty P_q(r)dr} \equiv 1 - \frac{S_q(t+\Delta t)}{S_q(t)} \tag{16}$$

Here $\int_t^{t+\Delta t} P_q(r)dr$ is the probability that an extreme event occurs between t and $t + \Delta t$, and $\int_t^\infty P_q(r)dr$ is a normalization factor that ensures $W_q(t;\infty) = 1$.

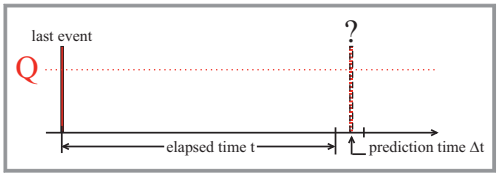

Figure 6. Hazard function: Given that at $t = 0$ there is an extreme event above q and there is no further extreme in the next t time steps. What is the probability that the next extreme event occurs between t and $t + \Delta t$?

For white noise,

$$W_q(t;\Delta t) = 1 - \exp\left(-\Delta t/R_q\right) \tag{17}$$

is independent of how much time t elapsed since the last event. Accordingly, an extreme event can never be overdue (when t exceeds the period R_q of the event). For the most relevant case $\Delta t \ll R_q$, we obtain $W_q(t; \Delta t) = \Delta t/R_q$.

For long-range correlated records, we approximate S_q by (15) and obtain

$$W_q(t;\Delta t) \approx 1 - \frac{e^{-b_\gamma\left((t+\Delta t)/R_q\right)^\gamma}}{e^{-b_\gamma\left(t/R_q\right)^\gamma}} \tag{18}$$

For $t = 0$, $W_q(0;\Delta t) = 1 - \exp(-b_\gamma(\Delta t/R_q)^\gamma)$. Accordingly, just after an extreme event the probability for observing another extreme event is drastically enhanced compared with white noise. For $\Delta t/R_q = 1/1000$, for example, $W = 0.001$ for white noise, while for $\gamma = 1/2$ where $b_\gamma = \sqrt{2}$, $W \cong 0.044$. For $\gamma = 1/4$, $b_\gamma \cong 2.2$ and thus $W_q(0;\Delta t) \cong 0.32$.

On the other hand, for $t/R_q = n$ and $\Delta t/R_q \ll 1$, W_q decreases with n as $(\Delta t/R_q)b_\gamma/n^{1-\gamma}$. Accordingly, when within the first n mean return intervals no further extreme event has occurred, the occurrence probability of a second event is strongly reduced compared with white noise.

6.4 Detour: Return Intervals in Literary Texts

Literary texts consist of words, some are rare, some are frequent. The question is how literary texts are structured and to which extent the structure depends on the language and the author. Here we are interested in the way rare words are incorporated in a text. To this end we rank the words, the most frequent word has rank 1, the next frequent word rank 2, and so on. Punctuations are considered as words. When we specify each word in a text by its rank, the text is like a record of numbers and can be analyzed in the same way.

It is well known that the frequency of a word as function of its rank follows approximately a power law (the famous Zipf's law [60]), and also the number of different words in a text increases with its length roughly by a power law [61].

To characterize the rhythm of a text by the arrangement of lower and higher ranked words we follow [46] and consider, in a text with L words, the fraction $f_Q \equiv L_Q/L$ of the rarest words that by definition have a rank above Q. We focus on the return intervals between consecutive words with a rank above Q. The mean interval length $R_Q = 1/f_Q$ represents the natural length scale. According to [46], the return intervals are long-range correlated with an exponent γ around 0.3 but contain also a large fraction of white noise ($\gamma = 1$).

