Confidence intervals for time averages in the presence of long range correlations, a case study on earth surface temperature anomalies

M. Massah[1], H. Kantz[1]

M. Massah, Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Str. 38, D 01187 Dresden, Germany (massah@mpg.pks.de)

[1]Max Planck Institute for the Physics of
Complex Systems, Nöthnitzer Str. 38, D
01187 Dresden, Germany

Time averages, a standard tool in the analysis of environmental data, suffer severely from long range correlations. The sample size needed to obtain a desired small confidence interval can be dramatically larger than for uncorrelated data. We present quantitative results for short and long range correlated Gaussian stochastic processes. Using these, we calculate confidence intervals for time averages of surface temperature measurements. Temperature time series are well known to be long range correlated with Hurst exponents larger than 1/2. Multi-decadal time averages are routinely used in the study of climate change. Our analysis shows that uncertainties of such averages are as large as for a single year of uncorrelated data.

Key Points.

- In the presence of long range correlation, time averages converge very slowly as a function of sample size.
- We show how to calculate error bars for finite time averages over long range correlated data.
- The uncertainty of a 30y average surface temperature at Potsdam (DEU) is plus minus 0.5 C and at Darwin (AUS) is plus minus 0.4 C.

1. Introduction

A commonly used and very easy to apply tool in time series analysis is to compute sample means. E.g., climate change, although being a very complex phenomenon with many facets, in the first place is analyzed in terms of temporal averages of meteorological relevant quantities (temperatures, precipitation) over 20 or 30 years [*Stocker et al.*, 2013] in order to level out intradecadal fluctuations. Much more detailed studies [see e.g. *Stainforth et al.*, 2013] are feasible and reveal many more details, which, however, might be much harder to understand and more difficult to integrate into a general perspective. Therefore, time averages are and will remain an essential tool for understanding climate change.

The precise value of a sample average depends on the details of the sample and hence varies from sample to sample generated by the very same process. If the process is stationary, such averages converge with growing sample size and become independent of the details of the chosen sample. Only then they represent a system property rather than a sample property. The issue of this paper is how slow such convergence can be, i.e., how far an average over a finite time window can be off the asymptotic limit. We do this specifically in view of the common practice to use a 30 year time window to quantify climate.

To be specific, we study time averages of time series data,

$$S_N = \frac{1}{N}\sum_{i=1}^{N} x(t_i) \ , \quad t_i = i\delta t \ . \tag{1}$$

For stationary ergodic processes, S_N in the limit $N \to \infty$ converges to a well defined value, the ensemble mean μ [*Birkhoff*, 1931]. In practice, N is finite, and we want to study how close S_N is to μ as a function of N. This can be expressed by the concept of *confidence interval*.

An $r\%$ confidence interval is defined as the largest interval in the range of a random variable (here: S_N) such that its value is found inside with $r\%$ probability (probability $= r/100$). Commonly used confidence intervals are for probabilities 95%, 99% and 99.9%. Often, the statement is inverted: Given the estimated value of some quantity, its true value (given as ensemble mean of the unknown underlying probability distribution) is with probability $r/100$ within the $r\%$ confidence interval around the estimated value. Hence, confidence intervals are used to quantify the statistical estimation errors and are shown as error bars.

We will analyse the N dependence of confidence intervals, and hence the speed of convergence of S_N to μ, for two classes of correlated Gaussian processes. Whereas short range correlations require a simple correction of sample size in the well known $1/\sqrt{N}$-behaviour, long range correlations lead to a much slower decay of the statistical errors. Based on these results, we will then calculate confidence intervals for time averages of measured temperature time series data. Such data represent in some approximation Gaussian stochastic processes with long range correlations. We compute the confidence intervals of the time averaged temperatures as a function of the time window size. The error bars of thirty year averages of these data are larger than that of a one year average of independently identically distributed (iid) data. The 95% confidence interval in Potsdam/Germany is ±0.5°C, and in Darwin/Autralia is ±0.4° C. We conclude that our knowledge about the true climate is much less precise than one might naively expect but can be quantified precisely.

