Statistical analysis of hydroclimatic time series: Uncertainty and insights

Demetris Koutsoyiannis[1] and Alberto Montanari[2]

Received 4 October 2006; revised 19 December 2006; accepted 2 January 2007; published 22 May 2007.

[1] Today, hydrologic research and modeling depends largely on climatological inputs, whose physical and statistical behavior are the subject of many debates in the scientific community. A relevant ongoing discussion is focused on long-term persistence (LTP), a natural behavior identified in several studies of instrumental and proxy hydroclimatic time series, which, nevertheless, is neglected in some climatological studies. LTP may reflect a long-term variability of several factors and thus can support a more complete physical understanding and uncertainty characterization of climate. The implications of LTP in hydroclimatic research, especially in statistical questions and problems, may be substantial but appear to be not fully understood or recognized. To offer insights on these implications, we demonstrate by using analytical methods that the characteristics of temperature series, which appear to be compatible with the LTP hypothesis, imply a dramatic increase of uncertainty in statistical estimation and reduction of significance in statistical testing, in comparison with classical statistics. Therefore we maintain that statistical analysis in hydroclimatic research should be revisited in order not to derive misleading results and simultaneously that merely statistical arguments do not suffice to verify or falsify the LTP (or another) climatic hypothesis.

Citation: Koutsoyiannis, D., and A. Montanari (2007), Statistical analysis of hydroclimatic time series: Uncertainty and insights, *Water Resour. Res.*, *43*, W05429, doi:10.1029/2006WR005592.

> Even if we would know everything, we should still have to derive statistical information from this knowledge in order to answer what are essentially statistical problems, such as explaining gas pressure or the intensity of spectral lines.
>
> Karl Popper to Albert Einstein

1. Introduction

[2] In the last decades, the hydrologic and water resources community goes behind the trails of the climatological community in an attempt to trace the future of water resources under climate change. Climate research uses two main routes, numerical models (known as general circulation models, GCMs) for future projections and paleoclimatology for past reconstructions. By definition of climate (as an average behavior of a long time), statistics play an important role in climatology. Statistics is a key tool in both routes. Even in GCMs, the manipulation of data inputs, the assessment of performance of models and the estimation of the uncertainty of projections are all statistical problems. A fortiori, paleoclimatology, which compiles and interprets numerous data series, is totally dependent on statistics. However, it has been recently argued [*Wegman et al.*, 2006] that the paleoclimatic community, even though relies heavily on statistical methods, does not seem to be interacting much with the statistical community. As a result, certain statistical methods are in some cases misused in climatology [see also *von Storch*, 1995; *von Storch and Zwiers*, 1999]. This may be a sign to hydrologic and water resources community to carefully evaluate the assumptions behind climatologists' results.

[3] Dominant doctrines behind paleoclimatic methodologies, which may influence the validity of results, are the (Manichean) dichotomy of natural time series into deterministic and random components (“signal” and “noise”), and the (procrustean) suppression of low-frequency fluctuations of time series so that they comply with an ab initio postulate of a Markovian behavior [see also *Wegman et al.*, 2006]. The dichotomy “signal” versus “noise” has been borrowed from electronics, where indeed is meaningful, but lacks meaning in geophysics (unless noise is used to describe errors, either in measurements or in models). All natural processes are nature's signals, not noises even when they “look like” noise. To describe these signals one may use either a deterministic or a stochastic approach but this is totally different from admitting that natural processes consist of two types of components. Obviously, a stochastic approach can incorporate any known deterministic dynamics (cf. the modeling of periodicity by cyclostationary stochastic models) but again this should not be confused as separation of components [*Koutsoyiannis*, 2003, 2006]. Such a separation, unfortunately very commonly performed, entails risk of misrepresentation of low-frequency natural fluctuations as “deterministic trends” [e.g., *Smith*, 1993; *Craigmile et al.*, 2004a, 2004b; *Koutsoyiannis*, 2006].

[1]Department of Water Resources, National Technical University of Athens, Zographou, Greece.
[2]Dipartimento di Ingegneria delle Strutture, dei Trasporti, delle Acque, del Rilevamento del Territorio, Faculty of Engineering, University of Bologna, Bologna, Italy.

0043-1397/07/2006WR005592$09.00

[4] On the other hand, the Markovian postulate (dating back to 1960s [*Gilman et al.*, 1963] and still in use today [e.g., *Mann and Lees*, 1996]) can be paralleled, in our opinion, with the pre-Keplerian fallacy that the astral bodies should follow cyclical orbits and that any deviation from the cycle should be modeled by another cycle, the epicycle. As climatic records do not verify a Markovian behavior, its adoption has been combined with a decomposition of a climatic series into components, one of which is Markovian (e.g., *Mann and Lees* [1996, equation [6]] perform such a decomposition on stochastic grounds, by spectral methods, whose physical fundament may be disputable).

[5] The Markovian dependence (also known as autoregressive of order 1, or AR(1)) is the most typical and simple example of the so-called short-term persistence (STP, also known as short-term dependence). STP is contrasted with long-term persistence (LTP, also known as Hurst phenomenon, Joseph effect, long memory, long-range dependence, scaling behavior, and multiscale fluctuation). From a practical point of view, LTP indicates that the process is compatible with the presence of fluctuations on a range of timescales, which may reflect the long-term variability of several factors such as solar forcing, volcanic activity and so forth. LTP can be also conceptualized as a tendency of clustering in time of similar events (droughts, floods, etc.).

[6] In statistical terms, the presence in a time series of long-term fluctuations implies dramatically increased uncertainty, especially on long timescales, in comparison to classical statistics. This is easy to understand, as the observed record could be a small portion of a longer cycle whose characteristics might be difficult to infer on the basis of the available observations. In this respect, in processes characterized by LTP, the results of the statistical analysis may be difficult to decipher. As a consequence, the application of statistical tools to climatic time series should be carefully considered and classical statistics should be carefully revisited to locate points that may produce misleading or incorrect results.

[7] In stochastic terms, STP and LTP are conceptualized in terms of conditional probabilities for the future given past observations. Thus, in a Markovian process the future is not influenced by the past when the present (a time instant) is known whereas in a process exhibiting LTP the influence of the past (the entire history) never ceases. Both Markovian dependence and LTP can result from physical principles. For example, the maximum entropy principle results in Markovian dependence if the maximization of entropy is done on a particular timescale and in LTP if the maximization is done on a range of timescales [*Koutsoyiannis*, 2005b]. Despite dominance of the Markovian behavior in climatologists' views, its two aforementioned features (non influence of the past, exclusiveness of a single scale of fluctuation) and other features discussed below might make it implausible, in our opinion.