We are interested in the probability $S_Q(r)$ that in a given text an interval between consecutive words with rank above Q, is longer than a certain interval length r. Fig. 7 shows $S_Q(r)$ for five texts and R_Q = 2, 4, 8, 16, 32 and 64. The dashed lines show S_Q for the shuffled texts where $S_Q(r) = \exp(-r/R_Q)$ for $R_Q \gg 1$. Accordingly, deviations from a simple exponential can be viewed as measure of the complexity of a text. For $R_Q = 2$, S_Q still follows, for most texts, a simple exponential. But already for $R_Q \geq 4$, $S_Q(r)$ follows the *stretched* exponential form Eq. (15), with an exponent (here denoted by β) roughly between 0.7 and 0.75. As a consequence, just after a rare word the probability for observing another rare word is strongly enhanced compared with white noise, see the discussion following Eq. (18). The reason for the universal behavior here (independence of language and author) is certainly the universal behavior of the human brain. One may speculate that when the brain uses a rare word, it is in some kind of "excited" state which makes it more likely to produce rare words again, and when it is in a "quiet" state it prefers to continue with not so rare words. The mathematical reason for the stretched exponential behavior is certainly long-range memory in the return intervals.

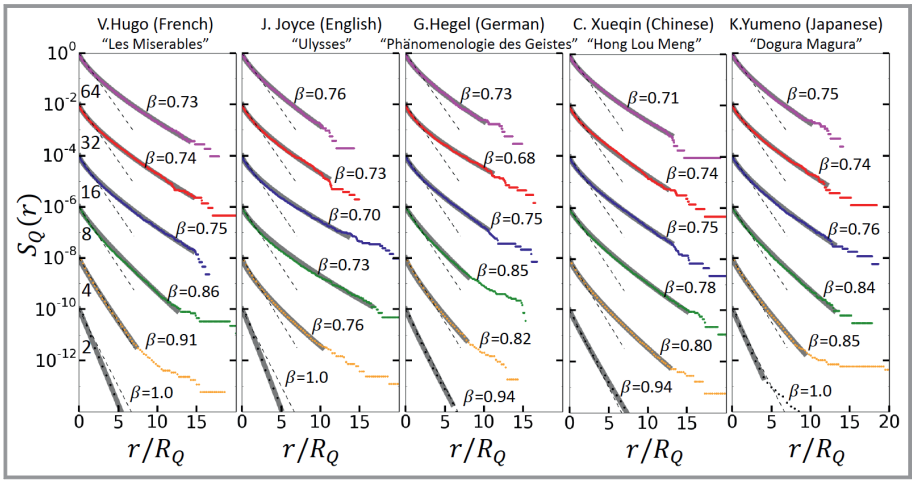

Figure 7. The probability $S_Q(r)$ that in a text the return intervals between words with rank above Q (see Fig. 1) exceed a certain length r, for five texts and six R_Q-values. By definition, $S_Q(0) = 1$. For transparency, we have multiplied S_Q for R_Q = 32, 16, 8, 4, and 2 by 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, and 10^{-10}, respectively. The dots are the numerical results. The colored lines are the best fit to Eq. (15). The quality of the fits is striking. The dashed straight lines are for the shuffled texts (from [46]).

7 Detection of External Trends

Finally, we consider how the presence of an external deterministic trend can be discovered in noisy records W_i, for example in those shown in Fig. 2. This question is particularly relevant for the detection of anthropogenic effects in climate data. The normal approach is as follows: first we subtract the periodic seasonal trend from the data set (see Fig. 2) and obtain a record $\{y_i\}$, $i = 1\ldots, L$, with zero mean. Then we study the correlation properties of $\{y_i\}$ by DFA2.

To estimate a trend in this record, we perform a regression analysis: from the regression line $r_i = bi + d$ we obtain the magnitude of the trend $\Delta = b(L-1)$ as well as the fluctuations around the trend line, characterized by the standard deviation $\sigma = \left[(1/L)\sum_{i=1}^{L}(y_i - r_i)^2\right]^{1/2}$. The relevant quantity we are interested in is the (positively defined) *relative trend*

$$x = |\Delta|/\sigma \tag{19}$$

When a certain relative trend has been measured in a data set, the central question is, whether this trend can be explained by the natural variability of the data set or not ("detection problem", see, e.g. [20]). To answer this question, one needs to know the probability $P(x;\ L)dx$ that in model records with the same persistence properties as the considered data set, a relative trend between x and $x + dx$ occurs. From P we derive the *p-value*

$$p(x;L) = \int_x^\infty P(x';L)dx' \tag{20}$$

By definition, p is the probability that the relative trend in the record is above x, with $p(0; L) = 1$.