Effects of long range correlations (LRC) on statistical estimates have become a concern of many researchers in recent years. Using the technique called Detrended Fluctuation Analysis (DFA) [*Peng et al.*, 1992], a huge number of publications has shown the existence of LRC in many data sets: in surface temperature data [*Eichner et al.*, 2003], in many other meteorological time series such as river run-offs [*Koscielny-Bunde et al.*, 2013], in physiological data [*Penzel et al.*, 2003], and in economic data [*Carbone et al.*, 2004]. In the estimation of trends [*Ko et al.*, 2008], in the application of extreme value statistics [*Rust et al.*, 2011], and in the clustering of extreme events [*Bunde et al.*, 2008], LRC have been identified as causes for deviations from standard statistical behavior. In the context of climate change, the significance of estimated trends in the presence of LRC has attracted much attention in the past [*Vyushin et al.*, 2007; *Franzke*, 2012, 2011]. Our focus here, namely the analysis of effects of LRC on confidence intervals for time averages, is technically and conceptually less challenging than the above cited works, but a methodology for determining these has not yet been proposed. A numerical analysis for exponential correlations has been performed by [*Zwiers et al.*, 1995], and the slow convergence of finite sample mean values in the presence of long memory has been mentioned in the statistics literature [see e.g. *Box et al.*, 2016].

We finish this introduction by an interpretation of the above mentioned error bars. Climate data are not generated by Gaussian stochastic processes, but Gaussian models sometimes are good data models for them. When quantifying the "state" of the climate system by a time average, we interprete the observed signal as stochastic (plus seasonal) fluctuations around a given fixed value. If these fluctuations were white noise, time averages over these would converge to zero like $1/\sqrt{N}$. If these fluctuations are LRC, their time average relaxes much more slowly, and we have more difficulties to determine the true off-set. Comparing different time windows, disregarding how little fluctuations are averaged out, we could mis-interpret changes of the mean value as changes of the true climate. For temperatures the assumption of LRC Gaussian fluctuations over a background signal is very realistic and offers a way to quantify the estimation errors of the true "climate state".

2. Confidence intervals of time averages for Gaussian processes

The time average S_N as defined in Eq.(1) is a random variable itself, and its probability distribution is determined by the statistical properties of the time series elements and their particular succession. Here, we analyse these for stationary Gaussian processes, which are completely defined by their mean value and their auto-correlation function. Since the x_t are Gaussian random variables, also S_N is Gaussian distributed around its mean μ. For these distributions, confidence intervals are proportional to their standard deviations σ with the following correspondence (see, e.g., [*Weisstein*, 2016]): The 68% confidence interval is given by $[\mu-\sigma, \mu+\sigma]$, the 95% interval by $[\mu-2\sigma, \mu+2\sigma]$, the 99.7% interval by $[\mu-3\sigma, \mu+3\sigma]$. Therefore, for Gaussian S_N, we have to calculate the standard deviation $\sigma(N)$ of its distribution as a function of sample size N.

2.1. AR(1) process

The *auto-regressive process of first order*, AR(1), is a stationary Gaussian stochastic process with the following iteration rule:

$$x_{i+1} = ax_i + \xi_i \ , \tag{2}$$

where the parameter $|a| < 1$ for stability and ξ_i are Gaussian white noises, $\langle \xi_i \xi_j \rangle = \delta_{i,j}$. The variance is $\langle x^2 \rangle = \sigma^2_{\mathrm{AR(1)}} = 1/(1-a^2)$ and the normalized auto-correlation function is $\langle x_n x_{n+k} \rangle / \sigma^2_{\mathrm{AR(1)}} = a^{|k|}$. Hence, x_n does not diverge to ∞ if $|a| < 1$, and the auto-correlation time is $-1/\ln|a|$.