[8] Probability, statistics and stochastic processes provide mathematical tools to describe LTP conveniently and efficiently. To fight a common misconception, it should be stressed that the use of such tools should not be confused with admitting that things happen spontaneously and randomly or without a cause. It is well known today (from the chaos literature) that even a simple nonlinear system with purely deterministic dynamics may trace an irregular trajectory, whose future may be unpredictable in deterministic sense. Unpredictability or future uncertainty depends then on the degree of nonlinearity and the dimensionality (number of degrees of freedom) of the deterministic system as well as the time horizon of prediction. For chaotic systems, the deterministic dynamics cannot produce a good prediction for a large time horizon. This is particularly the case in high-dimensional systems, where a stochastic approach may give better results (mean predictions and uncertainty limits) than a pure deterministic approach. This is reflected for instance in the recently developed method of ensemble weather predictions, a method based on the idea of Monte Carlo (i.e., stochastic) simulation, whose use has now been generalized in meteorological services.

[9] An example of this type, more closely related to the subject of this study, has been proposed by *Koutsoyiannis* [2006]. This example deals with a simple toy model that was devised to mimic the evolution of long hydroclimatic time series. The model is purely deterministic (involving no random component) and nonlinear, and has only two degrees of freedom. Application of the model demonstrates that (1) extremely simple deterministic dynamics can produce trajectories exhibiting LTP, (2) large-scale climatic fluctuations (like upward or downward trends) can emerge without any apparent reason, and (3) deterministic dynamics do not help predict climatic evolution, even in the case of the caricature model with only two degrees of freedom. Thus this demonstration justifies (1) the utility of a stochastic description even for systems with perfectly known purely deterministic dynamics and (2) the presence of LTP in all examined hydroclimatic series.

[10] To date there is considerable empirical evidence for the presence of LTP in hydroclimatic and other geophysical records, as well as time series from other fields. In fact, the history of LTP started more than half a century ago, after its discovery in geophysics by *Hurst* [1951], although, in a mathematical (stochastic processes) and physical context (turbulence) the concept has been pioneered a decade earlier by *Kolmogorov* [1940; see also *Shiryaev*, 1989].

[11] Throughout these decades the studies providing indications that LTP may be omnipresent in several natural (hydroclimatic, geophysical, biological) and human-associated (social, economical and technological) processes have been so numerous that it is difficult to shape a complete picture; yet it is worth giving some recent examples (which contain additional references). LTP properties of temperature (which is the focus of this paper) at a point, regional or global basis, have been studied by *Bloomfield* [1992], *Koscielny-Bunde et al.* [1996, 1998], *Pelletier and Turcotte* [1997], *Koutsoyiannis* [2003], *Monetti et al.* [2003], and *Koutsoyiannis et al.* [2007]. Similar analyses have been conducted for other climatological time series including wind power and indexes of North Atlantic Oscillation [*Haslett and Raftery*, 1989; *Stephenson et al.*, 2000] as well as proxy series such as tree ring widths or isotope concentrations [*Pelletier and Turcotte*, 1997; *Koutsoyiannis*, 2002; *Beran and Feng*, 2002; *Craigmile et al.*, 2004b]. Numerous studies have indicated LTP in hydrological time series and particularly in river flows [*Eltahir*, 1996; *Montanari et al.*, 1997; *Pelletier and Turcotte*, 1997; *Koutsoyiannis*, 2002, 2003; *Radziejewski and*

Kundzewicz, 1997; *Montanari*, 2003; *Sakalauskienë*, 2003; *Yue and Gan*, 2004; *Koscielny-Bunde et al.*, 2006]. Similar findings have been reported in diverse scientific fields such as biology [*Peng et al.*, 1994], ecology [*Halley and Inchausti*, 2004], physiology [*Hausdorff et al.*, 1997], psychology [*Wagenmakers et al.*, 2004], economics [*Ray and Tsay*, 2000], politics [*Byers et al.*, 2000], and Internet computing [*Karagiannis et al.*, 2004]. The similarity of behaviors in such diverse complex systems should not be regarded as coincidence; rather some fundamental explanation behind this should be investigated, as is for instance the central limit theorem (CLT) for the emerging of the normal distribution in diverse processes. Perhaps this explanation is the principle of maximum entropy, which also produces the normal distribution independently of CLT [*Koutsoyiannis*, 2005a, 2005b].

[12] Most recently, the presence of LTP in temperature data has been considered by *Cohn and Lins* [2005] and *Rybski et al.* [2006]. Both have found that instrumental records and reconstructed time series of temperature are compatible with the hypothesis of LTP and therefore suggested that this property should be taken into account in statistical tests. Earlier, *Koutsoyiannis* [2003] arrived at similar conclusions, arguing that there is the need in hydroclimatic research to adapt classical statistics, which is based on the independent identically distributed (IID) paradigm, so as to account for the observed LTP behavior. Also a variety of methods shed light to the statistical problems related to LTP [*Beran and Feng*, 2002; *Kantelhardt et al.*, 2002; *Craigmile et al.*, 2004a, 2004b].

[13] In this respect, *Cohn and Lins* [2005] and *Rybski et al.* [2006] have suggested a necessary rectification of the prevailing incorrect practices. Both have proposed adapted statistical tests, which they have illustrated essentially on the same climatic record, the instrumental temperature record of the Northern Hemisphere between 1856 and 2004 (due to Climatic Research Unit (CRU)). The LTP and trend properties of this record had been studied earlier by *Smith* [1993], *Beran and Feng* [2002], and *Craigmile et al.* [2004a]. *Cohn and Lins* [2005] and *Rybski et al.* [2006] focused on the well-known detection problem (whether or not climatic changes have occurred) and attribution problem (whether or not observed changes are related to anthropogenic forcings of the climate system). Interestingly, however, their conclusions on these problems are opposite. *Rybski et al.* [2006] conclude that the hypothesis that at least part of the recent warming cannot be solely related to natural factors can be accepted with a very low risk. *Cohn and Lins* [2005] state that given what we know about the complexity, long-term persistence, and nonlinearity of the climate system, this warming can be due to natural dynamics. This disagreement may indicate, in our opinion, that our understanding of the behavior of LTP and its consequences in climatic analyses and statistical testing is not complete yet and that additional insights are needed.