If the p-value of the observed relative trend x is below 0.05, one usually assumes that the considered trend cannot be fully explained by the natural variability of the record, and the trend is called *significant*. For p below 0.01, the trend is called *highly significant*. The values x_{05} and x_{01} where $p = 0.05$ or 0.01 are called quantiles. Accordingly, for x between x_{05} and x_{01}, the trend is significant, while for x above x_{01}, the trend is highly significant. For both white and red noise, $P(x; L)$ follows [62, 63]) a student-t distribution for large record length L,

$$P(x;L) \propto \left(1 + \frac{(x/a)^2}{\ell(L)}\right)^{-\frac{\ell(L)+1}{2}} \tag{21}$$

Here, $l(L) = L[1 - C(1)]/[1 + C(1)] - 2$ describes the effective degrees of freedom and $a = \sqrt{12/l(L)}$ is a scaling parameter; $C(1)$ is the lag-1 autocorrelation. In climate data, the typical record length L is between 500 and 2000 months. For l above 100, (21) can be very well approximated by a Gaussian with the same argument x/a. Then the quantiles x_{05} and x_{01} are simply the quantiles of a Gaussian, multiplied by a.

In long-term correlated records characterized by the Hurst exponent α, P can be also approximated by (21) [64], but with different l and a. Here we follow a different route. Using scaling arguments, it has been shown [23] that for L between 500 and 2000, p decays exponentially below $p = 0.1$ and thus can be approximated by

$$p(x;L) = 0.05 \times (1/5)^{\frac{x-x_{05}}{x_{01}-x_{05}}}, \quad p \le 0.1 \tag{22}$$

The quantiles x_{05} and x_{01} depend on the Hurst exponent α and on L. They have the form

$$x_q = C(L,q)\alpha^{\delta(L)}, \quad q = 05 \text{ or } 01. \tag{23}$$

The exponent $\delta(L)$ increases logarithmically with L and ranges from $\delta(500) \cong = 3.22$ to $\delta(2000) \cong 4.06$. The power-law dependence of x_q on α implies

$$x_q(\alpha_2) = x_q(\alpha_1)\left(\frac{\alpha_2}{\alpha_1}\right)^{\delta(L)} \tag{24}$$

Accordingly, the significance of an observed trend x depends strongly on the Hurst exponent α of the record. Let us consider two records of length 2000, with $\alpha_1 = 0.6$ and $\alpha_2 = 0.9$. Since $(0.6/0.9)^{4.06} \approx 0.19$, a trend in record 2 must be 5.2 times larger than in record 1 to be significant. This shows how important it is to analyze properly the correlation features of a record.

In climate science, this fact has been overlooked for a long time, see, e.g., the IPCC report from 2013 [65] or [66]. Until recently it was quite popular among climate scientists to *tacitly assume* (instead of carefully analyzing the data) that climate records are governed by sort-range correlated (red) noise, and to use (21) to determine the p-values simply by measuring $C(1)$. Critical evaluations of this approach can be found in [28, 31]. Tab. 1 taken from [28] compares the results of this procedure with the result based on the proper correlation analysis, for stations on the Antarctic Peninsula and the South Pole. The Antarctic Peninsula is of particular interest since, in contrast to East Antarctica, its temperature increased in the past decades. The table shows that the popular approach (last two columns) clearly overestimates the significance of the trend. When averaged over all stations, the temperature increase of the Antarctic Peninsula was far from significant ($p = 0.172$), while the popular approach suggested a nearly highly significant warming.