As can be verified explicitly, the standard deviation $\sigma(N)$ of the distribution of time averages S_N for an AR(1) process behaves asymptotically for large N as

$$\sigma(N) \approx \frac{\sigma_{\mathrm{AR(1)}}}{\sqrt{N/2\tau}} \quad \text{with } 2\tau := \frac{1+a}{1-a} \, . \tag{3}$$

For small N, $\sigma(N)$ is bounded by $\sigma_{\mathrm{AR(1)}}$, the standard deviation of the individual random variable x_i. Asymptotically, the N-dependence is the same as for iid data, but the sample size has to be replaced by an effective sample size $N\frac{1-a}{1+a}$. This is illustrated in Fig.1. One can show that for $a > 0$ and sufficiently different from 0, $\frac{1+a}{1-a} \approx -2/\ln a$, which is twice the correlation time of the process. Hence, we simply have to measure the time window N over which the average is taken in multiples of twice the auto-correlation time. Correlations among the data reduce the effective sample size, since correlated data points contain redundant information and therefore do not reduce the statistical uncertainty.

2.2. ARFIMA model

A class of Gaussian models with long range correlations has been introduced by *Granger et al.* [1980]; *Hoskig et al.* [1981] and is often called *auto-regressive fractionally integrated moving average* model, ARFIMA(p, d, q). The simplest version, ARFIMA(0,d,0), which is also called *fractionally integrated white noise*, can be mapped onto an MA(∞)-model with a particular choice of the coefficients [see *Box et al.* [2016] for details]. This mapping allows us to generate time series data for further numerical analysis, quite easily. The auto-correlation function of ARFIMA(0,d,0) decays like $\tau^{-(1-2d)}$, and the Hurst exponent of this process is $H = d + 1/2$.

The time average S_N is a linear combination of Gaussian random variables and hence its probability density is Gaussian with zero mean. We calculate the N dependence of the standard deviation of this distribution by considering the mean squared displacement MSD(N). A pseudo-Brownian path $W(N)$ is obtained by adding up the output of the ARFIMA model as noises, $W(N) = \sum_{i=1}^{N} \xi_i$, $\langle \xi_i \xi_j \rangle = \delta_{ij}$. Its essential feature is that the Mean Squared Displacement MSD(N) $:= \langle (W(N) - W(0))^2 \rangle$, scales like N^{1+2d}, for $d \in [-1/2, 1/2]$. This reflects the fact that ARFIMA(0,d,0) models create a time discrete version of fractional Gaussian noise (fGn) whose integration leads to fractional Brownian motion (fBm) [*Mandelbrot et al.*, 1968]. The relationship between MSD(N) and S_N is evident: MSD(N) $= N^2 \langle S_N^2 \rangle$. With $\langle S_N \rangle = 0$, the MSD is N^2 times the variance of the probability distribution of S_N. Hence we know that the standard deviation of the distribution of S_N scales like

$$\sigma(N) \approx N^{d-1/2} = N^{H-1} \tag{4}$$

with an unknown pre-factor. Numerical simulations show that if d is only moderately larger than 0, the asymptotic behavior extends to $N = 1$ and hence this pre-factor is given by the standard deviation of x_i, which is $\sigma_{\mathrm{ARFIMA}} = \Gamma(1-2d)/\Gamma^2(1-d)$ [*Box et al.*, 2016]. For d close to 1/2, which is the maximum for a stationary process, the asymptotic behavior sets in only at larger N, see Fig.1. Hence, for $d > 0$ ($H > 1/2$, persistence), the decay of the standard deviation in sample size N is slower than for the AR(1) process, Eq.3 [see also Sec.10.3.2 in *Box et al.* [2016] for an exact expression].

3. Analysis of surface temperature time series

The slowness of convergence of S_N to the true ensemble mean of the underlying process leads to dramatically increased confidence intervals of time averages, if data are LRC. Indeed, for geophysical data, there exists a huge number of results obtained by DFA where authors identified Hurst exponents bigger than 1/2 and hence LRC. We referenced only few of these articles in the introduction.

In this section we investigate data from two different meteorological stations. One is Potsdam Telegraphenberg located at 52.3813 latitude, 13.0622 longitude in 81m above sea level. There exists an excellent record of uninterrupted daily temperature measurements reaching back to January 1st, 1893. The data set of almost 45000 daily values has been validated by the German Weather Service [*DWD*, 2016] and is free for download from their climate data center [*DWD*, 2016]. The station is located close to Berlin in Germany. Due to its mid-latitude position and under the influence of continental air masses from eastern and northern Europe, Potsdam experiences a pronounced seasonality and rather large fluctuations around it. We use here the series of daily average temperatures. We will show that the temperature anomalies are well described by an ARFIMA(1,d,0) model and that we can therefore infer the magnitude of error bars of time averages of these data using the before mentioned results for ARFIMA-models.