[14] Such insights are sought in this study using simple analytical methods, rather than complicated numerical methods. The justification for this choice is that analytical methods are more insightful (albeit less accurate for reasons that we will discuss) than numerical ones. As an empirical basis we use the same basic data set as in the two aforementioned recent studies, the CRU record (now extended up to 2005; http://www.cru.uea.ac.uk/ftpdata/tavenh2v.dat) and, as auxiliary information, the six recently reconstructed temperature records of the Northern Hemisphere analyzed by *Rybski et al.* [2006]. These are abbreviated here as J98, M99, B00, E02, M03, and M05 and refer to records of *Jones et al.* [1998], *Mann et al.* [1999], *Briffa* [2000], *Esper et al.* [2002], *McIntyre and McKitrick* [2003], and *Moberg et al.* [2005], respectively; note that M03 was not proposed as a reconstruction but only as a modification of M99 to illustrate lack of robustness of methodology. These series have annual resolution (time step) and therefore their statistical analysis cannot (and need not) describe the effects of seasonality. The LTP properties of some of these and some other proxy series have been also studied in other works recently (D. Stockwell, Scale invariance for dummies, http://landshape.org/enm/?p = 13) and earlier (see *Koutsoyiannis* [2003] for J98).

[15] Our focus is on providing insight on uncertainty rather than on proposing accurate statistical tests. In this respect, our study of the detection/attribution problem is carried out on a conceptual basis and therefore we avoid proposing categorical results. In addition, we try to locate potential pitfalls, which may appear if this uncertainty is not explicitly considered and may have also influenced previous studies.

[16] The uncertainty is studied under both STP and LTP hypotheses, also in comparison to the IID case, but the emphasis is given to the LTP case. It is not our target to prove or disprove the LTP hypothesis here; in contrast, we demonstrate below that (because of high uncertainty) such a target cannot be achieved by merely statistical arguments. However, by summarizing the above discussion, we believe that several indications have been already accumulated (see the references cited above) that make the LTP hypothesis very plausible in contrast to the implausibility of common alternative hypotheses such as IID (usually implicit in most statistical analyses of hydroclimatic processes albeit not explicitly admitted) and Markovian (claimed by some climatological studies).

2. Detecting the Presence and Intensity of Long-Term Persistence

[17] Since *Hurst* [1951] discovered LTP, several formalisms and conceptualizations have been used to study it, on which the algorithms to detect this behavior are based [*Taqqu et al.*, 1995; *Montanari et al.*, 1997; *Kantelhardt et al.*, 2002]. Among these, the most popular are the original formalism by *Hurst* [1951], based on the so-called rescaled range statistic (R/S) and the detrended fluctuation analysis (DFA) introduced by *Peng et al.* [1994; see also *Vjushin et al.*, 2001]. However, we choose to use in our analysis of climatic series the formalism based on the aggregated standard deviation (ASD). The latter has several advantages such as (1) easy understandability and transparency that enables better perception of the behavior and does not hide its implications, (2) simplicity and minimal parameterization (it does not involve any other concept than standard deviation), which enables a probabilistic description of the concepts it uses and hence a statistical framework of estimation and testing, and (3) appropriateness, in terms of producing estimates within the interval (0, 1). The method is based on the analysis of the variability of the data aggregated at different timescales. Specifically, let X_i be a stationary

process on discrete time i (referring to years in our case) with (true, or population) standard deviation σ and let

$$X_i^{(k)} := (X_i + \ldots + X_{i-k+1})/k \tag{1}$$

denote the aggregate (average) process at timescale k, with (true) standard deviation $\sigma^{(k)}$ (the notation implies that $X_i^{(1)} \equiv X_i$). For sufficiently large k, $X_i^{(k)}$ represents the climatic process; typically, the convention $k = 30$ is used to standardize the climatic timescale (number of years). Now, LTP is expressed by the elementary scaling property

$$\sigma^{(k)} = \frac{\sigma}{k^{1-H}} \tag{2}$$

where H is the Hurst exponent, which for stationary and positively correlated processes varies in the range (0.5, 1). $H = 0.5$ means independence and increasing values represent increasing LTP intensities. The reader interested to further details about the range of values for H is referenced to *Mandelbrot and van Ness* [1968]. It is worth pointing out that the preliminary assumption of stationarity for X_i is necessary in a statistical testing framework for climatic change (see section 5) because the null hypothesis to be tested (based on a given climatic record) is precisely the stationarity of the process.

[18] The simple equation (2) can support (1) a definition of LTP, (2) a definition of a stochastic process having this property, that is the simple scaling stochastic (SSS) process (also known as stationary intervals of a self similar process), and (3) the estimation of H using sample estimates of $\sigma^{(k)}$ at several scales k. Equation (2) implies that the autocorrelation $\rho_j^{(k)}$ for scale k and lag j (defined as $\rho_j^{(k)} := \mathrm{Cov}[X_i^{(k)}, X_{i+kj}^{(k)}]/\mathrm{Var}[X_i^{(k)}]$) is independent of scale [e.g., *Koutsoyiannis*, 2002]:

$$\rho_j^{(k)} = \rho_j = (1/2)\left(|j+1|^{2H} + |j-1|^{2H}\right) - |j|^{2H} \tag{3}$$

[19] LTP is more precisely defined as an asymptotic property for large scales, in which case (2) should be replaced by $\sigma^{(k)} = \sigma^{(l)}/(k/l)^{1-H}$ for $k/l > 1$ and $l \to \infty$; also SSS is more precisely defined in terms of scaling properties of the distribution function. It is important to note that, even though the same equation (2) can serve as a basis for the definition of the LTP as well as the SSS process, these two are totally different notions: LTP is a behavior that can be investigated in any type of time series, such as series of observations of a natural process, output of a deterministic model, or synthetic series generated by a stochastic process. In contrast, SSS is a stochastic process.

[20] For comparison, in the case of the simplest STP model, which is the AR(1), (2) and (3) become respectively [*Koutsoyiannis*, 2002]:

$$\sigma^{(k)} = \frac{\sigma}{\sqrt{k}}\sqrt{\frac{(1-\rho^2) - 2\rho(1-\rho^k)/k}{(1-\rho)^2}} \tag{4}$$

$$\rho_1^{(k)} = \frac{\rho\left(1-\rho^k\right)^2}{k(1-\rho^2) - 2\rho(1-\rho^k)}, \quad \rho_j^{(k)} = \rho_1^{(k)}\rho^{k(j-1)},\ j \geq 1 \tag{5}$$

where $\rho \equiv \rho_1^{(1)}$. These indicate that (1) for large k, $\sigma^{(k)} \sim \sigma/\sqrt{k}$; (2) $\rho_j^{(k)}$ is a decreasing function of scale k, and (3) only at scale $k = 1$ (annual) is the process Markovian (i.e., $\rho_j = \rho^j$); at all other scales the autocorrelation structure in (5) (i.e., $\rho_j^{(k)} = \rho_1^{(k)}\ (\rho^k)^{j-1}$) is identical to that of an autoregressive moving average (ARMA) process of order (1, 1), another classical example of STP. Note that both AR(1) and SSS involve a single parameter each and that the equations (2) and (3) of SSS are simpler than (4) and (5) of AR(1), even though the former has been regarded by many as very complicated.