8 Conclusions

In this tutorial mini-review dedicated to Jörg Kärger, I have discussed the most important types of noise and shown how important it is to know the correct type of noise when analyzing data sets. For example, data that are long-term correlated like atmospheric or sea surface temperatures, exhibit a pronounced mountain valley structure with mountains and valleys on all length (= time) scales. When one falsely assumes, without having analyzed the data properly, that the structure consists of white (or short-range correlated "red") noise plus external trends, one is led to the wrong conclusion that strong external trends are superimposed to the noise. The same happens with the clustering of rare events, which is not only the consequence of an increasing external trend but is also a natural consequence of long-range correlations. A natural clustering of rare events can be seen in the historic water levels of the Nile River. It can be also observed in the daily volatility of financial assets, where long periods with low volatility can be followed by

Table 1. Persistence and trend characteristics of temperature records from the Antarctic Peninsula and the South Pole (Amundsen-Scott station), ending in 2014: The first 3 columns describe the record length N in months, the trend Δ and the standard deviation σ around the trend line in °C. Columns 4 and 5 show the Hurst exponent α obtained by DFA2 and the proper p-value. The last 2 columns show the lag-1 auto-correlation $C(1)$ and the p-value based on (21). Redrawn from [28].

Station	N [month]	Δ [°C]	σ [°C]	α [–]	p [–]	$C(1)$ [–]	p[–]
Bellingshausen	537	0.83	1.61	0.77	0.352	-0.05	0.030
Vernadsky	752	3.34	2.29	0.83	0.042	0.36	0.0004
Rothera	432	2.39	2.26	0.82	0.146	0.25	0.019
Peninsula average	432	1.82	1.78	0.83	0.172	0.18	0.012
Amundsen-Scott	671	–0.06	2.41	0.51	0.868	–0.01	0.917

long periods with high volatility. In both examples, the long-range correlations in the data drive the clustering of the rare events.

In the climate system, there is a general consensus that in the past 70 years the mean earth temperature has significantly increased by an anthropogenic trend driven by greenhouse gases, and also the advanced methods discussed here confirm the high significance of the trend. But this does not mean that every part of the globe has warmed equally. While the Arctic has warmed up drastically in the past century, in Antarctica a significant warming has not yet been observed. But this does not mean that it will not come; indications for a beginning anthropogenic reduction of the sea ice coverage around Antarctica have already been noticed [67].

Acknowledgment

I have benefited strongly from numerous discussions with Eva Koscielny-Bunde, Shlomo Havlin, H. Eugene Stanley, John Schellnhuber, Jan Kantelhardt, Sabine Lennartz, Mikhail Bogachev, and Josef Ludescher. Open access funding enabled and organized by Projekt DEAL.