We extract a seasonal cycle c_i, ignoring potential non-stationarities, in two ways. One is determined by the Fourier components of ω, 2ω, 3ω with $\omega = 2\pi/365.2425$, the other by low-pass filtering the series of the 123-year mean daily temperature for each calendar day. The latter shows some fluctuations around the former, but these two cycles differ by less than 0.2° C on average, and the distributions and the auto-correlations of these two sets of anomalies agree almost perfectly. So the arbitrariness of how the seasonal cycle is determined does not influence the following analysis.

Subtracting either of these cycles from the measured temperatures leads to the *anomalies*: The deviation of the temperature on a given day from the long year average of this calendar day. These anomalies are to a good approximation Gaussian distributed with zero mean and a standard deviation of about 4°C, see the left panel of Fig.2. The skewness is -0.15, the kurtosis is 3.45, in agreement with the small but visible deviations from Gaussianity. We consider this to be small enough to use a Gaussian data model for these data.

By DFA [*Peng et al.*, 1992] we find a Hurst exponent of $H \approx 0.65$ for this data set, with a scaling behavior over three decades in window size s, see Fig.2, main panel. So these data represent in good approximation an LRC Gaussian process which can be understood as a realization of an ARFIMA model.

A visual inspection of the auto-correlation function shows that in addition to the power law tail, there is some non-trivial short range correlation (not shown here). When ignoring the long range part, we could fit an exponential decay with an auto-correlation time of about 4-5 days. From the meteorological point of view, this correlation time reflects the average lifetime of circulation systems drifting across Germany. A minimal model which can generate both the short range correlations and the power law tail is an ARFIMA(1,d,0) model. We find (by manual adjustment) that the parameter values $a = 0.65$ and $d = 0.15$ yield a time series which reproduces well all properties of the data. The auto-correlation function calculated for a series of equal length as the measured data agrees well with the auto-correlations of the measurement series, within supposed statistical fluctuations.

Hence, we can compare the observed data to four models, where only one reproduces all properties of the data well: By simple random re-shuffling of all data items, we would create iid data with no correlations at all. The standard deviation of the distribution of time averages S_N would decay as $1/\sqrt{N}$. A better model, respecting the short term correlations, would be the AR(1) process with $a \approx 0.8$, also leading to a $1/\sqrt{N}$ behavior, where, however, the sample size has to be divided by twice the correlation time, so that the effective sample size is about 1/10 of the true one. The long range correlations of the data can be well modeled by the ARFIMA(0,d,0) model, which leads to a decay of the standard deviation like $1/\sqrt{N^{1-2d}}$. Finally, as shown before, the most realistic model is the ARFIMA(1,d,0). Its asymptotic behavior is also like $1/\sqrt{N^{1-2d}}$, but due to the fact that the short range correlations require to measure sample size in units of two auto-correlation times, N has to normalized.

We create time series of length 45000 of each of the latter three models. All model data are rescaled to the standard deviation of the measured temperature anomalies. For the measurement data and for these nontrivial models, we now calculate empirically the standard deviations of the time average distributions for various time window sizes N. We do so by segmenting the full 45000 data into disjoint parts and calculate time averages on every segment, limiting segment length to $N = 1000$ for statistical reasons. Thereby we obtain a set of (at least 45, depending on N) values of time averages, whose standard deviation is calculated as usual. Knowing the asymptotic behavior from the previous two sections, we can extrapolate to larger N and specifically to $N = 10950$, which is about 30 years. The results of this analysis are shown in Fig.3.

Only due to the results of Sec.2 we (straight lines in Fig.3), we can assess, e.g., the uncertainty of an average over a 30 year time window. Error bars representing a 95% confidence interval have a size of 2σ, as discussed above. If the Potsdam anomalies were identically independently distributed data, this error would be about 0.07°C. For AR(1) data with the given auto-correlation time, it would be about 0.2°C, for a pure long range correlated data set it would be about 0.3°C, and for the model which fits best the data it is 0.5°C. The same error would be found for an average over only about 260 days of iid data. We also stress that when increasing the time window N even beyond 30 years, the uncertainties of the time average shrink much more slowly for the Potsdam data than for short range correlated data.