[21] Obviously, the different formalisms to LTP imply different estimates of H. This is demonstrated in Table 1 for the seven time series and for three formalisms: the DFA as derived by *Rybski et al.* [2006], the R/S and the ASD. In the latter we used an algorithm by *Koutsoyiannis* [2003], which by construction ensures appropriate estimates ($0 < H < 1$). Generally, all methods result in very high but different H values.

3. Statistical Uncertainty

[22] Given a sample $X_1, \ldots, X_n$ of size n and observations $x_1, \ldots, x_n$, clearly $X_1^{(n)}$ is the standard estimator of the mean μ of the process (most typically denoted as $\overline{X}$) and $x_1^{(n)}$ is the estimate of μ. The standard deviation $\mathrm{StD}[\overline{X}] \equiv \mathrm{StD}[X_1^{(n)}]$ is a convenient indicator of uncertainty, and according to the scaling property (2), $\mathrm{StD}[\overline{X}] = \sigma^{(n)} = \sigma/n^{1-H}$. (Here $\mathrm{StD}[.] := \sqrt{\mathrm{Var}[\cdot]}$ denotes the standard deviation of a random variable). If we compare it to the classical statistical law $\mathrm{StD}[\overline{X}] = \sigma/\sqrt{k}$ (also valid asymptotically for STP processes as shown in (4)), the differences are dramatic as H grows away from 0.5. To demonstrate it, for a series of length n with LTP we can calculate the "equivalent" (or "effective") sample size n' in the classical statistics (IID) sense, so that $\sigma/n^{1-H} = \sigma/n'^{0.5}$. Clearly,

$$n' = n^{2(1-H)} \tag{6}$$

[23] As shown in Table 1, the equivalent sample sizes resulting by this equation for the seven time series are as low as 2–5. For instance in the SSS sense, the longest sample size (1979), is equivalent to a classical statistical sample size of about 3. Thus a record with length of 1979 years, which certainly would be called a long record having in mind classical statistics, is a very short record in the SSS framework. Only this example suffices to demonstrate that the Hurst behavior has astonishing effects in the foundation of climatology and hydrologic statistics, provided that the LTP hypothesis is true.

[24] Even the AR(1) model implies reduction of sample size; in this case using (4) and a similar logic, we obtain that

$$n' = n\frac{(1-\rho)^2}{(1-\rho^2) - 2\rho(1-\rho^n)/n} \tag{7}$$

[25] Values estimated from (7) are also given in Table 1 and show that the reduction is not as dramatic as in the SSS case. However, it is noteworthy that in the CRU data set the effective sample size reduces to 14.

[26] However, the implications are perhaps even worse than described above, because the analysis was based on the

Table 1. Comparisons of Estimates of Statistics for Different Methods and Data Sets

Data Series	CRU	J98	M99	B00	E02	M03	M05
			All Data				
Sample size	150	992	981	994	1162	581	1979
s, standard estimate	0.27	0.23	0.13	0.14	0.14	0.17	0.22
H by DFA[a]	1.09	0.82	0.97	0.93	1.04	0.83	0.86
H by R/S	1.07	0.90	0.89	0.89	0.93	0.97	0.92
H by ASD	0.93	0.88	0.91	0.91	0.94	0.92	0.94
ρ	0.84	0.53	0.65	0.64	0.81	0.66	0.91
Equivalent sample size							
SSS	1.9	5.0	3.4	3.3	2.5	2.8	2.7
AR(1)	13.8	307.5	205.0	221.1	120.8	119.3	95.3
			Period 1400–1855				
Sample size		456	456	456	456	456	456
s, standard estimate		0.20	0.10	0.13	0.09	0.16	0.21
H by ASD		0.86	0.88	0.91	0.93	0.92	0.93
ρ		0.54	0.62	0.59	0.77	0.65	0.88

[a]Values from *Rybski et al.* [2006], except in the CRU series, which was estimated in this study.

assumption that H is known a priori. In reality, H is estimated from the data, so there is additional sampling uncertainty (statistical estimation error). The sampling uncertainty applies also to all other statistics and we can anticipate that all confidence bands are wider than in classical statistics, as will be discussed below. In addition, because LTP is eventually an asymptotical property of the process (which should be detected on the tail, i.e., on the largest scales), even the detection of LTP is highly uncertain when dealing with time series with short length [*Taqqu et al.*, 1995].

[27] This point has already been made in some studies. For example, *Koutsoyiannis* [2002] showed that the sum of three Markovian processes (whose behavior, rigorously speaking is STP) is virtually indistinguishable from a process with LTP for lags as high as of the order of 1000. To demonstrate this point further, we fitted to the E02 series an ARMA(1, 1) process. Testing the autocorrelation function of the residuals of this, we concluded that they are indistinguishable from white noise; this means that the series is compatible with the ARMA(1, 1) process, i.e., it exhibits STP with Hurst coefficient 0.50. Furthermore, we generated with the fitted ARMA(1, 1) a synthetic series with sample size 2000, and all estimation methods we tried gave incorrect values of H on the order 0.79–0.93. Continuing this experiment, we also found that we need a series with length of about 20 000 to correctly estimate H, namely, to find a value around 0.50. These examples clearly point out that even the distinction between the extreme cases $H = 0.5$ and $H \to 1$ is not statistically decidable with typical sample sizes.

4. Observation Uncertainty

[28] It is well known that observations of hydrometeorological processes involve several inaccuracies; even in the instrumental CRU series, some observation uncertainty exists, mainly because of spatial integration of point measurements whose number and locations differ through history. However, in the case of proxy data, there is an extra source of high uncertainty because the data are not instrumental. In fact, all six proxy series considered here are supposed to represent roughly the same process, that is, the evolution of the Northern Hemisphere temperature (even though, from their construction, which joins different data sources with varying time span, seasonal representativeness, and methodological assumptions, one may have doubts about what they exactly represent [see also *Wegman et al.*, 2006]). The different values assigned for the same year in the different series manifest none other than the uncertainty in reconstruction of the past climate. This is well known and is related to the subjectivity of dendroclimatology on which the given proxy series are primarily based. The subjectivity originates from sampling procedures (e.g., in picking and choosing which samples to use) and from the differing statistical calibration approaches. Recall, for instance, that M03 and M99 are based on the same original data. The differences in seasonal and spatial representativeness of the various reconstructions is an additional source of uncertainty. For most series, uncertainty bands are also given but as pointed out by the *Board on Atmospheric Sciences and Climate* [2006, p. 113] they have been underestimated. For additional discussions, see *Esper et al.* [2003], *McIntyre and McKitrick* [2003], *Jones and Mann* [2004], *von Storch et al.* [2004], *Zorita and von Storch* [2005], S. McIntyre (A quote from Esper et al., http://www.climateaudit.org/?p = 365), and D. Stockwell (A new temperature reconstruction, http://landshape.org/enm/?p = 15).