References

[1] F. N. David, *Games, Gods, and Gambling: A History of Probability and Statistical Ideas*, Courier Dover Publications, Dover **1998**.
[2] *Leipzig, Einstein, Diffusion* (Ed: J. Kärger), Leipziger Universitätsverlag, Leipzig **2007**.
[3] *Diffusive Spreading in Nature, Technology, and Society* (Eds: A. Bunde, J. Caro, J. Kärger, G. Vogl), 2nd ed., Springer, Berlin **2023**.
[4] D. L. Turcotte, *Fractals and Chaos in Geology and Geophysics 2nd ed.*, Cambridge University Press, Cambridge **1997**.
[5] H. E. Hurst, *Trans. Am. Soc. Civ. Eng.* **1951**, *116*, 770.
[6] B. B. Mandelbrot, J. R. Wallis, *Wat. Resour. Res.* **1968**, *4*, 909.
[7] Y. Tessier, S. Lovejoy, B. Hubert, D. Schertzer, S. Pecknold, *J. Geophys. Res.: Atmos.* **1996**, *101*, 26427.
[8] A. Montanari, R. Rosso, M. S. Taqqu, *Wat. Resour. Res.* **2000**, *36*, 1249.
[9] M. Mudelsee, G. T. Börngen, U. Grünwald, *Nature* **2003**, *425*, 166.
[10] E. Koscielny-Bunde, J. W. Kantelhardt, P. Braun, A. Bunde, S. Havlin, *J. Hydrology* **2006**, *322*, 120.
[11] M. Mudelsee, *Wat. Resour. Res.* **2007**, *43*, W01202.
[12] V. Livina, Y. Ashkenazy, P. Braun, A. Monetti, A. Bunde, S. Havlin, *Phys. Rev. E* **2003**, *67*, 042101.
[13] B. B. Mandelbrot, *Gaussian Self-Affinity and Fractals*, Springer, Berlin **2001**.
[14] E. Koscielny-Bunde, A. Bunde, S. Havlin, H. E. Roman, Y. Goldreich, H.-J. Schellnhuber, *Phys. Rev. Lett.* **1998**, *81*, 729.
[15] R. Weber, P. Talkner, *J. Geophys. Res.* **2001**, *106*, 0131.
[16] R. Monetti, S. Havlin, A. Bunde, *Physica A* **2003**, *320*, 581.
[17] J. Eichner, E. Koscielny-Bunde, A. Bunde, S. Havlin, H. J. Schellnhuber, *Phys. Rev. E* **2003**, *68*, 046133.
[18] K. Fraedrich, R. Blender, *Phys. Rev. Lett.* **2003**, *90*, 108501.
[19] A. Kiŕaly, I. Bartos, I. Ján̈osi, *Tellus A* **2006**, *5*, 593.
[20] D. Rybski, A. Bunde, S. Havlin, H. v. Storch, *Geophys. Res. Lett.* **2006**, *33*, L06718.
[21] E. Giese, I. Mossig, D. Rybski, A. Bunde, *Erdkunde* **2007**, *61*, 86.
[22] D. Rybski, A. Bunde, H. v. Storch, *J. Geophys. Res.: Atmos.* **2008**, *113*, D02106.
[23] S. Lennartz, A. Bunde, *Phys. Rev. E* **2011**, *84*, 021129.
[24] C. L. E. Franzke, *J. Climate* **2012**, *25*, 4172.
[25] S. Lovejoy, D. Schertzer, *The Weather and Climate: Emergent Laws and Multifractal Cascades*, Cambridge University Press, Cambridge **2013**.
[26] A. Bunde, J. Ludescher, C. L. E. Franzke, U. Büntgen, *Nat. Geosci.* **2014**, *7*, 246.
[27] N. Yuan, Z. Fu, S. Liu, *Sci. Rep.* **2014**, *4*, 6577.
[28] J. Ludescher, A. Bunde, C. L. E. Franzke, H. J. Schellnhuber, *Clim. Dyn.* **2016**, *46*, 263.
[29] S. Blesić, D. Zanchettin, A. Rubino, *J. Climate* **2019**, *32*, 349.
[30] C. L. E. Franzke et al., *Reviews of Geophysics* **2020**, *58*, e2019RG000657. DOI: https://doi.org/10.1029/2019RG000657
[31] A. Bunde, J. Ludescher, H. J. Schellnhuber, *Clim. Dyn.* **2022**, *58*, 1349.
[32] A. Beretta, H. E. Roman, F. Raicich, F. Crisciani, *Physica A* **2005**, *347*, 695.
[33] S. Dangendorf, D. Rybski, C. Mudersbach, A. Müller, E. Kaufmann, E. Zorita, J. Jensen, *Geophys. Res. Lett.* **2014**, *41*, 5530.
[34] M. Becker, M. Karpytchev, S. Lennartz-Sassinek, *Geophys. Res. Lett.* **2014**, *41*, 5571.
[35] M. Santhanam, H. Kantz, *Physica A* **2005**, *345*, 713.
[36] R. Blender, C. Raible, F. Lunkeit, *Q. J. R. Meteorolog. Soc.* **2015**, *141*, 249.
[37] S. Lennartz, A. Bunde, D. L. Turcotte, *Geophys. J. Int.* **2011**, *184*, 1214.
[38] C.-K. Peng et al., *Nature* **1992**, *356*, 168.
[39] A. Arneodo, E. Bacry, P. V. Graves, J. F. Muzy, *Phys. Rev. Lett.* **1995**, *74*, 3293.
[40] P. C. Ivanov, M. G. Rosenblum, C. K. Peng, J. Mietus, S. Havlin, H. E. Stanley, *Nature* **1996**, *383*, 323.
[41] A. Bunde, S. Havlin, J. Kantelhardt, T. Penzel, J. H. Peter, K. Voigt, *Phys. Rev. Lett* **2000**, *85*, 3736.
[42] J. M. Hausdorff, Y. Ashkenazy, C. K. Peng, P. C. Ivanov, H. E. Stanley, *Physica A* **2001**, *302*, 138.
[43] B. B. Mandelbrot, *J. Polit. Econ.* **1963**, *71*, 420.
[44] R. N. Mantegna, H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge University Press, Cambridge **1999**.
[45] W. Ebeling, A. Neiman, *Physica A* **1995**, *215*, 233.
[46] K. Tanaka-Ishii, A. Bunde, *PlosOne* **2016**, *11*, e0164658. DOI: https://doi.org/10.1371/journal.pone.0164658
[47] S. Lennartz, A. Bunde, *Phys. Rev. E* **2009**, *79*, 066101.
[48] K. Funke, *Prog. Solid State Chem.* **1993**, *2*, 111.
[49] J. Kantelhardt, E. Koscielny-Bunde, H. Rego, S. Havlin, A. Bunde, *Physica A* **2001**, *295*, 441.
[50] J. Kantelhardt, S. Zschiegner, E. Koscielny-Bunde, A. Bunde, S. Havlin, H. E. Stanley, *Physica A* **2002**, *316*, 87.
[51] H. O. Peitgen, H. Jürgens, D. Saupe, *Chaos and Fractals*, Springer, Berlin **1992**.
[52] *The science of disasters – climate disruptions, heart attacks, and market crashes* (Eds: A. Bunde, J. Kropp, H. J. Schellnhuber), Springer, Berlin **2002**.
[53] R. Govindan, D. Vjushin, S. Brenner, A. Bunde, S. Havlin, H.-J. Schellnhuber, *Phys. Rev. Lett.* **2002**, *89*, 028501.
[54] D. Vjushin, I. Zhidkov, S. Brenner, S. Havlin, A. Bunde, *Geophys. Res. Lett.* **2004**, *31*, L10206.
[55] J. Ludescher, A. Bunde, U. Büntgen, H. J. Schellnhuber, *Clim. Dyn.* **2021**, *55*, 3017.