Our data analysis was done for temperature anomalies, since only these are to a good approximation consistent with an ARFIMA model. In particular, the temperatures themselves are not Gaussian distributed. Often, these are modeled by a seasonal auto-regressive process (SAR), and we could straightforwardly generalize this to a seasonal ARFIMA process, simply adding the seasonal cycle to the ARFIMA output. However, if we use for the time windows N only values which are multiples of 365.25 days, then the seasonal cycle by construction averages out to a constant off-set. Hence, the time average over these is identical to the time average over anomalies plus this constant. For time averages of, e.g., 30 years, the confidence interval for the true temperatures is identical to the confidence interval calculated for the anomaly time series. The specific result of this analysis is shown in Fig.3 (right) where the mean temperatures over disjoint blocks of 30 years are presented together with their 95% confidence interval. We conclude that the correct error bars for these LRC data are so big that an analysis ending 5 years earlier (thin lines) would not be significant for warming.

We perform the same analysis for temperature recordings from Darwin Airport in Australia, where the data are supplied by Australian bureau of meteorology [*Australian bureau of meteorology*, 2016]. These data represent a tropical climate. The seasonal cycle is more complicated (two maxima) and cannot be well represented by a few Fourier components, hence we use the low-pass filtered empirical calendar day mean values. The total data set covers only 74 years and has several shorter and a few longer gaps, we therefore discard the first 4 years and fill the remaining gaps by randomly drawing anomalies from the distribution obtained from the time series itself. The distribution of the anomalies is Gaussian for positive values but has an exponential tail for negative values (see Fig.2). Its variance, skewness, and kurtosis are 1.5, -0.68, and 4.15, respectively. We nonetheless use the ARFIMA model and find both by a direct parameter fit with the program `forecast` [*CRAN*, 2016] and by manual adjustment trying to match the auto-correlation function the parameters $a \approx 0.26$ and $d \approx 0.28$. As also DFA confirms (Fig.2), these data with $H = 0.78$ have stronger long range correlations than the Potsdam anomalies. The N-dependence of the standard deviation of time averages behaves approximately like $\sigma(N) \approx 1.6N^{-0.22}$ where the exponent is determined through Eq.(4). Hence, a 30 year time average ($N = 10950$) has an uncertainty of $\pm 0.4°$ C (95% confidence interval, $\pm 2\sigma$). The difference between the average temperature in the period 1956-1985 and the period 1985-2015 is 0.29° C, and between 1946-1975 and 1976-2005 is 0.28°C, both well within the error bars. Even though the data model produced Gaussian distributed data and the observed anomalies are not perfectly Gaussian, the empirical error bars which we can determine up to $N \approx 500$ agree well with the ARFIMA model results. This suggests that the Gaussian model works sufficiently well to take the extrapolation to the 30 year error bars seriously.

4. Summary and conclusion

We have discussed a theoretical framework which allows us to calculate the confidence intervals of time averages for short and long range correlated observation data. In practice, the following 3 steps have to be performed:

• construct the series of anomalies, or, more generally, make the sequence as stationary as possible.

• Check for long range correlations, e.g., by DFA, and determine the Hurst exponent.

• While the asymptotic decay of the error bar will be $\propto N^{H-1}$, its pre-factor can be determined by calculating the standard deviations of the distributions of time averages over many short segments of the data as a function of segment length, and matching the asymptotic decay to these.

Hence, fitting of an ARFIMA model is not needed, and also non-Gaussian data can be treated this way. The theoretical foundations are based on a Gaussian data model. For strongly non-Gaussian data such as precipitation or wind speeds, such a model is inappropriate, but the methodology can be applied and leads to a rough error estimate, which is expected to be useful since even for strongly non-symmetric data the distribution of time averages S_N tends towards a Gaussian for large time window N.