[29] An interesting piece of information conveyed by all proxies is the compatibility of all of them with the LTP hypothesis, even if we do not include in the analyses the years of instrumental observations (which one may argue that are affected by global warming). To make this clearer, we redid the LTP analysis for the period 1400–1855, which is the common period of all proxy series prior to the period of instrumental records. The results, shown in Table 1, indicate that the H values obtained for this period are virtually identical to those for the complete data set and close to each other, averaging to 0.91, a value close to that of CRU (0.93). On the other hand, the standard deviations on the annual scale, even though they do not depart significantly from the values of the whole period of each sample, are very different to each other (ranging in 0.09°–0.21°C versus 0.27°C of the CRU series).

[30] It is interesting to compare the above range of values with the sampling uncertainty of the standard deviation of the CRU series. Combining known results [*Matalas*, 1967; *Salas*, 1993, p. 19.11; *Beran*, 1994, p. 156; *Koutsoyiannis*, 2003], it is observed that, when there is temporal dependence in the process of interest, the standard estimator S of the standard deviation σ is not unbiased and that an approximately unbiased estimator for both the LTP and STP cases is

$$\stackrel{\approx}{S} := \sqrt{\frac{n'}{n'-1}} S \tag{8}$$

[31] This assumes that n (the actual sample size) is large enough; for a more accurate expression for small n see *Koutsoyiannis* [2003]. Notice the dependence of $\stackrel{\approx}{S}$ on the effective sample size n' rather than the n and that for small n' the correction factor (the square root in (8)) can be much larger than 1. Thus, in the SSS case the estimate $\stackrel{\approx}{s}$ may differ dramatically from the standard estimate s (notice the notational convenience of lower case letters for estimates, i.e., numerical values, and upper case ones for estimators, i.e., random variables). Also, combining results from *Koutsoyiannis* [2003] (based on systematic Monte Carlo simulations) and using $\stackrel{\approx}{s}$ as an estimate of the true standard deviation σ, it can be obtained that in the SSS case,

$$\frac{\mathrm{StD}\left[\stackrel{\approx}{S}\right]}{\stackrel{\approx}{s}} = \frac{\mathrm{StD}[S]}{s} \approx \sqrt{\frac{(0.1n+0.8)^{\lambda(H)}}{2(n-1)}}, \quad \text{with } \lambda(H) := 0.088(4H^2-1)^2 \tag{9}$$

[32] Now using the statistics of the CRU series, it is computed that the estimate of StD[S] is 0.033°C (versus 0.015°C in classical statistics). Loosely speaking, this justifies a difference in standard deviation between the different series of about 0.08°C (at significance 1%; even though the distribution of StD[S] is not normal). Consequently, from the values in Table 1, we can conclude that the variability of the J98 and M05 series is compatible with the variability of the CRU record. The same result does not apply to all other series. Thus, if one accepts one of the other four series as representative of the past climate, one can readily conclude that the observed temperature variation in the last years is not a result of natural dynamics. In other words, there is a statistical significance in the change of standard deviation, so no additional statistical test is needed. Furthermore, with simple statistical calculations with the standard deviation estimates shown in Table 1, we can easily classify the proxies in two groups (one is J98, M03, M05 and the other one M99, B00, E02), each of which contains series compatible to each other but the two groups are incompatible to each other. This makes unrealistic the possibility to use all series simultaneously in a global statistical approach and highlights once again the uncertainty involved in the use of proxy series.

5. Statistical Testing for Climatic Change

[33] The above findings highlight the potential lack of reliability of statistical tests performed on climatic records, especially proxy ones. Some of these critical behaviors are not known and not immediately evident. It is interesting to inspect with deeper detail the potential effects on the statistical detection of climatic change.

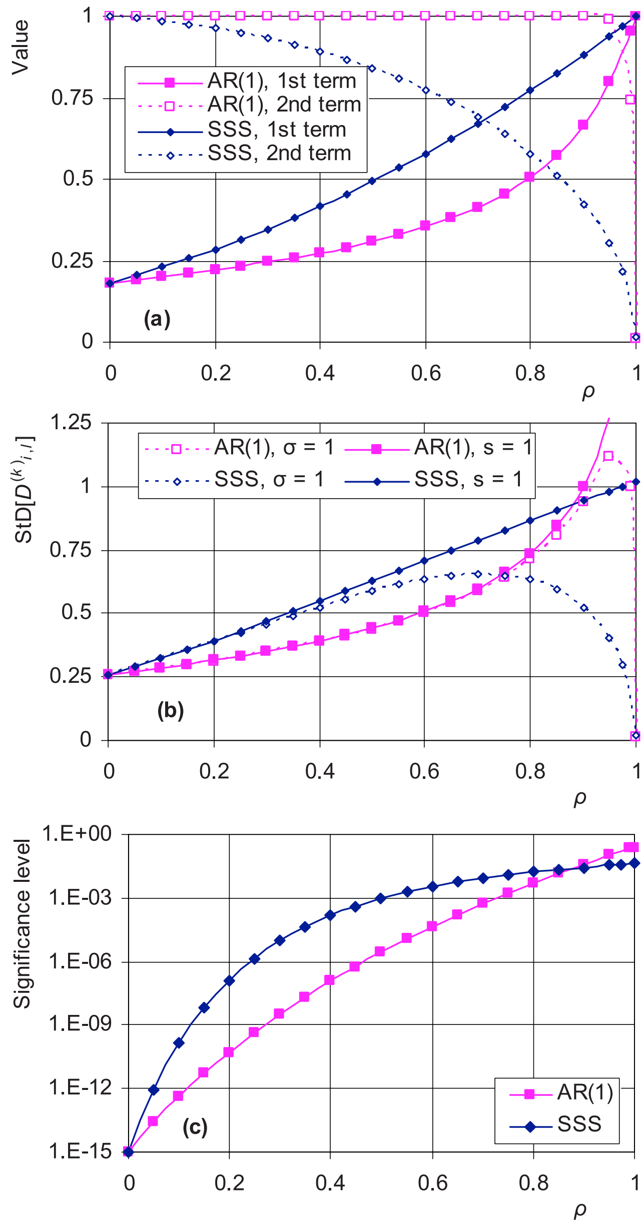

Figure 1. Variation with ρ of (a) the two multiplicative terms comprising StD[$D_{i,l}^{(k)}$] assuming σ = 1, (b) StD[$D_{i,l}^{(k)}$] assuming σ = 1 or s = 1 as indicated, and (c) the implied significance in rejecting the null hypothesis assuming that s = 1 and that in classical IID statistics the significance level is 10^{-15}. Common assumptions are k = 30, l/k = 3, and n = 150.