[56] A. Bunde, J. Eichner, S. Havlin, J. Kantelhardt, *Phys. Rev. Lett.* **2005**, *94*, 048701.
[57] C. Pfisterer, *Wetternachhersage: 500 Jahre Klimavariationen und Naturkatastrophen 1496-1995*, Verlag Paul Haupt, Bern **1998**.
[58] R. Glaser, *Klimageschichte Mitteleuropas*, Wissenschaftliche Buchgesellschaft, Darmstadt **2001**.
[59] K. L. Ngai, C. Cramer, T. Saatkamp, K. Funke, in *Non-Equilibrium Phenomena in Supercooled Fluids, Glasses and Amorphous Materials* (Eds: M. Giorano, D. Leporini, M. P. Tosi), World Scientific, Singapore **1996**.
[60] G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*, Hafner Publishing Company, New York **1966**.
[61] G. Herdan, *Quantitative Linguistics*, Butterworths, Washington **1964**.
[62] I. Bronstein et al., *Handbook of Mathematics*, Springer, Berlin **2004**.
[63] B. D. Santer et al., *J. Geophys. Res.: Atmos.* **2000**, *105*, 7337.
[64] A. Tamazian, J. Ludescher, A. Bunde, *Phys. Rev. E* **2015**, *91*, 032806.
[65] *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Eds: T. F. Stocker et al.), Cambridge University Press, Cambridge **2013**.
[66] J. Turner et al., *Int. J. Climatol.* **2019**, *40*, 2986.
[67] J. Ludescher, N. Yuan, A. Bunde, *Clim. Dyn.* **2019**, *53*, 237.