As particular results, we found that the 95% confidence intervals for 30-year averages of the Potsdam temperatures (Darwin temperatures) are almost as large as (even larger than) the whole warming effect of the past 100 (50) years. Although the physical consequences of increasing greenhouse gas concentrations are undebatable, this work shows that a quantitative assessment of climate change from observed data is still challenging. The results shown for Potsdam and Darwin admit with low but finite probability both the absence of climate change as well as a warming of already much more than 1° C, and without the past 5 years of data, the same analysis shows much less significance for warming.

As an anonymous referee pointed out to us, it would be very interesting to deconvolve the external driving to the Earth's global mean temperature from observed temperature fluctuations, within in a linear stochastic energy balance model with long range memory, as it was proposed in [*Rypdal et al.*, 2014, 2015]. Apart from issues such as whether an additive decomposition into short term weather fluctuations and long term drivers would be reasonable at all (see [*Lucarini et al.*, 2016] for arguments in favor of this), the clear difficulty in such an endeavor would lie in the fact that time averages over the correlated noise would not as nicely level off to zero as it would be for white noise. So also there the effect of LRC on the uncertainties of mean values would be to be respected.

Acknowledgments

The Potsdam temperature series is publicly available at *DWD* [2016]. The Darwin temperature series is publicly available at *Australian bureau of meteorology* [2016]. We acknowledge helpful discussions with Anna Deluca and Carl Schertel.

References

Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex V. and Midgley P.M. (eds.). (2013). IPCC, Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press*, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.

Stainforth, D. A., Chapman, S. C., Watkins, N. W. (2013). Mapping climate change in european temperature distributions. *Environmental Research Letters.* 8, 3, 034031.

Gardiner, C.W. (1985). Handbook of Stochastic Methods. *Springer*

Birkhoff, G.D. (1931). Proof of the ergodic theorem. *PNAS* 17, 656.

Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M. and Stanley, H.E. (1992). Long-range correlations in nucleotide sequences. *Nature* 356, 168.

Eichner, J.F. et al. (2003). Power-law persistence and trends in the atmosphere: a detailed study of long temperature records. *Phys. Rev. E* 68, 046133.

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

Penzel, T., Kantelhardt, J.W., Grote, L., Peter, J.-H. and A. Bunde. (2003). Comparision of Detrended Fluctuation Analysis and Spectral Analysis for Heart Rate Variability in Sleep and Sleep Apnea. *IEEE Transact. Biomed. Engin.* 50, 1143.

Carbone, A., Castelli, G., Stanley, H.E. (2004). Time-dependent Hurst exponent in financial time series. *Physica A: Stat Mech.* 344 267-271.

Liu, Y., Parameswaran, G. and Stanley, H.E. (1999). Statistical properties of the volatility of price fluctuations. *Phys. Rev. E* 60 1390.

Ko, K., Lee, J. and Lund, R. (2008). Confidence intervals for long memory regressions. *Statistics & Probability Letters* 78, 1894.

Rust, H.W., Kallache, M., Schellnhuber, H.-J. and Kropp, J.P. (2011). Confidence Intervals for Flood Return Level Estimates assuming Long-Range Dependence; in: Kropp, J. and Schellnhuber, H.-J.(Eds.), In Extremis-Disruptive Events and Trends in Climate and Hydrology. *Springer, Berlin.*

Bunde, A., Eichner, J.F., Havlin, S., and Kantelhardt, J.W. (2004). Return intervals of rare events in records with long-term persistence. *Physica A: Stat. Mech.* 342, 308 and more publications from the same group.

Zwiers, F. W. and Storch, H. (1995). Taking Serial Correlation into Account in Tests of the Mean. *J. Climate.* 8 (2) 336351.

Vyushin, D. I., Fioletov, V. E. and Shepherd, T. G. (2007). Impact of long-range correlations on trend detection in total ozone. *J. Geophys. Res.* 112 (D14) D14307.

Franzke, C. (2012). On the statistical significance of surface air temperature trends in the Eurasian Arctic region, *Geophys. Res. Lett.* 39 L23705.

Franzke, C., (2011). Nonlinear trends, long-range dependence, and climate noise properties of surface temperature, *J. Climate.*, 25 4172-4183

Granger, C.W.J. and Joyeux, R. (1980). An Intorduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1, 15-29.