[34] *Cohn and Lins* [2005] used as a test statistic the slope of a linear fit to the time series to test whether or not a climate variable has changed in a statistically significant sense, over the available observation period. *Rybski et al.* [2006] proposed essentially the statistic $D_{i,l}^{(k)} := X_i^{(k)} - X_{i-l}^{(k)}$ to test whether a or not a climate variable, defined on a timescale k, has changed in a statistically significant sense, over a period of l years

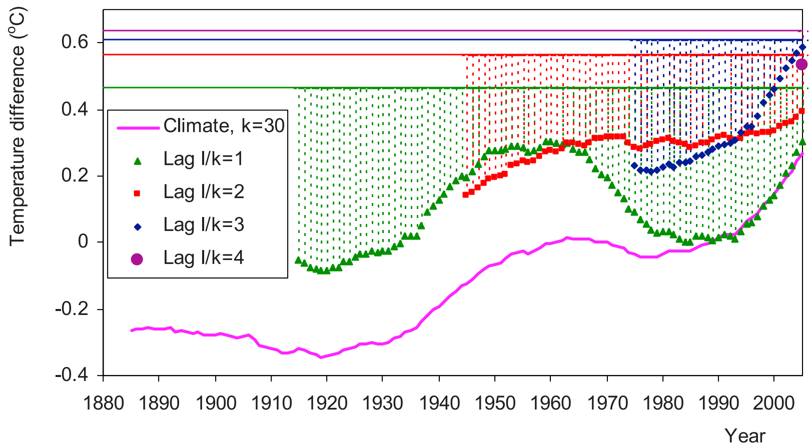

Figure 2. Graphical depiction of the pseudotest based on StD[$D_{i,l}^{(k)}$] with known H. The solid curve represents the CRU time series averaged over climatic scale $k = 30$. The series of points represent values of $D_{i,l}^{(k)}$ for the indicated lags l/k. Horizontal lines represent the critical values of the pseudotest, which are the estimates of StD[$D_{i,l}^{(k)}$] times a factor 2.58, corresponding to a double-sided test with significance level 1% and assuming normality (only the positive critical values are plotted).

(starting from year i). This is indeed an interesting statistic and we wish to discuss it further (noting that similar analyses apply to any type of statistical test). First, $D_{i,l}^{(k)}$ does not depend on a fitted model (as e.g., a linear fitting to the data). Second, it is flexible and convenient as it allows choosing the climatic timescale k and the lag $l/k \geq 1$ (defined on scale k). Third, and more important, it yields a simple, general (not dependent on the process type), convenient and exact expression for the standard deviation of the test statistic, which we have obtained from (1):

$$\mathrm{StD}\left[D_{i,l}^{(k)}\right] = \sqrt{2}\sigma^{(k)}\sqrt{1-\rho_{l,k}^{(k)}} \tag{10}$$

[35] This does not depend on the mean of the process and includes two multiplicative terms, the first ($\sigma^{(k)}$, computed by (2) or (4)) depending on the standard deviation and the autocorrelation structure of the process, and the second (computed by (3) or (5)) depending merely on the autocorrelation structure.

[36] The variation of the two terms with ρ for both the SSS and AR(1) processes is depicted in Figure 1a for the assumptions indicated in the caption. The two terms have opposing effects. The first term increases with ρ, faster in the SSS than in the AR(1) case. The second term is a decreasing function of ρ but in AR(1) it practically equals 1 unless ρ takes very high values (>0.95). The combined effect of the two terms is demonstrated in Figure 1b for $\sigma = 1$. In the SSS case, for relatively low ρ (or H), StD[$D_{i,l}^{(k)}$] is an increasing function of H but for $\rho > \sim 0.70$ it becomes a decreasing function tending to zero as $\rho \to 1$ (because the second term dominates). The situation is similar in the AR(1) case but StD[$D_{i,l}^{(k)}$] becomes decreasing function of ρ only for $\rho > 0.95$.

[37] In all this demonstration it was assumed that both σ and ρ are known. In practice, however both are unknown and estimated from the sample. The picture changes in this case. To estimate StD[$D_{i,l}^{(k)}$], one may be tempted to use the standard estimate s of σ that is used in classical statistics (for example, *Rybski et al.* [2006] do not mention this problem at all). However, as explained above (equation (8)), in SSS statistics, s is strongly biased and $\tilde{\tilde{s}}$ should be used instead; thus, if $s = 1$ then, according to (8) and (6), an approximately unbiased estimate of σ is $[n^{2(1-H)}/(n^{2(1-H)} - 1)]^{1/2}$. It can be seen in Figure 1b that in this case StD[$D_{i,l}^{(k)}$] is an increasing function for virtually the entire domain of ρ.

[38] The effects of autocorrelation to the significance of rejecting the null hypothesis of no change in climate is demonstrated in Figure 1c, assuming that a classical statistical test has already resulted in rejection of the null hypothesis with extremely low risk (i.e., significance level) 10^{-15}. It is observed that the significance level increases dramatically with ρ. For $\rho = 0.7$ the significance level becomes 10^{-2} in the SSS case and 10^{-3} in the AR(1) case. For higher values of ρ both the SSS and the AR(1) processes yield significance levels that are very close to each other; this may be interesting to those who do not appreciate the LTP hypothesis and prefer to assume an STP behavior.

[39] Yet this modified analysis was based on the tacit assumption that the true value of H or ρ is known. However, this assumption is not true and thus the above methodology does not consist a formal test, so we call it a "pseudotest" and anticipate that it only yields a lower bound of the significance level. For unknown H, the estimate of StD[$D_{i,l}^{(k)}$] is anticipated to be greater but its calculation may be intractable by analytical means (given that the estimators of H and σ are dependent [*Koutsoyiannis*, 2003]). A Monte Carlo testing framework becomes then the method of choice (such a framework was proposed in a different context by *Cohn and Lins* [2005], which results in even greater escalation of orders of magnitude of significance level). However, as explained above, the focus of this study is on understanding so we preferred the analytical discussion, even though it yields a pseudotest rather than a formal one.

[40] It may be of some interest to apply this pseudotest to the CRU data series. The application is shown graphically in Figure 2, for a double-sided test for significance level 10^{-2} and for the SSS case, using all possible integer lags l/k

from 1 (l = 30) to 4 (l = 120). In neither case the pseudotest resulted in rejection of the null hypothesis (no change), although it comes close to rejection for 2005 for l/k = 3. As noted above, a real test would be even more tolerant in rejecting the null hypothesis. This result agrees with *Cohn and Lins* [2005] rather than with *Rybski et al.* [2006], who perhaps underestimated some uncertainty factors, as discussed above.

6. Conclusion and Discussion

[41] The above analysis shows that several hydroclimatic tasks, including the detection and attribution problem and the characterization of trends, should be studied in a framework properly recognizing and characterizing the dependence structure of hydroclimatic records, and that the classical IID framework should be abandoned. It also shows that the statistical uncertainty is dramatically increased in the presence of dependence, especially if this dependence is LTP.

[42] Certainly, statistical problems in hydroclimatic research will continue attracting attention in the years to come, as newer data accumulate. Before concrete conclusions can be drawn, a rigorous methodological framework, based on both physical and statistical arguments, should be built. Obviously, the aim of this study was neither to provide such a framework nor to give an answer to the detection and attribution problems. We hope, however, that our remarks may be useful in building this framework.

[43] The answer to the very important question whether the dependence structure of hydroclimatic processes is LTP or STP is very relevant to the detection and attribution problems. However, a categorical answer to this question cannot be based on merely statistical arguments, because, as we demonstrated above, even the presence of LTP can be disputable on purely statistical grounds. Certainly, better physical understanding and theoretical analyses are strongly needed to illustrate and verify or falsify the LTP hypothesis or other climatic hypotheses.

[44] This emphasizes the need of a theory, in addition to statistical tools, to assess the natural behaviors. Without a concrete theoretical framework the situation can be summarized by the following quotation from *Cohn and Lins* [2005, paragraph 1]: "From a practical standpoint . . . it may be preferable to acknowledge that the concept of statistical significance is meaningless when discussing poorly understood systems."

[45] **Acknowledgments.** We are grateful to the Editor, Marc Parlange, the Associate Editor, Timothy Cohn, reviewer Harry Lins, and two anonymous reviewers for their positive and encouraging critiques, suggestions, and comments. D.K. wishes to thank A.M., the research team in DISTART, and the Institute of Advanced Studies of the University of Bologna for their warm hospitality during his visit in Bologna and their collaboration, a result of which was this paper.

References

Beran, J. (1994), *Statistics for Long-Memory Processes*, *Monogr. Stat. Appl. Probab.*, vol. 61, CRC Press, Boca Raton, Fla.

Beran, J., and Y. Feng (2002), SEMIFAR models—A semiparametric approach to modelling trends, *Comput. Stat. Data Anal.*, *40*(2), 393–419.

Bloomfield, P. (1992), Trends in global temperature, *Clim. Change*, *21*, 1–16.

Board on Atmospheric Sciences and Climate (2006), *Surface Temperature Reconstructions for the Last 2,000 Years*, Natl. Acad. Press, Washington, D. C.

Briffa, K. R. (2000), Annual climate variability in the Holocene: Interpreting the message of ancient trees, *Quat. Sci. Rev.*, *19*(1–5), 87–105.

Byers, D., J. Davidson, and D. Peel (2000), The dynamics of aggregate political popularity: Evidence from eight countries, *Electoral Stud.*, *19*(1), 49–62.

Cohn, T. A., and H. F. Lins (2005), Nature's style: Naturally trendy, *Geophys. Res. Lett.*, *32*, L23402, doi:10.1029/2005GL024476.

Craigmile, P. F., P. Guttorp, and D. B. Percival (2004a), Wavelet-based parameter estimation for trend contaminated fractionally differenced processes, *NRCSE Tech. Rep. Ser. 77*, Univ. of Wash., Seattle.

Craigmile, P. F., P. Guttorp, and D. B. Percival (2004b), Trend assessment in a long memory dependence model using the discrete wavelet transform, *Environmetrics*, *15*(4), 35–313.

Eltahir, E. A. B. (1996), El Niño and the natural variability in the flow of the Nile River, *Water Resour. Res.*, *32*(1), 131–137.

Esper, J., E. R. Cook, and F. H. Schweingruber (2002), Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability, *Science*, *295*(5563), 2250–2253.

Esper, J., E. R. Cook, P. J. Krusic, K. Peters, and F. H. Schweingruber (2003), Tests of the RCS method for preserving low-frequency variability in long tree-ring chronologies, *Tree Ring Res.*, *59*(2), 81–98.

Gilman, D. L., F. J. Fuglister, and J. M. Mitchell, Jr. (1963), On the power spectrum of "red noise", *J. Atmos. Sci.*, *20*, 182–184.

Halley, J. M., and P. Inchausti (2004), The increasing importance of 1/f-noises as models of ecological variability, *Fluctuation Noise Lett.*, *4*(2), R1–R26.

Haslett, J., and A. E. Raftery (1989), Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource, *Appl. Stat.*, *38*(1), 1–50.

Hausdorff, J. M., S. L. Mitchell, R. Firtion, C. K. Peng, M. E. Cudkowicz, J. Y. Weil, and A. L. Goldberger (1997), Altered fractal dynamics of gait: Reduced stride-interval correlations with aging and Huntington's disease, *J. Appl. Physiol.*, *82*(1), 262–269.

Hurst, H. E. (1951), Long term storage capacities of reservoirs, *Trans. Am. Soc. Civ. Eng.*, *116*, 776–808.

Jones, P. D., and M. E. Mann (2004), Climate over past millennia, *Rev. Geophys.*, *42*, RG2002, doi:10.1029/2003RG000143.

Jones, P. D., K. R. Briffa, T. P. Barnett, and S. F. B. Tett (1998), High resolution palaeoclimatic records for the last millennium: Interpretation, integration and comparison with GCM control-run temperatures, *Holocene*, *8*(4), 455–471.

Kantelhardt, J. W., S. A. Zschiegnera, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. E. Stanley (2002), Multifractal detrended fluctuation analysis of nonstationary time series, *Physica A*, *316*, 87–114.

Karagiannis, T., M. Molle, and M. Faloutsos (2004), Long-range dependence: Ten years of Internet traffic modeling, *IEEE Internet Comput.*, *8*(5), 57–64.

Kolmogorov, A. N. (1940), Wienersche Spiralen und einige andere interessante Kurven in Hilbertschen Raum, *Dokl. Akad. Nauk URSS*, *26*, 115–118.

Koscielny-Bunde, E., A. Bunde, S. Havlin, and Y. Goldreich (1996), Analysis of daily temperature fluctuations, *Physica A*, *231*, 393–396.

Koscielny-Bunde, E., A. Bunde, S. Havlin, H. E. Roman, Y. Goldreich, and H.-J. Schellnhuber (1998), Indication of a universal persistence law governing atmospheric variability, *Phys. Rev. Lett.*, *81*(3), 729–732.