Hosking, J.R.M. (1981). Fractional differencing. *Biometrika* 68,165-176.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (2016). Time Series Analysis: Forecasting and Control. *5th ed., Hoboken, NJ: Wiley*

Mandelbrot, B.B. and van Ness, J. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review* 10 (4) 422.

DWD Climate Data Center (CDC): Historical daily station observations (temperature, pressure, precipitation, wind, sunshine duration, etc.) for Germany, version v004, 2016: Data set description.

The complete meteorological data set of the station Potsdam (station ID 3987) can be downloaded from the Climate Data Center of the German Weather Service DWD via: `ftp://ftp-cdc.dwd.de/pub/CDC/observations_germany/climate/daily/kl/historical/`.

The complete meteorological data set of the Darwin Airport station (station number 14015) can be downloaded from the website of Australian bureau of meteorology via: `www.bom.gov.au/jsp/ncc/cdio/weatherData/`

Forecast package by CRAN R project, which is a package of ”forcasting functions for time series and linear models”. `https://cran.r-project.org/web/packages/forecast/index.html`

Siegert, S., Broecker, J. and Kantz, H. (2015). Skill of data based predictions versus dynamical models - case study on extreme temperature anomalies; in: Chavez, M., Ghil, M. and Fucugauchi, J. U. (Eds.) Extreme Events: Observations, modeling and economics. *AGU Monograph, Washington, DC.*

Weisstein, E. W. ”Confidence Interval.” From MathWorld–A Wolfram Web Resource. `http://mathworld.wolfram.com/ConfidenceInterval.html`

Eichner J.F. et al. (2013). Power-law persistence and trends in the atmosphere: a detailed study of long temperature records. *Phys. Rev. E* 68, 046133.

Rypdal, M. and Rypdal, K. (2014). Long-Memory Effects in Linear Response Models of Earths Temperature and Implications for Future Global Warming. *J. Climate* 27 (14) 52405258.

Rypdal, K., Rypdal, M. and Fredriksen H. B. (2015). Spatiotemporal Long-Range Persistence in Earths Temperature Field: Analysis of StochasticDiffusive Energy Balance Models. *J. Climate* 28 (21) 83798395.

Lucarini, V., Ragone, F. and Lunkeit, F. (2016). Predicting Climate Change Using Response Theory: Global Averages and Spatial Patterns. *J Stat Phys* 129.

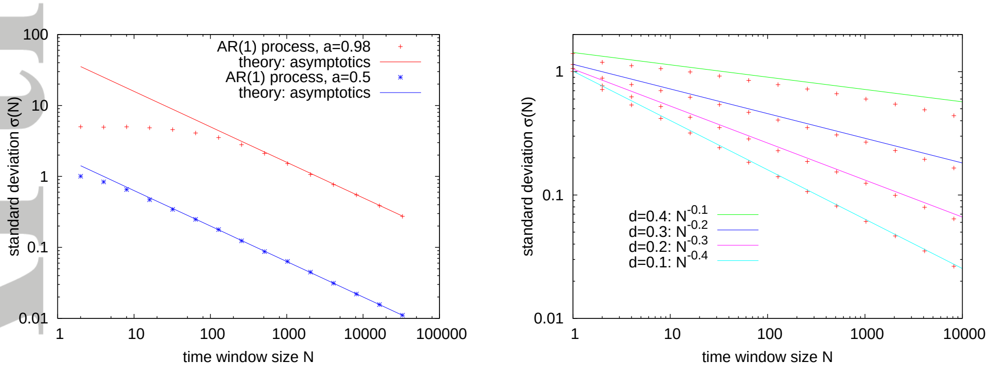

Figure 1. Comparison of AR(1) and ARFIMA(0,d,0) models: Left panel : Numerical estimates of the standard deviation $\sigma(N)$ of the distribution of S_N as a function of window size N obtained from samples of 1000 time averages each, for AR(1) processes with parameters $a = 0.98$ (upper) and $a = 0.5$ (lower). The asymptotic behavior $\propto 1/\sqrt{N(1-a)/(1+a)}$ is shown as lines. Right panel : the same for ARFIMA(0,d,0) processes for different $H = d + 1/2 > 1/2$ (persistent paths) as a function of N. Continuous lines show the asymptotic theoretical predictions $\frac{\sqrt{\Gamma(1-2d)}}{\Gamma(1-d)} N^{-(1/2-d)} \propto N^{-(1-H)}$. For d much larger than 1/2 (H close to 1), the decay in sample size is really slow. Deviations between numerics and asymptotics, in particular for $d = 0.4$, are due to statistical errors in these highly correlated samples.