Koscielny-Bunde, E., J. W. Kantelhardt, P. Braun, A. Bunde, and S. Havlin (2006), Long-term persistence and multifractality of river runoff records: Detrended fluctuation studies, *J. Hydrol.*, *322*, 120–137.

Koutsoyiannis, D. (2002), The Hurst phenomenon and fractional Gaussian noise made easy, *Hydrol. Sci. J.*, *47*(4), 573–595.

Koutsoyiannis, D. (2003), Climate change, the Hurst phenomenon, and hydrological statistics, *Hydrol. Sci. J.*, *48*(1), 3–24.

Koutsoyiannis, D. (2005a), Uncertainty, entropy, scaling and hydrological stochastics, 1, Marginal distributional properties of hydrological processes and state scaling, *Hydrol. Sci. J.*, *50*(3), 381–404.

Koutsoyiannis, D. (2005b), Uncertainty, entropy, scaling and hydrological stochastics, 2, Time dependence of hydrological processes, *Hydrol. Sci. J.*, *50*(3), 405–426.

Koutsoyiannis, D. (2006), A toy model of climatic variability with scaling behaviour, *J. Hydrol.*, *322*, 25–48.

Koutsoyiannis, D., A. Efstratiadis, and K. Georgakakos (2007), Uncertainty assessment of future hydroclimatic predictions: A comparison of probabilistic and scenario-based approaches, *J. Hydrometeorol.*, in press.

Mandelbrot, B. B., and J. W. van Ness (1968), Fractional Brownian motion, fractional noises and applications, *SIAM Rev.*, *10*, 422–437.

Mann, M. E., and J. M. Lees (1996), Robust estimation of background noise and signal detection in climatic time series, *Clim. Change*, *33*, 409–445.
Mann, M. E., R. S. Bradley, and M. K. Hughes (1999), Northern Hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations, *Geophys. Res. Lett.*, *26*(6), 759–762.
Matalas, N. C. (1967), Mathematical assessment of synthetic hydrology, *Water Resour. Res.*, *3*(4), 937–945.
McIntyre, S., and R. McKitrick (2003), Corrections to the *Mann et al.* (1998) proxy data base and Northern Hemispheric average temperature series, *Energy Environ.*, *14*(6), 751–771.
Moberg, A., D. M. Sonechkin, K. Holmgren, N. M. Datsenko, and W. Karlen (2005), Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data, *Nature*, *433*(7026), 613–617.
Monetti, R. A., S. Havlin, and A. Bunde (2003), Long-term persistence in the sea surface temperature fluctuations, *Physica A*, *320*, 581–589.
Montanari, A. (2003), Long-range dependence in hydrology, in *Theory and Applications of Long-Range Dependence*, edited by P. Doukhan et al., pp. 461–472, Springer, New York.
Montanari, A., R. Rosso, and M. S. Taqqu (1997), Fractionally differenced ARIMA models applied to hydrologic time series, *Water Resour. Res.*, *33*(5), 1035–1044.
Pelletier, J. D., and D. L. Turcotte (1997), Long-range persistence in climatological and hydrological time series: Analysis, modeling and application to drought hazard assessment, *J. Hydrol.*, *203*(1–4), 198–208.
Peng, C.-K., S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger (1994), Mosaic organization of DNA nucleotides, *Phys. Rev. E*, *49*, 1685–1689.
Radziejewski, M., and Z. W. Kundzewicz (1997), Fractal analysis of flow of the river Warta, *J. Hydrol.*, *200*, 280–294.
Ray, B. K., and R. S. Tsay (2000), Long-range dependence in daily stock volatilities, *J. Bus. Econ. Stat.*, *18*(2), 254–262.
Rybski, D., A. Bunde, S. Havlin, and H. von Storch (2006), Long-term persistence in climate and the detection problem, *Geophys. Res. Lett.*, *33*, L06718, doi:10.1029/2005GL025591.
Sakalauskienë, G. (2003), The Hurst phenomenon in hydrology, *Environ. Res. Eng. Manage.*, *3*(25), 16–20.
Salas, J. D. (1993), Analysis and modeling of hydrologic time series, in *Handbook of Hydrology*, edited by D. Maidment, chap. 19, pp. 19.1–19.72, McGraw-Hill, New York.
Shiryaev, A. N. (1989), Kolmogorov: Life and creative activities, *Ann. Probab.*, *17*(3), 866–964.
Smith, R. (1993), Long-range dependence and global warming, in *Statistics for the Environment*, edited by V. Barnett and F. Turkman, pp. 141–161, John Wiley, Hoboken, N. J.
Stephenson, D. B., V. Pavan, and R. Bojariu (2000), Is the North Atlantic Oscillation a random walk?, *Int. J. Climatol.*, *20*, 1–18.
Taqqu, M. S., V. Teverovsky, and W. Willinger (1995), Estimators for long-range dependence: An empirical study, *Fractals*, *3*, 785–798.
Vjushin, D., R. B. Govindan, R. A. Monetti, S. Havlin, and A. Bunde (2001), Scaling analysis of trends using DFA, *Physica A*, *302*, 234–243.
von Storch, H. (1995), Misuses of statistical analysis in climate research, in *Analysis of Climate Variability: Applications of Statistical Techniques*, edited by H. von Storch and A. Navarra, pp. 11–26, Springer, New York.
von Storch, H., and F. Zwiers (1999), *Statistical Analysis in Climate Research*, Cambridge Univ. Press, New York.
von Storch, H., E. Zorita, J. M. Jones, Y. Dimitriev, F. González-Rouco, and S. F. B. Tett (2004), Reconstructing past climate from noisy data, *Science*, *306*(5696), 679–682.
Wagenmakers, E. J., S. Farrell, and R. Ratcliff (2004), Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition, *Psychon. Bull. Rev.*, *11*(4), 579–615.
Wegman, E. J., D. W. Scott, and Y. H. Said (2006), Ad hoc committee report on the "hockey stick" global climate reconstruction, report, U.S. Comm. on Energy and Commerce, Washington, D. C. (Available at http://republicans.energycommerce.house.gov/108/home/07142006_Wegman_Report.pdf)
Yue, S., and T. Y. Gan (2004), Simple scaling properties of Canadian annual average streamflow, *Adv. Water Resour.*, *27*, 481–495.
Zorita, E., and H. von Storch (2005), Methodical aspects of reconstructing non-local historical temperatures, *Mem. Soc. Astron. Ital.*, *76*, 794.

D. Koutsoyiannis, Department of Water Resources, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zographou, Greece.

A. Montanari, Department DISTART, Faculty of Engineering, University of Bologna, Via del Risorgimento 2, I-40136 Bologna, Italy. (alberto.montanari@unibo.it)