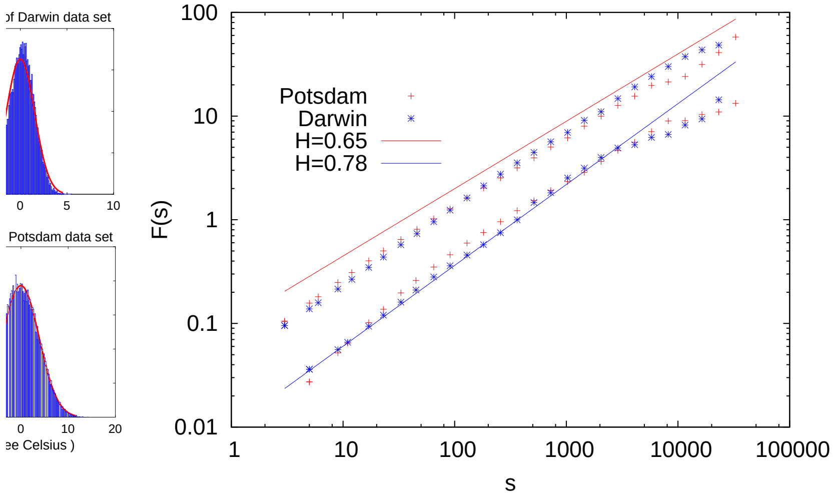

Figure 2. Main panel: The fluctuation function $F(s)$ for DFA0 and DFA2 for the Potsdam for the Darwin temperature anomalies. The data follow power laws $F(s) \approx s^H$ with $H \approx 0.65$ for Potsdam and $H \approx 0.78$ for Darwin.

Small panels: The histograms of the temperature anomalies from Potsdam station (upper) and Darwin (lower). Both series of anomalies are almost stationary across the seasons.

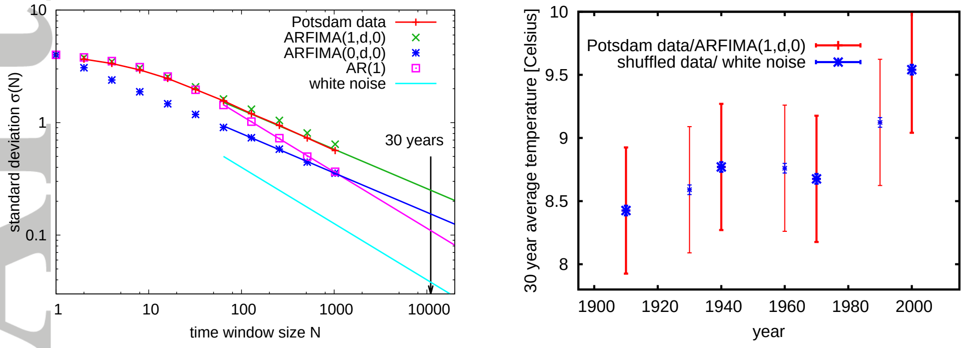

Figure 3. Left panel: The standard deviation of the distribution of finite time averages for Potsdam temperature anomalies and models. Symbols are calculated on data sets of length 45000 each, whereas straight lines represent the theoretical, asymptotic behavior. The arrow indicates N for 30 years. Right panel: Mean temperatures at Potsdam station averaged over 30 years each with non-overlapping windows centered at 1910, 1940, 1970, 2000 (bold), together with their 95% confidence intervals as error bars (red). Blue error bars indicate the precision of the same type of average for iid/shuffled data. Error bars are calculated on the basis of left panel. The same type of average, but windows centered at 1935, 1965, 1995 is printed with thin lines.